Protein Conformational Dynamics In Genomic Analysis

by

Brandon Mac Butler

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2016 by the
Graduate Supervisory Committee:

Sefika Banu Ozkan, Chair
Sara Vaiana
Giovanna Ghirlanda
Robert Ros

ARIZONA STATE UNIVERSITY

December 2016

ABSTRACT

Proteins are essential for most biological processes that constitute life. The function of a protein is encoded within its 3D folded structure, which is determined by its sequence of amino acids. A variation of a single nucleotide in the DNA during transcription (nSNV) can alter the amino acid sequence (i.e., a mutation in the protein sequence), which can adversely impact protein function and sometimes cause disease. These mutations are the most prevalent form of variations in humans, and each individual genome harbors tens of thousands of nSNVs that can be benign (neutral) or lead to disease. The primary way to assess the impact of nSNVs on function is through evolutionary approaches based on positional amino acid conservation. These approaches are largely inadequate in the regime where positions evolve at a fast rate. We developed a method called dynamic flexibility index (DFI) that measures site-specific conformational dynamics of a protein, which is paramount in exploring mechanisms of the impact of nSNVs on function. In this thesis, we demonstrate that DFI can distinguish the disease-associated and neutral nSNVs, particularly for fast evolving positions where evolutionary approaches lack predictive power. We also describe an additional dynamics-based metric, dynamic coupling index (DCI), which measures the dynamic allosteric residue coupling of distal sites on the protein with the functionally critical (i.e., active) sites. Through DCI, we analyzed 200 disease mutations of a specific enzyme called GCase, and a proteome-wide analysis of 75 human enzymes containing 323 neutral and 362 disease mutations. In both cases we observed that sites with high dynamic allosteric residue coupling with the functional sites (i.e., DARC spots) have an increased susceptibility to harboring disease nSNVs. Overall, our comprehensive proteome-wide analysis suggests that incorporating

these novel position-specific conformational dynamics based metrics into genomics can complement current approaches to increase the accuracy of diagnosing disease nSNVs. Furthermore, they provide mechanistic insights about disease development. Lastly, we introduce a new, purely sequence-based model that can estimate the dynamics profile of a protein by only utilizing coevolution information, eliminating the requirement of the 3D structure for determining dynamics.

ACKNOWLEDGMENTS

I graciously thank my PhD mentor, Dr. Banu Ozkan, for her support through my graduate school career and teaching me many important lessons. It is only through her guidance that I have been able to achieve success as a PhD student. She provided me with the foundation to grow as a scientific researcher and we had many constructive discussions. Her knowledge and enthusiasm about the field of biophysics was very inspiring and motivational. I also appreciate her compassion and understanding at a personal level that helped me maintain a positive attitude.

I thank Dr. Avishek Kumar for his help with projects that we worked on together, help enhancing my coding skills, and friendship that all significantly contributed to my success in graduate school. His meticulous nature and programming abilities was an inspiration to break old habits and implement much better practices that will benefit me in the future professionally. I also recognize other members of our research group Dr. Nevin Gerek, Dr. Ashini Bolia, Dr. Taisong Zou, Paul Campitelli, Can (John) Kazan, and Tushar Modi that have all contributed to my success by their support with my projects and valuable discussions during group meetings. Moreover, I would like to acknowledge my undergraduate mentor Dr. Mohammad Huda and friend Dr. Pranab Sarker for teaching me valuable skills that would eventually help me in graduate school.

I am thankful to my parents and sister for their support throughout graduate school. I also thank all of my true friends that have all been instrumental in graduate school. I thank Stuart Sevier for being a good life-long friend and being a part of my inspiration to do physics and obtain a PhD as well as being strongly supportive in my personal life. I also thank Edwin Baldelomar for being a good life-long friend and being

extremely supportive throughout undergraduate and graduate school. I also sincerely thank every one of my true friends that I have made here at ASU for their support.

Dedicated to my family, friends, and cousin, Jason.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

1 INTRODUCTION

> "... 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 jiggling and wiggling of atoms."
> –Richard Feynman, The Feynman Lectures on Physics

Some parts of this chapter are excerpted from:

Kumar, A., Butler, B., Kumar, S., and Ozkan, S.B. "Integration of structural dynamics and molecular evolution via protein interaction networks: a new era in genomic medicine," Current Opinion in Structural Biology 35, 135-142 (2015).

The genome contains the blueprints for the synthesis of proteins, which carry out crucial biological functions. Protein synthesis occurs in the exome (coding region of the genome), in which transcribed DNA (mRNA) is translated on the ribosome to produce a chain of amino acids (polypeptide), which then folds into a unique 3D protein structure. This tertiary structure is defined by the specific sequence of amino acids. Thus, every amino acid sequence encodes a specific 3D protein structure with a particular function. This sequence-structure-function relationship has long been at the epicenter of biological research. Advanced high-throughput sequencing of individual genomes has led to the burgeoning discovery of new sequences, with millions of unique sequences in public databases. Moreover, for the past two decades, scientists have been profiling genomic

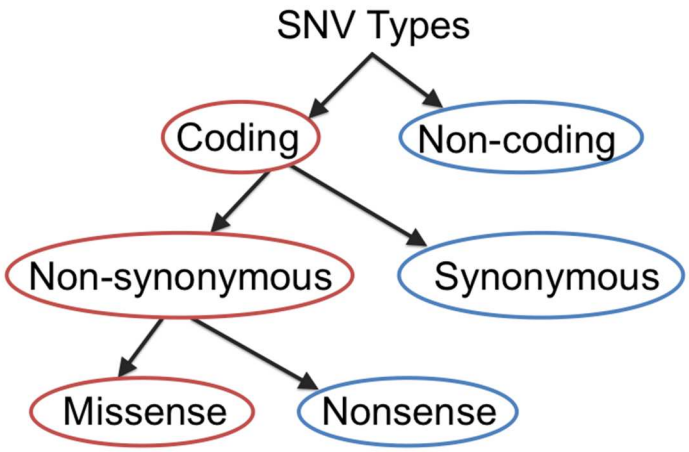

Figure 1.1: Types of single nucleotide variants in the genome. Each personal genome contains around 3 million single nucleotide variants (SNVs), most of which occur in the non-coding region. Comparatively few SNVs are found in the coding region due to strong purifying selection effects. Within the coding region, SNVs can be synonymous (silent) mutations or non-synonymous (nSNVs) which can affect protein function. Most coding SNVs are non-synonymous. In addition, non-synonymous SNVs can either be nonsense mutations (i.e. they result in a stop codon that halts the synthesis of the protein which can lead to disease) or missense mutations, which can lead to the development of various Mendelian or complex diseases.

variations in healthy and diseased individuals. These variations are responsible for the uniqueness between individual genomes (i.e. variations give rise to differences in the population) according to the Neutral Theory of Molecular Evolution (NTME) (Nei *et al.*, 2010). Genome-wide association studies, whole-genome sequencing, and exome sequencing have shown that each personal genome contains millions of genetic variants, thousands of which are related to the development of Mendelian (monogenic) disease or complex (polygenic) disease (Hamosh *et al.*, 2005; Green and Guyer, 2011; Stenson *et al.*, 2014; Sidore *et al.*, 2015). Thus, interpreting these variants and assessing their potential harm is at the forefront of personalized medicine.

The different types of genetic variants will be discussed, as depicted in Figure 1.1. A *single nucleotide variant* (SNV) arises when a single nucleotide base in the three-base-

pair codon is mutated (e.g. GAG to GTG), which can occur in the coding or non-coding region of the genome. In the coding region, the mutated codon can encode a different amino acid or the same amino acid. A coding SNV that leads to an amino acid substitution is termed a *non-synonymous* SNV (nSNV), because it results in a different polypeptide sequence in its corresponding protein (see Figure 1.2). Conversely, a *synonymous* SNV (silent mutation) does not affect the encoded amino acid due to degeneracy in the genetic code (i.e. different codons can code for the same amino acid). Generally, most synonymous variants do not lead to harmful phenotypes, and therefore are usually considered inconsequential. However, synonymous variants should not be overlooked, as they can sometimes impact phenotype by disrupting transcription, translation, splicing, or mRNA stability (Chamary *et al.*, 2006; Goymer, 2007; Kimchi-sarfaty *et al.*, 2007). Moreover, insertions and deletions can also have harmful functional effects and play a role in genetic variation (Mullaney *et al.*, 2010). Most emphasis is currently on non-synonymous SNVs, which can occur in two different forms. A *nonsense* nSNV leads to a premature stop-codon that obstructs the synthesis of the protein, which produces a non-functional protein. Although uncommon, nonsense nSNVs have been implicated in several genetic disorders such as the blood disorder Thalassemia (Chang and Kan, 1979) and several types of muscular dystrophy (Flanigan *et al.*, 2011). The most common non-synonymous variant, a *missense* nSNV (depicted in Figure 1.2), yields an amino acid substitution that results in an altered polypeptide chain. This amino acid change can disrupt post-translational modification, protein folding, stability, binding affinity, and other functional properties (Katsonis *et al.*, 2014). Thus, missense nSNVs

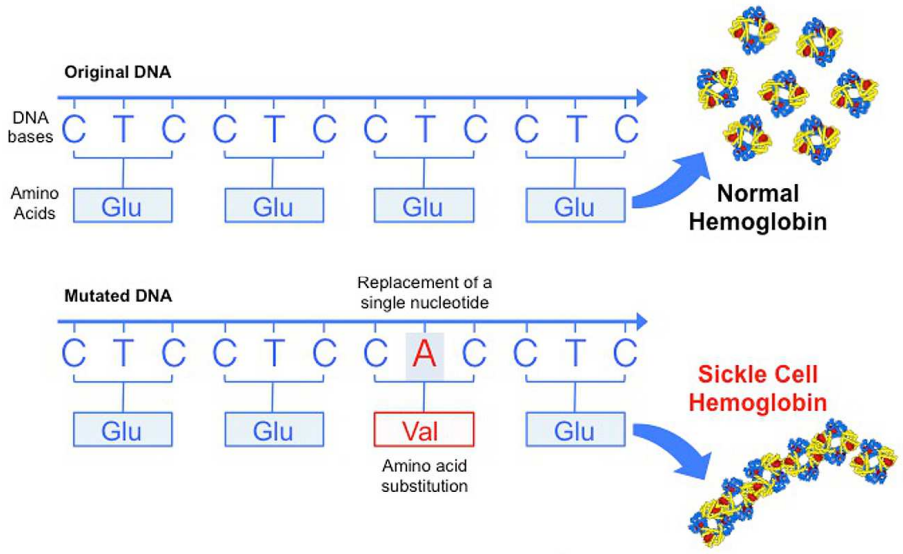

Figure 1.2: A missense mutation of the hemoglobin protein that leads to sickle cell disease. On the top, the wild type nucleotide base pairs on the original DNA template strand (CTC) are transcribed to mRNA, in which the codon GAG (not shown here) encodes the amino acid Glutamic Acid (Glu) in the protein sequence. At the protein level, this results in normal functional hemoglobin that forms into quaternary structures. On the bottom, the mutated DNA contains a missense mutation, where a single nucleotide is mutated from T to A in the template sequence and A to G in the transcribed mRNA sequence (not shown here). This yields a different codon GUG in the transcribed DNA, which encodes a different amino acid Valine (Val). This is known as an amino acid substitution caused by a non-synonymous single nucleotide variant or nSNV (also referred to as missense mutation). The amino acid substitution leads to a new sequence, which expresses as a dysfunctional protein that forms clumps and cannot function. [Source: National Library of Medicine, National Institutes of Health (left panel) and Understanding Evolution. 2016. University of California Museum of Paleontology (www.evolution.berkeley.edu) (right panel)].

can manifest as disease in humans by producing dysfunctional proteins that lead to various ailments. For instance, the genetic disorder Sickle Cell Hemoglobin is attributed to a missense variant in the gene that codes for Hemoglobin, the protein responsible for carrying oxygen in red blood cells. Figure 1.2A depicts the occurrence of this missense variant at the genomic level. A single nucleotide on the DNA template strand is mutated from A to T in the codon GAG, yielding a different codon GTG. The codon GTG is transcribed on the mRNA strand as GUG. Finally, the codon GUG translates to Valine

(Val), which is a different amino acid than without the mutation (wild type). In the wild type, the original codon GAG is translated as Glutamic Acid (Glu), which results in the normal Hemoglobin protein. This particular amino acid substitution gives rise to the disease Sickle Cell Anemia by altering the shape of Hemoglobin, allowing it to form aggregates that inhibit its normal function (Figure 1.2B). Indeed, non-synonymous SNVs are responsible for at least half of all known Mendelian diseases (Hamosh *et al.*, 2005; Stenson *et al.*, 2014). Thus, the challenge is discriminating between nSNVs that will severely impact function (deleterious) and those that are benign (neutral). It is now established that each personal genome contains tens of thousands of nSNVs, most of which are rare (Dudley *et al.*, 2012; Sidore *et al.*, 2015). Deleterious variants are rapidly purged from the population through purifying selection before they have the opportunity to become fixed (Dudley *et al.*, 2012; Tennessen *et al.*, 2012). That is, variants that cause detrimental effects are swiftly eliminated by natural selection in order to preserve the population. The frequency of a variant in a population is given by the minor allele frequency (MAF), where rare variants have MAF <0.5% and common variants MAF >5%. Sorting based on MAF provides a first approximation in isolating potentially harmful variants, which typically have MAF <1% (M. X. Li *et al.*, 2012). Analysis of genomic data has revealed that nSNVs are highly abundant in the non-coding region and less abundant in the coding region (Dudley *et al.*, 2012). Essentially, nSNVs in the highly functional coding region are under intense selection pressure, since they are likely to impact protein function. For this reason, the coding region is disproportionately skewed toward containing rare (MAF <1%) deleterious nSNVs (Marth *et al.*, 2011; Dudley *et al.*, 2012). The common variants (MAF >5%) in the coding region are usually found to be

synonymous SNVs (Dudley *et al.*, 2012). Therefore, predictive studies focus their efforts on rare nSNVs, since they represent the strongest candidates for disease development.

With the sequencing of each new personal exome, the constellation of known nSNVs is expanding at a remarkable rate. But the translation of a personal exome variation profile into biomedically relevant information remains a challenge, especially since the majority of novel nSNVs are rare and hard to detect (Tennessen *et al.*, 2012; Dudley *et al.*, 2012). Multiple databases have been cataloguing human genetic variation in a systematic way that can be utilized in functional studies. For instance, The Human Gene Mutation Database (HGMD) is a collection of nSNVs that are associated with human inherited disease (Stenson *et al.*, 2014). The Online Mendelian Inheritance In Man (OMIM) database contains nSNVs related to all known Mendelian disorders (Hamosh *et al.*, 2005). The dbSNP database (Sherry *et al.*, 2001) contains population nSNVs, including variation data collected from the 1000 Genomes Project (Abecasis *et al.*, 2012). The Genome-Wide Associate Studies (GWAS) project has characterized thousands of nSNVs that are associated with human disease (M. J. Li *et al.*, 2012). Finally, The Cancer Genome Atlas (Chang *et al.*, 2013) and the Catalogue of Somatic Mutations In Cancer (COSMIC) (Forbes *et al.*, 2011) are specific to variants associated with cancer. Experimental studies that elucidate the functional effects of nSNVs are sparse, mainly due to prohibitive cost limitations and time inefficiency. Based upon the current wealth of genomic variation data, efficient and reliable *in silico* tools are needed to interpret the effects of nSNVs, which can be integrated into personalized medicine to diagnose disease susceptibility. To this aim, computational approaches that leverage

genomic variation databases are emerging as the primary way to assess the functional impacts of nSNVs.

Many computational methods that estimate the effect of nSNVs exploit evolutionary information, particularly by using probabilistic scoring functions that leverage positional amino acid conservation and/or phylogenetics based on known sequences (S. Kumar *et al.*, 2009; Liu *et al.*, 2015). A position that is highly conserved in a multiple sequence alignment (MSA) of related homologs is assumed to be essential for function, and thus the occurrence of an nSNV at that position is likely to have a severe effect. Indeed, it has been evinced that deleterious nSNVs are overabundant at highly conserved positions and underabundant at variable positions (Miller and Kumar, 2001; Vitkup *et al.*, 2003; Kumar *et al.*, 2011). Some computational methods to predict deleterious variants based on evolutionary conservation include SIFT (Ng and Henikoff, 2003), Gumby (Prabhakar *et al.*, 2006), and GERP++ (Davydov *et al.*, 2010). In addition to conservation, sequence-based features such as amino acid physicochemical properties (i.e., composition, polarity, and size) is used to quantify the severity of a given amino acid change (Grantham, 1974). Amino acid substitutions that are associated with disease typically exhibit large physicochemical changes between the wild type and mutant amino acid (i.e., radical substitutions), whereas neutral substitutions are less radical (Miller and Kumar, 2001; Vitkup *et al.*, 2003). Several approaches use a combination of sequence conservation and amino acid physicochemical properties such as MAPP (Stone and Sidow, 2005) and Align-GVGD (Tavtigian, 2005). These sequence-based features can also be coupled with machine learning algorithms based on training sets to make enhanced predictions as in PhD-SNP (Capriotti *et al.*, 2006), Parepro (Tian *et al.*, 2007),

and MutationTaster (Schwarz *et al.*, 2010). Although evolutionary conservation has proven to be a very effective tool in nSNV diagnosis, it has posed limitations for evolutionary methods that are dependent on it. Conservation scores lead to distinguished prediction accuracies for damaging nSNVs at highly conserved positions, but the accuracies decrease dramatically for damaging nSNVs at variable positions (S. Kumar *et al.*, 2009; Kumar *et al.*, 2011). In addition, conservation also struggles to correctly identify benign nSNVs at highly conserved positions, since they are presumed to be damaging. Moreover, the accuracy of evolutionary conservation also hinges on the ability to obtain accurate conservation scores from the multiple sequence alignment, which must contain a sufficient number of sequence homologs that are evolutionarily related. Thus, if the quality of the sequence alignment is not optimal, the conservation scores may not accurately portray the most functionally related positions.

Beyond the sequence-based methods, there have been many efforts to utilize structural properties in the diagnosis of nSNVs. The proliferation of available experimental structures in the Protein Data Bank (PDB) (Berman *et al.*, 2000) has allowed for the use of structural information to study human disease, especially with databases that enable mapping of missense mutations to three-dimensional (3D) structures (Luu *et al.*, 2012). As a result, a number of approaches leverage structural information to assess the functional effects of missense mutations such as stability, binding energy, and solvent accessibility. Stability has been widely used to study the effects of mutations. A protein must fold into a stable conformer and adopt a specific 3D structure in order to function. Mutations can destabilize a protein by disrupting essential interactions such as electrostatic interactions, hydrogen bond networks, binding affinities,

and hydrophobic interactions (Steward *et al.*, 2003; Stefl *et al.*, 2013), which can impact function in many ways, including obstructing protein folding, promoting aggregation, and inhibiting the required formation of protein-protein complexes. Many diseases, including Parkinson's disease, Alzheimer's disease, and cancer are associated with destabilizing missense mutations that inhibit normal protein functions (Stefl *et al.*, 2013; Stehr *et al.*, 2011; Shi and Moult, 2011; Lori *et al.*, 2013; Grant *et al.*, 2007; Moore *et al.*, 2003). Indeed, the majority of mutations involved in Mendelian diseases are destabilizing (Yue *et al.*, 2005, 2006; Wang and Moult, 2001), therefore stability is widely used in phenotypic prediction studies. The change in stability due to a mutation is quantified as the change in folding free energy $\Delta\Delta G$ upon mutation (i.e., $\Delta\Delta G = \Delta G_{MT} - \Delta G_{WT}$). This can be estimated using molecular dynamics (MD) simulations with a potential energy function, although it is computationally expensive and only suitable for small-scale studies (Worth *et al.*, 2011). Other *in silico* tools estimate $\Delta\Delta G$ using statistical/empirical potential energy functions based on known structures (Topham *et al.*, 1997; Parthiban *et al.*, 2006; Worth *et al.*, 2011; Guerois *et al.*, 2002) or machine learning algorithms based on structure and/or sequence (Cheng *et al.*, 2006; Capriotti, Fariselli, and Casadio, 2005; Capriotti *et al.*, 2004; Capriotti, Fariselli, Calabrese, *et al.*, 2005), or a combination of both (Pires *et al.*, 2014; Dehouck *et al.*, 2011; Masso and Vaisman, 2010). These integrated computational approaches are faster and more efficient in stability estimations than MD, making them more attractive for large-scale applications in nSNV prediction studies. However, recent surveys have revealed that their capacity in diagnosing nSNVs are quite limited, with modest accuracies around 60% (Potapov *et al.*, 2009; Khan and Vihinen, 2010). Another structural attribute, solvent accessibility, is a scoring metric of

sites according to their location on the 3D structure (e.g. buried or exposed) and is used to study the functional effect of nSNVs (Dobson *et al.*, 2006; Wei *et al.*, 2012; David *et al.*, 2012). A review of different structural attributes used in nSNV studies found that the feature based on nearest neighbors (13Å structural neighbor profile) was more successful at predicting disease association than solvent accessibility, and it was nearly as successful as sequence conservation, which emphasizes the importance of microenvironments around nSNVs in determining their functional effects (Ye *et al.*, 2007). Some methods use machine learning methods to integrate both evolutionary and structural features into their approaches such as PolyPhen-2 (Adzhubei *et al.*, 2010), SVM-3D (Capriotti and Altman, 2011), SNAP (Bromberg and Rost, 2007), SNPs3D (Yue *et al.*, 2006), and nsSNPAnalyzer (Bao *et al.*, 2005). Finally, there are also consensus methods such as PredictSNP (Bendl *et al.*, 2014) that classify nSNVs based on the combined results of many different prediction methods. The advent of using these structural features combined with already successful evolutionary features was optimistically thought to lead to advancements in disease prediction accuracies. However, the incorporation of structural information has only resulted in a marginal ~4% increase in disease prediction accuracies compared to evolutionary methods. This lackluster improvement is primarily because of two reasons. First, commonly used structural features (i.e., solvent accessibility, structural neighbor profiles, b-factors, and secondary structure) are based on a static protein structure and does not account for the intrinsic dynamic motions of a protein, which are critical for assessing functional importance. Thus, the inclusion of proteins dynamics is paramount in prediction analysis to further improve the accuracies of evolutionary methods. Second, proteins do not function in isolation; rather, they

interact with each other in order to function, thus, protein-protein interactions should also be considered. Although the importance of conformational dynamics is recognized, it has not been introduced to genomic analysis due to the lack of position-specific based metrics that probe dynamics. Here, we develop a dynamics-based technique to fill this gap.

In Chapter 2, the methodological details implemented in the work presented thesis will be outlined. An in-depth review of the theoretical approaches to investigate conformational dynamics of proteins will be presented. Full atomic models such as molecular dynamics (MD) and normal mode analysis (NMA) that are based on complex empirical potential energy functions will be discussed as well as their limitations and lack of applicability to large-scale proteomic studies. Then coarse-grained approaches based on elastic network models (ENM) will be discussed, specifically the Gaussian network model (GNM), anisotropic network model (ANM), and perturbation response scanning (PRS). Their broad range of applicability will be highlighted to motivate their use in the proteome-wide studies presented in this thesis, as well as some of their inherent limitations.

Currently, most machine learning methods that utilize structural features are based on static 3D structures, which neglect protein conformational dynamics. However, protein structure-encoded dynamics, which span a broad timescale of motion from atomic fluctuations and side chain rotations to collective domain movements, underlie a protein's biological function. Protein evolution studies of several different protein families have shown that changes in conformational dynamics through allosteric regulation lead to new functions (e.g., green fluorescent protein (GFP), beta-lactamase inhibitors, and nuclear receptors (Glembo *et al.*, 2012; Zou *et al.*, 2015; Kim *et al.*, 2015)). Moreover

evolutionary rates are strongly correlated with the flexibility of individual positions obtained from conformational dynamics (Nevin Gerek *et al.*, 2013; Liu and Bahar, 2012; Liberles *et al.*, 2012). Protein dynamics studies assert that protein function can be explained by analyzing the individual contribution of residues to the conformational dynamics and stability of a protein (Nevin Gerek *et al.*, 2013; Liu and Bahar, 2012; Butler *et al.*, 2015). Therefore, conformational dynamics-based metrics can also be utilized in predicting the impact of nSNVs on protein function. Gerek et al. used an amino acid site-specific *dynamic flexibility index* (DFI) metric to evaluate the effect of flexibility of individual sites on biological fitness and function. DFI is a position–specific metric that quantifies the resilience of each residue to a perturbation occurring at another part of the chain (i.e., all other residues in the network), thus identifying the flexible and rigid parts of a protein (Nevin Gerek *et al.*, 2013). Analysis of disease-associated and neutral nSNVs for more than 100 human proteins revealed that disease-associated nSNVs occur predominately at low DFI sites (i.e., rigid hinge sites), signifying the importance of hinge sites that control functionally crucial motions. In contrast, neutral variants are more abundant at positions with high DFI, suggesting that flexible sites are more robust to mutations (Nevin Gerek *et al.*, 2013). Furthermore, DFI profiles of over a thousand sites harboring mutations revealed that sites at protein interfaces have lower average DFI than those at non-interfaces, suggesting that protein–protein interfaces have less dynamic flexibility (Butler *et al.*, 2015). These results suggest that hinge positions at interfaces are crucial for binding, thus mutations at these hinge sites will likely be damaging.

Another way to assess the phenotypic effects of nSNVs is by considering the interactions among protein-protein complexes. A structural mapping of observed nSNVs

on human protein–protein interaction (PPI) networks revealed that damaging nSNVs are largely found at protein–protein interfaces (Wang *et al.*, 2012). For this reason, some methods have focused on modeling interfaces and predicting changes in binding affinities to distinguish damaging from benign nSNVs. The proliferation of available experimental structures in the Protein Data Bank (Berman *et al.*, 2000) and current advancements in homology modeling have facilitated the development of human structural interaction network (HSIN) databases of protein–protein and domain–domain interactions (Mosca *et al.*, 2012), and mapping nSNVs to three-dimensional (3D) structures (Luu *et al.*, 2012). The structural mapping of nSNVs has revealed that nSNVs at interfaces may disrupt or enhance protein–protein interactions, thus they are commonly implicated in pathogenesis (Schuster-Böckler and Bateman, 2008; David *et al.*, 2012). Similar to core residues, interface residues are generally more hydrophobic, thus mutations involving polar or charged residues may destabilize important interface interactions necessary for binding (Yates and Sternberg, 2013). The loss of obligate electrostatic interactions due to interface mutations may lead to complete loss of function of the complex. On the other hand, mutations that enhance binding interactions may cause aggregation or aberrant recognition, as observed in cancers (Yates and Sternberg, 2013). Because interface mutations can drastically affect binding interactions, efforts to predict the effects of mutations by measuring the difference between the free energy change upon binding of the wild type and the mutant ($\Delta\Delta G$) have shown success. Free energy differences upon binding calculated via thermodynamic integration and free energy perturbation approaches using molecular dynamics (MD) are computationally expensive, particularly for large–scale protein complexes. Therefore, *in silico* tools have been developed as a fast

alternative to estimate $\Delta\Delta G$ upon binding using statistical energy functions based on known protein structures (Dehouck *et al.*, 2013; Li *et al.*, 2014), empirical force fields (Schymkowitz *et al.*, 2005; Guerois *et al.*, 2002), and/or machine learning techniques using training sets (Berliner *et al.*, 2014; Zhao *et al.*, 2014). However, the accuracy of these calculations are not robust because local structural changes upon mutations are generally neglected (Potapov *et al.*, 2009; Khan and Vihinen, 2010). In particular, the change in physicochemical properties upon mutation, such as large changes in polarity and hydrophobicity, do not significantly alter the binding energy, making it challenging to assess nsSNVs (Teng *et al.*, 2009). Evaluating the importance of individual interface residues to binding can be used to predict the effect of nsSNVs, as only a small fraction of interface residues contribute significantly to binding. Experimental methods to identify critical binding sites involve mutating each site to alanine and measuring the corresponding change in binding affinity. The sites that contribute the most to binding energy are known as *hotspots* (Bogan and Thorn, 1998). These hotspots are often located at highly conserved positions with large changes in accessible surface area (ASA) upon binding (Keskin *et al.*, 2008; Tuncbag *et al.*, 2010). Thus, methods to predict hotspot residues at binding interfaces are largely based on ASA as in the webserver HotPoint (Tuncbag *et al.*, 2010). If an nSNV occurs at a hotspot site, it will likely be damaging since it will drastically affect crucial binding interactions. Incorporating hotspots into machine learning methods has been successful in predicting disease nSNVs at protein-protein interfaces (Schuster-Böckler and Bateman, 2008). It remains a challenge to predict the effect of interface nSNVs occurring at non-hotspots.

Chapter 3 will provide insight into how conformational dynamics can elucidate the mechanism of disease association of nSNVs in protein-protein complexes. We will introduce the dynamic flexibility index (DFI) as a tool to estimate site-specific conformational dynamics, which is based on the coarse-grained ENM and PRS method (discussed in Chapter 2). The majority of proteins must form complexes in order to function, thus interfaces residues are critical to the overall stability and function of interacting proteins. It was found using DFI that interface sites have lower flexibility compared to other parts of the protein, and that nSNVs at these sites are highly susceptible to disease. Using DFI, the phenotype of nSNVs at interfaces could be discriminated, whereas the static-based metric accessible surface area (ASA) commonly employed in evolutionary methods was not indicative of phenotype. Evolutionary methods alone are highly suitable for nSNV diagnosis for highly conserved residues, but their accuracy is remarkably low for less conserved residues. This weakness indicates a necessity for improvement for the purposes of overall more accurate predictions. We will show that DFI has the capacity to predict nSNV phenotypes in the less conserved regime, where evolutionary methods are inadequate, making it a useful tool that can complement existing evolutionary methods and increase overall prediction accuracies.

Allostery is the mechanism for regulation of cellular functions through the alteration of protein dynamics and structure based on an action at a distant site. Allostery is an inherent property in maintaining the function and stability of biomacromolecules, thus it is recognized as a crucial factor in disease development and is used to design allosteric drugs (Wagner *et al.*, 2016; Nussinov, 2016). Pathogenicity can result from the disruption of allosteric regulation in a number of ways. For instance, nSNVs can impair

allosteric post-translational modification as observed in driver mutations in cancer (Nussinov and Tsai, 2015; Nussinov *et al.*, 2013). Deleterious nSNVs can also change the ON/OFF populations in cell signaling by altering the stability of certain conformations and/or dynamics. Mutations may lead to disease by shifting allosteric pathways, as observed in the nSNV that gives rise to Hyperekplexia (Shan *et al.*, 2012). Conformational dynamics is directly connected to allostery in proteins, which was evinced by MD simulations and NMR spectroscopy (Guo and Zhou, 2016; Boulton and Melacini, 2016; Lisi and Loria, 2016). An MD analysis conducted for disease mutations of human ferritin (Kumar, Glembo, *et al.*, 2015) showed that mutations distally located to functionally critical sites can allosterically impair hinges of the protein (i.e., rigid parts), softening the functionally critical regions, which leads to the loss of allosterically-regulated conformational dynamics. Allostery also provides a mechanism for the severe impact of nSNVs that are located distally to hotspot residues but are dynamically linked to hotspots (Tennessen *et al.*, 2012). Hotspots evaluated by the HotPoint server (Tuncbag *et al.*, 2010) of the protein complexes in the dataset studied in Butler et al. (Butler *et al.*, 2015) indicated that most mutations occurring at hotspots are deleterious. However, among the hundred deleterious nSNVs at interfaces, only ~50% were located at hotspots. This raised the question as to how nSNVs at non-hotspot sites were so significant in disease development. This was studied using a new dynamic-based metric called the dynamic coupling index (DCI) (Kumar, Glembo, *et al.*, 2015). DCI quantifies the resilience of each residue upon the perturbation of only the functionally crucial sites (e.g., hotspots, catalytic sites, ligand binding sites). Thus, DCI enables the identification of dynamically linked sites that are important for allosteric regulation in the protein.

Intriguingly, ~80% of deleterious mutations at non-hotspots exhibited high DCI (i.e., strongly dynamically linked to hotspots), indicating that allosteric impairment of the hotspot residues resulted in a loss of function.

In chapter 4, we will focus on the allosteric impact of nSNVs in enzymatic proteins that have known catalytic functions. We will present a case study of the GCase protein that is implicated in Gaucher disease. Although it has been well-studied over the past century, a plausible mechanism has yet to be firmly established. All-atom molecular dynamics (MD) simulations were performed for the native protein as well as 4 known mutants in order to elucidate the change in dynamic flexibility (DFI) upon mutation. The MD results indicated an overall rigidification of the mutated proteins, including their active sites and ligand recognition sites, which are important for enzymatic function. The loss of flexibility inhibits the mobility of the ligand recognition sites, making it impossible for them to reorient themselves to accommodate binding substrates for function. Moreover, when the active sites become exceedingly rigid, their functional efficiency is hampered. This provides a mechanism to Gaucher disease, which is associated with a loss of catalytic efficiency. In addition, we examine the allosteric effects of the 4 mutations, which are distal to the active sites, on global conformational dynamics. We implement a new dynamic feature, dynamic coupling index (DCI) that measures the resilience of a site to a perturbation at functionally critical residues. The results showed a remarkable decrease in allosteric dynamic linking to the active sites for each mutation. Each mutation severely impaired the global allosteric coupling to the active sites, which altered the conformational dynamics of the protein. This highlights the relationship of conformational dynamics and allostery and points to the importance of

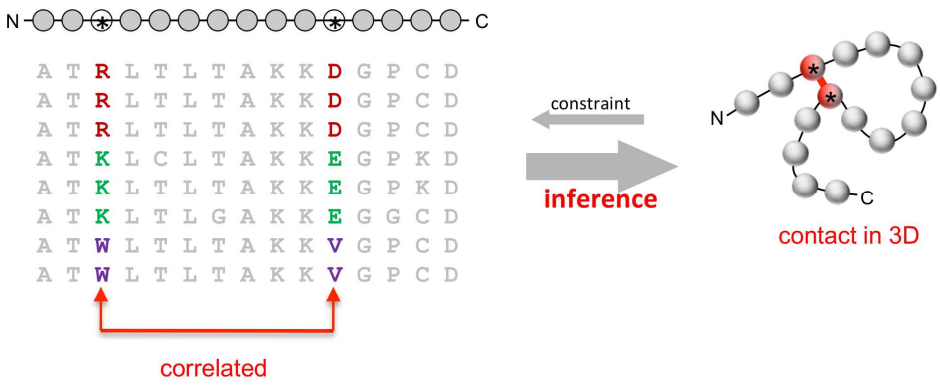

Figure 1.3: Coevolving residues in two columns of a multiple sequence alignment (MSA) (left) are used to infer structural contacts in the tertiary structure (right). Evolutionary couplings are often close in proximity in the 3D protein structure. They represent an evolutionary constraint on function between similar proteins (homologs) in a protein family. From an MSA of a given protein family, the 3D structure can be predicted using coevolving residues [Source: (Marks *et al.*, 2011)].

dynamics in disease development. We extended the analysis to a large and diverse set of enzymatic proteins to see if the DFI and DCI metric could be used to diagnose nSNVs on a large proteomic scale. Both DFI and DCI were able to distinguish between neutral and deleterious nSNVs. Moreover, for a subset of cases where common evolutionary methods (i.e., EvoD, PolyPhen-2, and SIFT) misdiagnosed a given nSNV, our dynamics based metrics were remarkably effective in diagnosing them with accuracies around 72%. This study reveals that protein conformational dynamics can be a complementary feature in genomic variation analysis and disease prediction, as also substantiated by the former analysis (Chapter 3) and the work of Gerek et. al.

It has recently become possible to utilize evolutionary sequence variation in protein families to examine structural and molecular properties of proteins. Specifically, the coevolution of amino acids throughout the evolution of sequences in a family gives insight to their relationship in the 3D structure (see Figure 1.3). Amino acid positions that exhibit concordant substitution patterns (coevolve) in an alignment of sequence homologs

in a given family are evolutionary couplings (ECs). The degree of coupling is indicative of their spatial proximity in the tertiary structure (i.e. strong couplings are representative of native 3D contacts). As shown in Figure 1.3, there are two columns *i, j* where amino acids coevolve in the multiple sequence alignment, which serve as an evolutionary constraint on function. From this, it can be inferred that these two residues *i, j* are in spatial proximity in the 3D folded structure. This technique is particularly valuable in the prediction of non-local contacts where the structure is unknown. The coevolution of amino acids from sequences of protein families has been used to successfully reproduce known 3D structures (Marks *et al.*, 2011; Morcos *et al.*, 2011) including notoriously difficult membrane proteins (Hopf *et al.*, 2012). It has also been used to predict the interactions between protein complexes (Hopf *et al.*, 2014). Moreover, coevolution can be used to determine important functional sites (Marks *et al.*, 2012), allosteric interactions in proteins for drug discovery (Wagner *et al.*, 2016), and functional landscapes of complex biomolecules (Jana *et al.*, 2014).

Chapter 5 will focus on the use of coevolution to approximate the dynamic behavior of a protein strictly from protein sequences and no *a priori* knowledge of the structure. Typically a crystal 3D structure must be provided when obtaining the conformational dynamics of a protein. In many cases, however, the structure is unavailable since there are vastly more known sequences than there are available structures in the PDB. To address this, we introduce a new way to obtain approximate protein dynamics with only knowledge of a protein sequence. The 3D structural contacts are estimated from the co-evolution of residues in a family of protein sequences. Then by taking the 3D contacts to be nodes in the Gaussian Network Model (GNM) the

vibrational dynamics are calculated from the mean-square fluctuations (B-factors). Using 3D contacts in known experimental crystal structures, the GNM has been used to accurately predict mean-square fluctuations of native proteins. A Kirchhoff matrix is formed according to network connectivity, where non-bonded contacts exist between residues within a specified cutoff distance surrounding a given residue in the 3D network. As an alternative to using the known structure, evolutionary couplings can equally serve as non-bonded contacts in the Kirchhoff matrix. The inverse of the Kirchhoff matrix yields the theoretical mean-square fluctuation values (B-factors) for each site, which compare well to experimental B-factors and those calculated using the known structure. Thus, in this novel approach, the protein vibrational dynamics can be explored using exclusively sequence information and no *a priori* knowledge of structure. Given the remarkably disproportionate number of sequences available compared to experimental structures, this could be an invaluable tool in genomic variation studies that focus on disease prediction for proteins that have unknown structures. Moreover, this technique can be utilized in aiding in *de novo* structure prediction and refinement.

In short, this thesis will elucidate the potential role of protein conformational dynamics in genomic variation analysis. The current methods for diagnosing the functional impact of variants are based on evolutionary methods and have well-known weaknesses. Therefore, there is clearly a need for improvement in order to optimize the accuracy of these diagnosis methods. The evolutionary methods have attempted to include structure in their predictions, which resulted in a disappointing ~4% increase in accuracy, mainly due to the use of a static protein structure. In this thesis, we show that protein conformational dynamics provides insights that cannot be obtained using the

static structure and can distinguish between disease and neutral phenotypes of known variants. In particular, our dynamics-based metrics have the capacity to diagnose variants in the regimes where evolutionary methods are inadequate, which provides another step forward in accurate variant diagnosis to assess an individual's predisposition to disease based on their genome. Lastly, the discovery of novel protein sequences continues to outpace the experimental determination of 3D crystal protein structures, therefore the use of conformational dynamics in genome variation studies is inherently limited. We address this dilemma by proposing a new method to approximate protein dynamics using only sequence information and no *a priori* knowledge of the 3D structure. This would extend the range of applicability of conformational dynamics to match evolutionary methods that also rely on sequence information. The ultimate goal in genomic analysis and phylomedicine is to better understand the impact of variants on phenotype in humans and assess an individual's predisposition to disease. The work presented in this thesis will make the case that conformational dynamics of proteins can be used in conjunction with existing methods to achieve this goal.

CHAPTER 2

2 COMPUTATIONAL METHOD FOR STUDYING PROTEIN FUNCTIONAL DYNAMICS

2.1 Introduction

A myriad of resolved 3D folded protein structures have been characterized and deposited in the Protein Data Bank (PDB) (Berman *et al.*, 2000) based on x-ray crystallography and NMR experiments. Despite the abundance of known structures, leveraging structural information to gain mechanistic insights about how a mutation can lead to disease development remains a challenge. Overall, the current methods exploring the effect of mutations on a protein employ a static structure, e.g. solvent accessibility, which quantifies the amount of solvent that is accessible to the surface (Tsodikov *et al.*, 2002). While this static-based metric has shown some success in studying the effect of mutations due to its simplicity and fast computation (Wei *et al.*, 2012; Adzhubei *et al.*, 2013), it does not account for the intrinsic dynamics of proteins which is important for function. Indeed, the biological function of a protein is based on collective motions that sample different conformational states around the equilibrium structure, such as the open and closed forms of an enzyme. We devised a dynamic flexibility index (DFI) as a measure to explore site-specific dynamics of proteins, and it has been used to elucidate the mechanism of disease development (Nevin Gerek *et al.*, 2013; Butler *et al.*, 2015; Kumar, Glembo, *et al.*, 2015). DFI uses the coarse-grained technique known as Perturbation Response Scanning, which is based on an Elastic Network Model (ENM)

and Linear Response Theory. The ENM, while rooted in the classical full-atomistic Normal Mode Analysis procedure, is much more computationally efficient and yields nearly identical results. In this chapter, the all-atom models of Molecular Dynamics and Normal Mode Analysis will be discussed, followed by the detailed formulation of the coarse-grained approaches utilizing ENMs: Gaussian Network Model, Anisotropic Network Model, and Perturbation Response Scanning.

All-atom molecular dynamics (MD) is an atomistic approach that solves Newton's equations of motion for all atoms in a protein using an empirical potential energy function (McCammon *et al.*, 1977; Levitt, 1983). MD predicts the intrinsic dynamics of a native protein around its equilibrium conformation (i.e., mean-square displacement of atoms) and has been widely used over the past 40 years to investigate protein dynamics. Despite its success, MD has some limitations: First, it often fails to predict large-scale collective motions of globular proteins (e.g. domain movements); Second, the complexity of the force fields used to compute equations of motion at the atom level make these calculations computationally expensive. These factors impose strict limitations on the size of proteins that can be used in MD calculations. Another approach to investigate protein dynamics, normal mode analysis (NMA), revealed that the low-frequency modes ($f < 30$ cm^{-1}) are responsible for cooperative motions (i.e., functionally-related motions) of a protein and also contribute significantly to the entropy of the system, whereas the fast modes correspond to localized movements of few atoms and are less crucial for function (Bahar and Rader, 2005; Go *et al.*, 1983; Brooks *et al.*, 1983; Eyal *et al.*, 2011). NMA has been successfully used predict the mean-square displacements in globular proteins, such as the bovine pancreatic trypsin inhibitor (Go *et*

al., 1983). Similar to MD, however, NMA calculations also depend on complicated potentials, which impose the same computational restrictions as MD. Moreover, NMA yields unstable modes that can be difficult to extract, leading to final configurations often disagree with experimental observations (Tirion, 1996).

Low-frequency modes of motion (soft modes), which usually correspond to functional motions, can be found using simplified mechanical models, independently of complicated and computationally expensive force-field potentials (Doruker *et al.*, 2000). For instance, the dynamics of G-actin found was obtained using a single-parameter Hookean potential and turned out to be in close agreement with the dynamics obtained by NMA (Tirion, 1996). Subsequently, coarse-grained approaches were developed that enjoy simplified potentials to obtain soft mode dynamics such as the Gaussian network model (GNM) (Bahar *et al.*, 1997; Haliloglu *et al.*, 1997), the anisotropic network model (ANM) (Atilgan *et al.*, 2001), and perturbation response scanning (PRS) (Atilgan and Atilgan, 2009; Atilgan *et al.*, 2010; Gerek and Ozkan, 2011). These coarse-grained models offer simplicity and fast computational times, yet they can effectively capture functional motions due to soft modes independently of complicated empirical potentials used MD and NMA (Atilgan *et al.*, 2010; Bahar and Rader, 2005; Cui and Bahar, 2005). Moreover, these simplified approaches have made it feasible for computational investigations of the dynamics of large proteins (>500 residues) (Hinsen, 1998; Hinsen *et al.*, 1999; Tama and Sanejouand, 2001) as well as proteome-wide studies to study the functional impact of nSNVs on a wide range of proteins (>100 proteins) (Nevin Gerek *et al.*, 2013; Butler *et al.*, 2015; Zheng and Tekpinar, 2009). This chapter will focus on the

details of the primary coarse-grained approaches–GNM, ANM, and PRS–and their applications in investigating functional protein dynamics.

2.2 Coarse-grained Approaches

2.2.1 Gaussian Network Model

The Gaussian network model (GNM) is an isotropic model for protein dynamics based on contact topology and is rooted in the early work of random polymer networks (Flory, 1976; Pearson, 1977; Kloczkowski *et al.*, 1989). The protein is coarse-grained such that the C-alpha atoms represent nodes in an elastic network, and the interactions between each node is approximated by an elastic spring if the distance between them is within a specified cutoff distance. In contrast to the empirical force-field potentials used in MD and NMA (Equation 2.1), the pairwise interaction between inter-connected C-alpha atoms are given by a simple Hookean single-parameter potential (Tirion, 1996) of the form

$$V_{ij} = \frac{\gamma}{2}\left(|\boldsymbol{R}_{ij}| - |\boldsymbol{R}_{ij}^{0}|\right)^{2} \tag{2.1}$$

Where $\boldsymbol{R}_{ij}$ and $\boldsymbol{R}_{ij}^{0}$ are the instantaneous and equilibrium separation vectors between connected pairs (see Figure 2.1) and γ is a uniform constant. In this topological model, sequentially-bonded residues (covalent bonds) are in contact similar to the Rouse chain model (Rouse, 1953); in addition, non-bonded residues that are sequentially and spatially distant can also be in contact (tertiary contacts) if the distance between them, $\boldsymbol{R}_{ij}$, is less than a specified cut-off distance, R_c. A sufficient cut-off distance for residue contacts was determined to be $R_c \leq 7$Å (Bahar *et al.*, 1997). Essentially, an interaction sphere with radius R_c can be imagined to surround a residue *i,* such that any other residue *j* within the

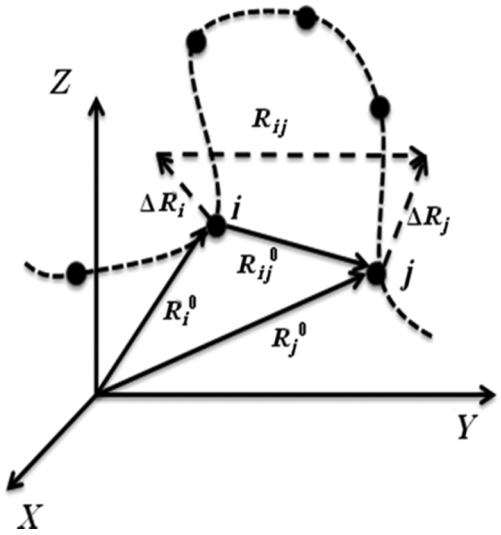

Figure 2.1: Equilibrium fluctuations of C-alpha atoms in the GNM. A section of a protein is shown as a dashed line with black dots representing the Cα atoms as nodes in the elastic network. The equilibrium position vectors are R_i^0 and R_j^0, which correspond to two C-alpha atoms i and j. Similarly, their instantaneous position vectors are R_i and R_j. Their separation distance vectors are R_{ij}^0 (solid line) and R_{ij} (dashed line) respectively. The fluctuation of the equilibrium position vectors are ΔR_i and ΔR_j in the X, Y, and Z directions between i and j and the change in inter-residue distance can be expressed as $\Delta R_{ij} = R_{ij} - R_{ij}^0 = \Delta R_j - \Delta R_i$. [Source: (Eyal et al., 2011)]

sphere will be in contact with residue *i.* A connectivity matrix Γ (Kirchhoff) is then constructed as

$$\Gamma_{ij} = \begin{cases} -1 & i \neq j \text{ and } \boldsymbol{R}_{ij} \leq R_c \\ 0 & \text{i} \neq j \text{ and } \boldsymbol{R}_{ij} > R_c \\ -\sum_{i,i\neq j} \Gamma_{ij} & i = j \end{cases} \tag{2.3}$$

Where $\boldsymbol{R}_{ij}$ is the separation distance between residue pairs *i* and *j*, and R_c is the specified cut-off distance (Haliloglu *et al.*, 1997). The Kirchhoff matrix has dimensions $N \times N$ with *N* being the total number of C-alpha atoms in the protein. The off-diagonal elements of each row are assigned –1 if the pairs are in contact and zero if they are not in contact. The diagonal element for each row is found by taking the negative sum of all other

elements in that row via the sum $\Gamma_{ii} = -\sum \Gamma_{ij}$. The diagonal of the Kirchhoff matrix represents the local packing density surrounding each residue i (Bahar *et al.*, 1997; Haliloglu *et al.*, 1997; Bagci *et al.*, 2002).

Residues in the native protein undergo Gaussian-distributed isotropic thermal fluctuations about their mean positions given by $\Delta\boldsymbol{R}_i$ and $\Delta\boldsymbol{R}_j$ as in Figure 2.1 (Bahar *et al.*, 1997; Haliloglu *et al.*, 1997). The conformational potential energy of the system is then

$$V_{GNM} = \frac{\gamma}{2}\left[\sum_{i,j}^{N}\left(\Delta\boldsymbol{R}_{ij} - \Delta\boldsymbol{R}_{ij}^{0}\right)^2\right] = \frac{\gamma}{2}\left[\sum_{i,j}^{N}\Delta\boldsymbol{R}_i\Gamma_{ij}\Delta\boldsymbol{R}_j\right] \tag{2.4}$$

Where γ is a uniform constant, Γ is the $N \times N$ Kirchhoff matrix, and $\Delta\boldsymbol{R}$ is a $N \times 1$ column vector comprised of all residue fluctuations $\{\Delta\boldsymbol{R}_1, \Delta\boldsymbol{R}_2, \Delta\boldsymbol{R}_3, \dots, \Delta\boldsymbol{R}_N\}$ (Eyal *et al.*, 2011). It follows that the configurational partition function can be written similarly as in the theory of random polymer networks (Flory, 1976; Pearson, 1977; Kloczkowski *et al.*, 1989) as

$$Z_N = \frac{\gamma}{2}\int exp[-V/k_BT]\, d\{\Delta\boldsymbol{R}\} \tag{2.5}$$

Where V is the GNM potential in Equation 2.4, k_B is the Boltzmann constant, T is the temperature of the system, and $d\{\Delta\boldsymbol{R}\} = \{d\Delta\boldsymbol{R}_1, d\Delta\boldsymbol{R}_2, d\Delta\boldsymbol{R}_3, \dots, d\Delta\boldsymbol{R}_N\}$. The cross-correlation of fluctuations $\Delta\boldsymbol{R}_i$ and $\Delta\boldsymbol{R}_j$ between C-alpha atoms i and j is given by their statistical ensemble average (Pearson, 1977; Kloczkowski *et al.*, 1989):

$$\begin{aligned}\langle\Delta\boldsymbol{R}_i \cdot \Delta\boldsymbol{R}_j\rangle &= \frac{1}{Z}\int\left(\Delta\boldsymbol{R}_i \cdot \Delta\boldsymbol{R}_j\right)exp[-V/k_BT]d\{\Delta\boldsymbol{R}\} \\ &= (3k_BT/\gamma)[\Gamma^{-1}]_{ij}\end{aligned} \tag{2.6}$$

The mean-square fluctuations of each residue *i* can be calculated by taking *i=j* in Equation 2.6 as

$$\langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_i \rangle = \langle (\Delta \boldsymbol{R}_i)^2 \rangle = (3k_B T/\gamma)[\Gamma^{-1}]_{ii} \tag{2.7}$$

Since the determinant of the Kirchhoff is zero, the inverse Kirchhoff Γ^{-1} cannot be evaluated directly. Instead, the pseudoinverse is evaluated by single value decomposition (SVD) using eigenvalues λ_k and eigenvectors $\boldsymbol{u}_k$ of Γ as

$$\langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_i \rangle = 3k_B T/\gamma \sum_k^{N-1} [\Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_i]_k = 3k_B T/\gamma \sum_k^{N-1} \left[\lambda_k^{-1} \boldsymbol{u}_k \boldsymbol{u}_k^{\mathrm{T}}\right]_{ij} \tag{2.8}$$

Where the summation is over all $N-1$ non-zero modes, *k*, of Γ ($1 \leq k \leq N-1$) (Atilgan *et al.*, 2001). The eigenvalues λ_k and eigenvectors $\boldsymbol{u}_k$, which correspond to the frequency and shape of each of the *k* modes of motion respectively, allow us to analyze the form of each mode separately. Particularly, the contribution of correlated motion by the *k*th mode is given by

$$[\Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_i]_k = \frac{3k_B T}{\gamma} \lambda_k^{-1} [\boldsymbol{u}_k]_i [\boldsymbol{u}_k]_j \tag{2.9}$$

Where $[\boldsymbol{u}_k]_i$ is the *i*th element of $\boldsymbol{u}_k$. A plot of $[\boldsymbol{u}_k]_i^2$ with respect to residue index *i* is the normalized distribution of mean-square fluctuations in the *k*th mode (*k*th mode shape). The slowest modes are related to broad collective motions of residues that are usually involved principally in biological function. They may be cooperative motions such as domain movements to accommodate a substrate or bind to a partner to function. In essence, the slow modes represent the most function-related modes. Conversely, the fastest modes correspond to highly constrained residues undergoing very small and fast movements in their local environment.

The cross-correlation of fluctuations $\Delta \boldsymbol{R}_i$ and $\Delta \boldsymbol{R}_j$ is the same as in an $N \times N$ covariance matrix (i.e. $C_{ij} = \langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j \rangle$), and the largest contribution to the covariance comes from the slowest modes (Eyal *et al.*, 2011). Thermal fluctuations in the GNM are isotropic such that

$$\langle (\Delta X_i)^2 \rangle = \langle (\Delta Y_i)^2 \rangle = \langle (\Delta Z_i)^2 \rangle = \langle (\Delta R_i)^2 \rangle / 3 \tag{2.9}$$

And are proportional to the crystallographic B-factors determined by experiment

$$B_i = \frac{8\pi^2}{3} \langle (\Delta \boldsymbol{R}_i)^2 \rangle = \frac{8\pi^2 k_B T}{\gamma} [\Gamma^{-1}]_{ii} \tag{2.10}$$

Several studies have shown the B-factors predicted by GNM are in good agreement with crystallographic B-factors for a diverse set of proteins (Bahar *et al.*, 1997; Haliloglu *et al.*, 1997). A proteome-wide analysis by Kundu et al. found that for 113 high-resolution structures (resolution <2.0 Å) the B-factors predicted by the GNM correlated significantly with observed B-factors (Kundu *et al.*, 2002). Moreover, mean-square fluctuations from GNM were also shown to correlate with ANM, MD simulations, crystallography, and NMR (Doruker *et al.*, 2000).

Equations 2.6 and 2.7 yield the magnitude of cross-correlations and mean-square fluctuations and, thus, have no direction dependence. The GNM does not require directionality or the 3D coordinates of residue displacements – only the contact topology of C-alpha atoms (the relative distances between pairs) is necessary to obtain $\langle (\Delta \boldsymbol{R}_i)^2 \rangle$. This simplistic model allows for fast, efficient calculations of protein dynamics with minimal input. The computational bottleneck of the GNM is inverting the Kirchhoff using SVD, which depends on the size of the protein. This still pales in comparison to the computational cost of MD or NMA calculations. The GNM describes isotropic

fluctuations of the $N-1$ modes of motion of a protein in *N*-dimensional configurational space. The incorporation of directional preferences of 3D motion is accounted for in the anisotropic network model, which will be elaborated on in the following section.

2.2.2 Anisotropic Network Model

The Anisotropic Network Model (ANM) is a coarse-grained approach where a protein is viewed as an elastic network model, akin to the GNM described in the previous section. Whereas in the GNM all residue fluctuations are isotropic, the ANM differs in that it accounts for the 3D directionality of fluctuations (i.e. fluctuations can be anisotropic). Therefore the ANM has the capacity to evaluate the 3D character of the normal modes of a protein. Consider two interacting residues *i* and *j* (Figure 2.1) connected by a an elastic spring with the ANM potential

$$\begin{aligned} V_{ij} &= \frac{\gamma}{2}\left(R_{ij}-R_{ij}^{0}\right)^{2} \\ &= \frac{\gamma}{2}\left(\left[\left(X_j-X_i\right)^{2}+\left(Y_j-Y_i\right)^{2}+\left(Z_j-Z_i\right)^{2}\right]^{1/2}-R_{ij}^{0}\right)^{2} \end{aligned} \tag{2.11}$$

Where γ is a uniform spring constant and R_{ij} and R_{ij}^{0} are the instantaneous and equilibrium separation between *i* and *j* ($R_{ij}=R_j-R_i$). X_i, X_j, Y_i, Y_j, Z_i, and Z_j are the vector components of R_i and R_j in Figure 2.1. This potential can also be written in terms of the network of *N* residue pairs as

$$\begin{aligned} V_{ANM} &= \frac{\gamma}{2}\sum_{i,j}^{N}\left(R_{ij}-R_{ij}^{0}\right)^{2} f\left(R_{ij}^{0}\right) \\ &= \frac{\gamma}{2}\sum_{i,j}^{N}\left[\left(X_j-X_i\right)^{2}+\left(Y_j-Y_i\right)^{2}+\left(Z_j-Z_i\right)^{2}\right]^{1/2} f\left(R_{ij}^{0}\right) \end{aligned} \tag{2.12}$$

Where $f\left(R_{ij}^0\right)$ is a Heaviside step function such that $f\left(R_{ij}^0\right) = \begin{cases} -1 & R_{ij}^0 \leq R_c \\ 0 & R_{ij}^0 > R_c \end{cases}$ (i.e. interacting pairs are assigned -1 whereas non-interacting pairs 0) (Bahar and Rader, 2005; Atilgan *et al.*, 2001; Tama and Sanejouand, 2001). Alternatively, $f\left(R_{ij}^0\right)$ can also represent an exponential decay function that attenuates with increasing separation distance (Hinsen, 1998). For the ANM, a cutoff distance of $R_c = 13\text{Å}$ was found to produce the most realistic vibrational frequency distribution (Atilgan *et al.*, 2001). The first-order derivative of V with respect to the X component of $\boldsymbol{R}_i$ in for two interacting residues i and j are

$$\begin{aligned} \frac{\partial V}{\partial X_i} &= -\frac{\partial V}{\partial X_j} \\ &= -\gamma\left(X_j - X_i\right)\left(1 - R_{ij}^0/R_{ij}\right) \end{aligned} \tag{2.13}$$

And the second-order derivative is

$$\begin{aligned} \frac{\partial^2 V}{\partial^2 X_i^2} &= \frac{\partial^2 V}{\partial^2 X_j^2} \\ &= \gamma\left[1 + R_{ij}^0\left(X_j - X_i\right)^2/R_{ij}^3 - R_{ij}^0/R_{ij}\right] \end{aligned} \tag{2.14}$$

Similar equations hold for the Y and Z components of $\boldsymbol{R}_i$. At equilibrium, $R_{ij} = R_{ij}^0$ such that these equations become

$$\frac{\partial V}{\partial X_i} = -\frac{\partial V}{\partial X_j} = 0 \tag{2.15}$$

$$\frac{\partial^2 V}{\partial^2 X_i^2} = \frac{\partial^2 V}{\partial^2 X_j^2} = \gamma\left(X_j - X_i\right)^2/R_{ij}^2 \tag{2.16}$$

And similarly the cross-derivatives between the X and Y components are

$$\frac{\partial^2 V}{\partial X_i \partial Y_j} = -\gamma\left(X_j - X_i\right)\left(Y_j - Y_i\right)/R_{ij}^2 \tag{2.17}$$

When considering all neighbors j of residue i then the second-order derivatives can be expressed as

$$\frac{\partial^2 V}{\partial X_i^2} = \gamma \sum_j \left(X_j - X_i\right)^2 / R_{ij}^2 \tag{2.18}$$

$$\frac{\partial^2 V}{\partial X_i Y_i} = \gamma \sum_j \left(X_j - X_i\right)\left(Y_j - Y_i\right)/R_{ij}^2 \tag{2.19}$$

Where the summation is over all neighbors j of residue i. For a protein with N residues, the second-order derivatives are stored in a $3N \times 3N$ Hessian matrix $\boldsymbol{H}$, which has the form

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{h}_{11} & \boldsymbol{h}_{12} & \cdots & \boldsymbol{h}_{1N} \\ \boldsymbol{h}_{21} & \boldsymbol{h}_{22} & \cdots & \boldsymbol{h}_{2N} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{h}_{N1} & \boldsymbol{h}_{N2} & \cdots & \boldsymbol{h}_{NN} \end{bmatrix} \tag{2.20}$$

Where each super-element $\boldsymbol{h}_{ij}$ is a 3×3 matrix containing all second-order cross-derivatives given by Equation 2.17 for off-diagonal super-elements $\boldsymbol{h}_{ij}$ $(i \neq j)$ as

$$\boldsymbol{h}_{ij} = \begin{bmatrix} \frac{\partial^2 V}{\partial X_i X_j} & \frac{\partial^2 V}{\partial X_i Y_j} & \frac{\partial^2 V}{\partial X_i Z_j} \\ \frac{\partial^2 V}{\partial Y_i X_j} & \frac{\partial^2 V}{\partial Y_i Y_j} & \frac{\partial^2 V}{\partial Y_i Z_j} \\ \frac{\partial^2 V}{\partial Z_i X_j} & \frac{\partial^2 V}{\partial Z_i Y_j} & \frac{\partial^2 V}{\partial Z_i Z_j} \end{bmatrix} \tag{2.21}$$

The diagonal super-elements $(i = j)$ of $\boldsymbol{H}$ are constructed as

$$\boldsymbol{h}_{ii} = \begin{bmatrix} \dfrac{\partial^2 V}{\partial X_i^2} & \dfrac{\partial^2 V}{\partial X_i Y_i} & \dfrac{\partial^2 V}{\partial X_i Z_i} \\ \dfrac{\partial^2 V}{\partial Y_i X_i} & \dfrac{\partial^2 V}{\partial Y_i^2} & \dfrac{\partial^2 V}{\partial Y_i Z_i} \\ \dfrac{\partial^2 V}{\partial Z_i X_i} & \dfrac{\partial^2 V}{\partial Z_i Y_i} & \dfrac{\partial^2 V}{\partial Z_i^2} \end{bmatrix} \tag{2.22}$$

Where the diagonal and off-diagonal elements are constructed using Equation 2.18 and 2.19 respectively. The Hessian matrix in the ANM is analogous to the Kirchhoff matrix in the GNM. The single value decomposition of the Hessian results in $3N - 6$ non-zero eigenvalues λ_k and eigenvectors $\boldsymbol{u}_k$ from which its pseudoinverse can be written as

$$\boldsymbol{H}^{-1} = \sum_{k}^{3N-1} \left[\lambda_k^{-1} \boldsymbol{u}_k \boldsymbol{u}_k^{\mathrm{T}}\right] \tag{2.23}$$

Now the cross-correlations between residue fluctuations in the ANM are calculated as

$$\begin{aligned} \langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j \rangle &= \langle \Delta X_i \Delta X_j \rangle + \langle \Delta Y_i \Delta Y_j \rangle + \langle \Delta Z_i \Delta Z_j \rangle \\ &= \frac{k_B T}{\gamma} \left[\boldsymbol{H}^{-1}_{3i-2,3j-2} + \boldsymbol{H}^{-1}_{3i-1,3j-1} + \boldsymbol{H}^{-1}_{3i,3j}\right] \end{aligned} \tag{2.24}$$

And the mean-square fluctuations are calculated as

$$\begin{aligned} \langle (\Delta \boldsymbol{R}_i)^2 \rangle &= \langle (\Delta X_i)^2 \rangle + \langle (\Delta Y_i)^2 \rangle + \langle (\Delta Z_i)^2 \rangle \\ &= \frac{k_B T}{\gamma} \left[\boldsymbol{H}^{-1}_{3i-2,3i-2} + \boldsymbol{H}^{-1}_{3i-1,3i-1} + \boldsymbol{H}^{-1}_{3i,3i}\right] \end{aligned} \tag{2.25}$$

Unlike the GNM, the fluctuations of ΔX_i, ΔY_i, and ΔZ_i are treated separately in the ANM to account for anisotropy. Similar to the GNM, theoretical B-factors can be calculated in the ANM in terms of the inverse Hessian as $B_i = \frac{8\pi^2 k_B T}{3\gamma} tr(H_{ii}^{-1})$. In a study of the α-amylase inhibitor protein, the 3D molecular motions predicted by ANM correlated well with all-atom MD simulations and experiment, which was superior to that of the GNM

(Doruker *et al.*, 2000). ANM has been used in multiple studies to probe the directionality of collective dynamic motions brought about by the low-frequency modes of proteins. For instance, ANM was used to predict fluctuation dynamics of the slow modes of the retinol binding protein that correlated with experimental values (Atilgan *et al.*, 2001).

The GNM and ANM are both based on elastic networks, and they each have applications in which one is more useful over the other. In general, the GNM is preferred when seeking the magnitude of fluctuations as well as analyzing the contribution of fluctuations to individual modes. Indeed, it has been evinced that B-factors estimated by the GNM are more comparable with experimental B-factors than those of ANM (Kundu *et al.*, 2002). However, the motions of biomolecules occur in all directions, thus incorporating anisotropic effects is important to for gaining mechanistic insights of function, particularly when analyzing directional components of motion. ANM accounts for the magnitude and direction of fluctuations resulting from the slowest modes, which represent an accurate depiction of natural cooperative motions of proteins undergoing biological function. Thus analyzing directional motion of biomolecules requires the use of ANM instead of the GNM to produce the most realistic model. While both GNM and ANM are exponentially less computationally intensive than NMA or MD, GNM offers a time-scale advantage over ANM, which only needs to invert a $N \times N$ matrix as compared to a $3N \times 3N$ matrix. An application of elastic network models to capture the effects of external perturbations on a protein structure will be discussed in the following section.

2.2.3 Perturbation Response Scanning

The elastic network models (ENM) described above are ways to obtain equilibrium fluctuation profiles of proteins by extracting the most essential normal modes

(i.e., low-frequency modes). Although they are simple and efficient, they are also restrictive since they only measure correlations between residue fluctuations around the equilibrium state. Inducing a perturbation in the elastic network and measuring the dynamic responses can reveal underlying information about energy landscapes that go beyond the equilibrium fluctuations. Thus, several studies have used a modified ENM approach in an attempt to measure the effect of perturbing the network, where structural perturbations were induced by adjusting the strength of force constants of springs between interacting residues (Zheng *et al.*, 2007) or changing the distance between interacting residues (Zheng and Brooks, 2005). Another ENM-based technique known as Perturbation Response Scanning (PRS) measures the residue fluctuations of the elastic network upon perturbation of a single node (i.e., C-alpha atom). This perturbation on a node acts to mimic the natural scenario of an approaching ligand exerting a force on the binding pocket. Accordingly, its application has been used in a wide range of studies, including elucidating mechanisms of ligand binding (Atilgan and Atilgan, 2009), binding-induced conformational changes (Atilgan *et al.*, 2010; Bolia and Ozkan, 2016), allostery in small PDZ domain proteins (Gerek and Ozkan, 2011), identifying critical regions for function/stability (Abdizadeh *et al.*, 2015), and improving flexible docking for ligand binding (Bolia and Ozkan, 2016). PRS is a coarse-grained approach in which the native protein is modeled as a 3D elastic network (ENM) as described above. A perturbation in the form of a random external force (i.e. Brownian kick) is sequentially applied on each C-alpha atom in the network (see Figure 2.2). This leads to a cascade of perturbations throughout the network, where the resulting displacements of all other residues are given by linear response theory.

Linear response theory (LRT) was used in structural biology to investigate the structural changes of a protein upon ligand binding (Ikeguchi *et al.*, 2005). The theory asserts that the mechanical response behavior upon a binding event is linearly related to the intrinsic equilibrium fluctuations in the unperturbed state (i.e., unbound ligand-free state). Thus, structural changes elicited by such a binding even can be estimated simply with the knowledge of the residue cross-fluctuations in the ENM, $\langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j \rangle$. Given an external force applied on a residue j in the unperturbed state, the response of another residue i is given by

$$\langle \Delta \boldsymbol{R}_i \rangle \cong \frac{1}{k_B T} \sum_j \langle \Delta \boldsymbol{R}_i \Delta \boldsymbol{R}_j \rangle_0 \boldsymbol{f}_j \tag{2.26}$$

Where $\langle \Delta \boldsymbol{R}_i \Delta \boldsymbol{R}_j \rangle_0$ is the fluctuation covariance in the unperturbed state, and $\boldsymbol{f}_j$ is a single-vector representing the external force acting on residue *j* (Ikeguchi *et al.*, 2005). The covariance term $\langle \Delta \boldsymbol{R}_i \Delta \boldsymbol{R}_j \rangle_0$ can be equivalently obtained from either an all-atom MD simulation or the inverse Hessian in ENM (Equation 2.23). The structural changes predicted by Equation 2.25 is not dependent on the direction or position of the applied external force (Ikeguchi *et al.*, 2005), so $\boldsymbol{f}_j$ is essentially a random external force. The approximation of LRT is realistic in the regime where the low-frequency modes of a protein contribute the most to conformational changes upon binding; Hence, if faster modes contribute more to conformational changes, LRT is not applicable due to unrealistically large external forces, $\boldsymbol{f}_j = k_B T \sum_i \langle \Delta \boldsymbol{R}_j \Delta \boldsymbol{R}_i \rangle_0^{-1} \langle \Delta \boldsymbol{R}_i \rangle_f$ (Yang *et al.*, 2014). This is usually not the case, however, since high-frequency modes generally correspond to miniscule changes such as side change rotations that do not drastically affect the overall structural conformation. Conversely, larger conformational changes that relate to

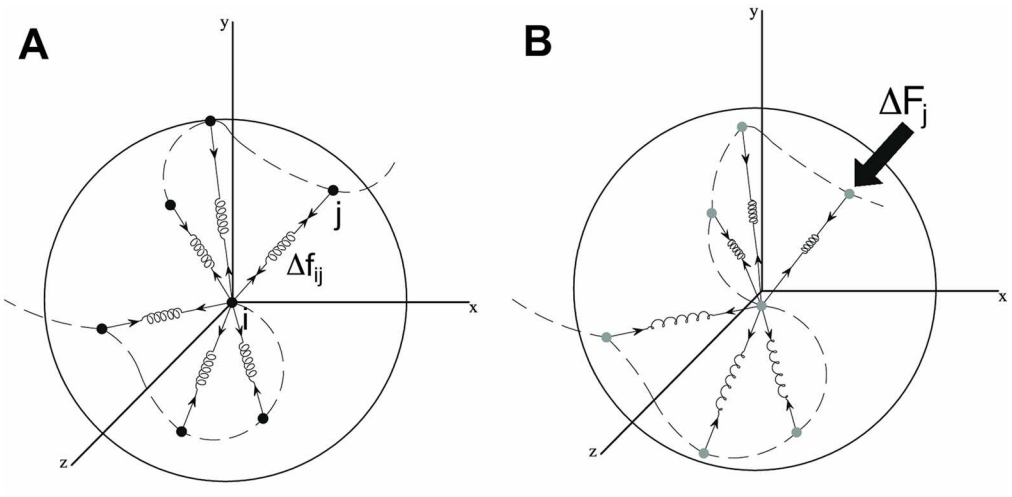

Figure 2.2: A free-body diagram illustrating the Perturbation Response Scanning method (PRS). Each sphere contains the pairwise interactions between C-alpha atoms (black dots in (A) and grey dots in (B)). The origin is taken to be a single residue *i*, which is connected to other residues *j* via elastic springs (given by Equation 2.1). (A) The unperturbed elastic network where no external forces (perturbation) are being applied, and the network is in the equilibrium state, where only the internal forces holding the network in equilibrium are at play. (B) An external force (perturbation), ΔF_j, is applied on residue *j* resulting in a net displacement of all C-alpha atoms from their equilibrium positions (as in A). [Source: (Gerek and Ozkan, 2011)]

function are almost always due to the low-frequency modes. Thus, the slowest modes relevant for function are used to calculate $\langle \Delta \boldsymbol{R}_i \Delta \boldsymbol{R}_j \rangle_0$ in Equation 2.26 which are sufficiently predicted by the ENM.

A recent study developed a generalized version of LRT that incorporates time-dependence (Yang *et al.*, 2014), which can model the mechanical propagation of a force throughout the protein (due to ligand binding) as a function of time. The response displacement of a residue i to a constant external force $\boldsymbol{f}_j$ on a residue j over time is

$$\langle \Delta \boldsymbol{R}_i(t) \rangle = \frac{1}{k_B T} \sum_j \left[\langle \Delta \boldsymbol{R}_i(0) \Delta \boldsymbol{R}_j(0) \rangle_0 - \langle \Delta \boldsymbol{R}_i(t) \Delta \boldsymbol{R}_j(0) \rangle_0 \right] \boldsymbol{f}_j \qquad (2.27)$$

Here, the time-dependent (td-LRT) equation is almost the same as the time-independent equation above except for the additional term $\langle \Delta \boldsymbol{R}_i(t) \Delta \boldsymbol{R}_j(0) \rangle_0$, which vanishes as $t \rightarrow \infty$ and it reduces to time-independent LRT.

As mentioned, LRT is used in conjunction with ENM in the PRS approach to elucidate conformational motions of proteins. Consider the free-body diagram of C-alpha atoms in the PRS model in Figure 2.2 where each C-alpha atom is connected by an elastic spring. In the absence of an external force (Figure 2.2A) all of the C-alpha atoms must be in equilibrium in accordance with force balance (see equations below). In the case of a 3D elastic network with N residues (C-alpha atoms) and M interactions between residues that are less than the specified cutoff distance R_c, then the force balance is

$$[\boldsymbol{B}]_{3N \times M} [\Delta \boldsymbol{f}]_{M \times 1} = 0 \tag{2.28}$$

Where $\boldsymbol{B}$ the direction cosine matrix, $\Delta \boldsymbol{f}$ contains the internal interaction forces aligned in the direction of the bond between two interacting residues (i.e., if a residue has 6 contacts as in Figure 2.2A, then $\Delta \boldsymbol{f}$ is a 6×1 column matrix).

In the presence of an external force ΔF_j as in Figure 2.2B, the nodes of the network undergo small displacements away from their original positions. Thus, force balance requires that the sum of all the interaction forces be equal to the external applied force

$$[\boldsymbol{B}]_{3N \times M} [\Delta \boldsymbol{f}]_{M \times 1} = [\Delta \boldsymbol{F}]_{3N \times 1} \tag{2.29}$$

Where the $\Delta \boldsymbol{F}$ matrix contains the components of the applied force acting on a particular residue. For a given residue i the applied force vector is

$$\Delta \boldsymbol{F}^{\mathrm{T}} = \left[000 \cdots \Delta F_x^i \Delta F_y^i \Delta F_z^i \cdots 000 \right]_{1 \times 3N} \tag{2.30}$$

Where the perturbed residue is $\Delta \boldsymbol{F}^i = \left(\Delta F_x^i \Delta F_y^i \Delta F_z^i\right)$ and all other residues entries are zero. As mentioned, the external force applied to a residue in Figure 2.2B causes a positional displacement $\Delta \boldsymbol{R}$ of each of the interacting residues from their equilibrium positions. Moreover, the distances between each connected residue $\Delta \boldsymbol{r}$ (bond distance) also change concordantly with $\Delta \boldsymbol{R}$, in which their relationship is given by

$$[\boldsymbol{B}^{\mathrm{T}}]_{M\times 3N}[\Delta \boldsymbol{R}]_{3N\times 1} = [\Delta \boldsymbol{r}]_{M\times 1} \tag{2.31}$$

Where $\boldsymbol{B}^{\mathrm{T}}$ is the transpose of $\boldsymbol{B}$.

For an elastic network that consists of nodes that are connected by springs, the internal interaction forces between nodes $\Delta \boldsymbol{f}$ are related to the bond distances $\Delta \boldsymbol{r}$ by Hooke's law as

$$[\boldsymbol{K}]_{M\times M}[\Delta \boldsymbol{r}]_{M\times 1} = [\Delta \boldsymbol{f}]_{M\times 1} \tag{2.32}$$

Where the coefficient matrix $\boldsymbol{K}$ is diagonal. Two different spring constants have been devised (Gerek and Ozkan, 2011) to account for bonded interactions, γ_b, and non-bonded interactions, γ_{nb}. For bonded interactions, $\gamma_b = 1$, whereas the non-bonded interactions between residues i and j is given by the inverse square of their separation distance R_{ij} as $\gamma_{np} = 8/R_{ij}^2$. Although the original PRS method used uniform spring constants (Atilgan and Atilgan, 2009), this can be problematic since the optimal values are different across different proteins (Hinsen *et al.*, 2000; Yang *et al.*, 2009). This modified version of PRS that delineates between bonded and non-bonded interactions was able to correctly predict the conformational changes on a benchmark set of 25 protein structures (Gerek and Ozkan, 2011).

Upon substituting Equations 2.31 and 2.32 into Equation 2.29, we obtain the external forces necessary to induce the sequential displacement of residues

$$([\boldsymbol{B}]_{3N\times M}[\boldsymbol{K}]_{M\times M}[\boldsymbol{B}^{\mathrm{T}}]_{M\times 3N})[\Delta\boldsymbol{R}]_{3N\times 1} = [\Delta\boldsymbol{F}]_{3N\times 1} \tag{2.33}$$

Where $\boldsymbol{BKB}^{\mathrm{T}}$ is equivalent to the Hessian matrix $\boldsymbol{H}$ (Atilgan *et al.*, 2001). Finally, to calculate the response of individual residues to an applied external force this expression can be rearranged as

$$[\Delta\boldsymbol{R}]_{3N\times 1} = ([\boldsymbol{B}]_{3N\times M}[\boldsymbol{K}]_{M\times M}[\boldsymbol{B}^{\mathrm{T}}]_{M\times 3N})^{-1}[\Delta\boldsymbol{F}]_{3N\times 1} \tag{2.34}$$

And substituting the Hessian for $[\boldsymbol{B}]_{3N\times M}[\boldsymbol{K}]_{M\times M}[\boldsymbol{B}^{\mathrm{T}}]_{M\times 3N}$ Equation 2.34 becomes

$$[\Delta\boldsymbol{R}]_{3N\times 1} = [\boldsymbol{H}]^{-1}_{3N\times 3N}[\Delta\boldsymbol{F}]_{3N\times 1} \tag{2.35}$$

Which describes the total response displacement of each residue of the protein due to a perturbation by an external random force.

The Hessian $\boldsymbol{H}$ is obtained from the interactions between C-alpha atoms in the coarse-grained ENM in using described in Section 2.2.2 and its inverse is evaluated by single value decomposition. Alternatively, $\boldsymbol{H}^{-1}$ can be replaced by the $3N \times 3N$ covariance matrix $\boldsymbol{G}$ as calculated by all-atom MD simulations as

$$[\Delta\boldsymbol{R}]_{3N\times 1} = [\boldsymbol{G}]_{3N\times 3N}[\Delta\boldsymbol{F}]_{3N\times 1} \tag{2.36}$$

Which as discussed is computationally intensive to obtain as compared to the coarse-grained approaches. We primarily use the coarse-grained PRS method to investigate protein dynamics for large datasets. However, the PRS method cannot calculate the difference between the wild type and mutant forms of a protein (since it is not amino acid specific). Thus, to study the changes in dynamics between the wild type and mutant forms of the GCase protein (presented in Chapter 4), we will use all-atom MD simulations to obtain the covariance matrix and use Equation 2.36 to get the dynamics.

CHAPTER 3

3 CONFORMATIONAL DYNAMICS ON PROTEIN-PROTEIN INTERATIONS INFORMS FUNCTIONAL IMPACT OF GENETIC VARIANTS

As excerpted from:

Butler, B., Gerek Z., Kumar, S., and Ozkan, S.B. "Conformational dynamics of nonsynonymous variants at protein interfaces reveals disease association," Proteins 83: 3, 428-435 (2015).

3.1 Introduction

Advances in sequencing technologies are providing a wealth of data on human genetic variation. It is now clear that any personal exome contains thousands of variants, the majority of which are non-synonymous single nucleotide variants (nsSNVs) (Kumar *et al.*, 2011). However, distinguishing between neutral variants (i.e., those with little or no effect on phenotype) from variants associated with disease still remains a major challenge for both monogenic (Mendelian) and complex diseases (S. Kumar *et al.*, 2009; Kumar *et al.*, 2011). The current state-of-the-art methods for diagnosing amino acid variants primarily employ evolutionary information obtained from multispecies sequence analysis in a variety of ways (S. Kumar *et al.*, 2009; Kumar *et al.*, 2011; P. Kumar *et al.*, 2009; Kumar *et al.*, 2012; Dudley *et al.*, 2012; Adzhubei *et al.*, 2010). While these methods have been used extensively, they often fail to correctly diagnose damaging

variants at evolutionarily variable positions and neutral variants at highly conserved positions (S. Kumar *et al.*, 2009).

Several methods have been proposed to incorporate structure-based information from protein structures. Two prominent methods are to use accessible surface area (ASA), which determines the surface area of a protein accessible to a solvent, and the change in protein stability, which utilizes the difference in free energy between the folded and unfolded state upon mutation through empirical calculation based on the 3-D structure (Cline and Karchin, 2010; Cheng *et al.*, 2008; Li *et al.*, 2011; Wei *et al.*, 2012; Yue *et al.*, 2005, 2006). Interestingly, the addition of these modalities has only produced a marginal 3-4% improvement in the rate of true positive diagnosis (Li *et al.*, 2011; Huang, Wang, *et al.*, 2010; Huang, Shi, *et al.*, 2010). A common feature among these methods is that they are based on the static 3-D structure of the protein, which fails to capture the dynamic motion of the protein structure. From the conformational transitions of allosteric proteins to the required flexibility of a ligand-binding site, proteins must fluctuate to achieve their function (Zheng *et al.*, 2007; Wang *et al.*, 2004; Velazquez-Muriel *et al.*, 2009; Tzeng and Kalodimos, 2011; Martin *et al.*, 2012; Liu *et al.*, 2010; Liberles *et al.*, 2012; Kalodimos, 2012; Jackson *et al.*, 2009; Glembo *et al.*, 2012; Eisenmesser *et al.*, 2005, 2002; Echave and Fernandez, 2009; Echave, 2008; Bhabha *et al.*, 2011; Bahar *et al.*, 2010).

A reason for the lack of methods incorporating protein dynamics into nsSNV diagnoses could be the absence of amino acid site-specific measures that can statistically quantify the contribution and impact of each position on the conformational dynamics of the protein in a fast and efficient way. We recently developed a *dynamic flexibility index*

(*dfi*), which measures the contribution of each position to functionally important dynamics (Nevin Gerek *et al.*, 2013). Through *dfi* analyses of more than 100 monomeric proteins, we found that the added feature of protein dynamics has the potential to distinguish between nsSNVs that impact biological function and those that have no effect on function (neutral nsSNVs) at a proteome scale (Nevin Gerek *et al.*, 2013). Moreover, this large-scale analysis including population variations implicated in diseases, functionally critical positions (catalytic and binding sites), and evolutionary rates of substitutions produced concordant patterns; it established that the preservation of dynamic properties of residues in a protein structure is critical for maintaining the protein/biological function (Nevin Gerek *et al.*, 2013).

The *dfi* metric has not yet been evaluated for biological assemblies. Many proteins form biological assemblies in order to perform their specific functions in the cell. Recent studies have shown that nsSNVs located at protein-protein interface sites are often associated with disease (Wei *et al.*, 2012; David *et al.*, 2012) where additional metrics beyond evolutionary information can be useful (Jordan *et al.*, 2010). Therefore, we report the *dfi* analysis for proteins that form biological assemblies and its relationship with evolutionary conservation. We also compare the difference between the *dfi* of disease-associated and neutral nsSNVs when it is calculated in biological assemblies and when it is calculated by using proteins as monomers in order to determine which is more informative at phenotypic prediction. Moreover, we compare *dfi* with the static measure of solvent accessible area, which has also been used to predict disease-associated nsSNVs in biological assemblies (Wei *et al.*, 2012).

3.2 Methods

3.2.1 Data Set

We generated a curated dataset of 1,174 protein nsSNVs using available databases, including HumVar that contains 301 disease-associated and 200 neutral population variants compiled for PolyPhen-2 (Adzhubei *et al.*, 2010), 383 neutral variants from the 1000 Genomes Project with those having population frequency greater than 10% (Abecasis *et al.*, 2012), and 290 disease-associated variants from the Human Gene Mutation Database (HGMD) (Stenson *et al.*, 2003). The set of 333 unique multimeric proteins containing 591 disease-associated and 583 neutral nsSNVs was modeled such that all the proteins formed assemblies and have 3-D structures in the Protein Data Bank (Bernstein *et al.*, 1977) with >80% sequence identity between the reference sequence and experimentally-derived protein structures and >80% sequence coverage using BLAST. The high constraints were imposed to ensure that the structures used in this study are real experimental human proteins rather than pure homology models.

3.2.2 The *dfi* Metric for Biological Assemblies

The dynamic flexibility index (*dfi*) is a metric to determine the structural flexibility at specific sites on a protein. We applied our original method (Nevin Gerek *et al.*, 2013) directly to biological assemblies (BAs) such that the dynamic flexibility for each position in the BA is considered. In brief, the method is based on the perturbation response scanning (PRS) method where the equilibrium structure of a protein is constructed as a 3-D elastic network model (ENM) in which the nodes are represented by C-alpha atoms (Tirion, 1996; Hinsen, 1998), and the pairwise potential between each atom is given by the potential of a harmonic spring. A small perturbation in form of

random Brownian kick is applied sequentially to each C-alpha atom in the elastic network. The perturbation on a single residue results in a cascade of perturbations to all other atoms in the network, thus inducing a global response. The fluctuation response profile of the positions upon perturbation of a single residue ($[\Delta\mathbf{R}]_{3N\times1}$) is obtained using linear response theory and given by the equation

$$[\Delta\mathbf{R}]_{3N\times1} = ([\mathbf{H}]_{3N\times3N})^{-1}[\Delta\mathbf{F}]_{3N\times1} \quad (3.1)$$

Where the $\Delta\mathbf{F}$ vector contains the components of the externally applied random unit force vectors ($\hat{f}$) on the selected residues, and $\mathbf{H}^{-1}$is the inverse of Hessian matrix (i.e., **H,** the Hessian, is a $3N \times 3N$ matrix composed of the second order derivatives of the harmonic potential with respect to the components of the position vectors for the chain of length *N*). To minimize the effects of randomness, this perturbation procedure is performed ten times to ensure that the applied force is isotropic with a zero angular average ($\langle\hat{f}\rangle = 0$), and the response vector $\Delta\mathbf{R}_j^{\mathrm{i}}$ is averaged.

In short, the application of the random Brownian kick to a given residue on the 3D elastic network perturbs the residue interaction network of the protein beyond fluctuations inherent in the system at equilibrium and elicits responses from all other residues in the structure. Through the perturbation response scanning method (PRS) (Gerek and Ozkan, 2011; Atilgan *et al.*, 2001), we compute the fluctuation response of residue *j*, $\Delta\mathbf{R}_j^{\mathrm{i}}$, both in direction and magnitude upon perturbation. We repeat this perturbation on each single residue for all positions in chain and obtain the response profiles of all other positions. The dynamic flexibility index, *dfi*, is then obtained by the equation

$$DFI_i = \frac{\sum_{i=1}^{N} \left|\Delta R^i\right|_j}{\sum_{j=1}^{N} \sum_{i=1}^{N} |\Delta R^i|_j} \tag{3.2}$$

Where $\left|\Delta R^i\right|_j = \sqrt{\langle \Delta R^2 \rangle}$ is the magnitude of positional displacements for residue *j* in response to a perturbation at residue *i* after averaging out the response vector $\Delta \mathbf{R}_j^i$ over ten different random directional unit forces, and *N* is the total number of positions on the biological assembly. Note that we compared the *dfi* values obtained from the coarse-grained ENM model with those obtained from all-atom replica exchange molecular dynamics simulations for several proteins in our earlier work (Nevin Gerek *et al.*, 2013) in which the *dfi* values obtained from these two different simulation approaches yield very high correlations, as Pearson correlation coefficients between PRS and all-atom MD ranged from 0.64 to 0.88 for 5 proteins.

For the monomeric analysis of biological assemblies, the *dfi* value is estimated using the monomeric unit alone (i.e., for a homomeric dimer with two units of 2*N* residues only the *N* residue position of the monomeric unit is considered). Thus, the impact of the interactions aroused due the interaction of interface residues between each unit in the BA is not considered. In estimating the *dfi* values for the BA, however, the whole complex (i.e., 2*N* residue positions of the two homomeric units) is used such that the interactions between the interface positions in the BA are explicitly included in the Hessian. Moreover, the flexibility response of residue *i* on unit 1 after perturbing residue *j* on unit 2 is computed and included in the *dfi* profile of unit 1. A workflow depicting the methodology for the *dfi* analysis of the BA and monomeric unit is provided in Figure 3.1.

Since we collectively analyze atomic positions for a wide variety of protein structures, *dfi* must be normalized. Thus, the *dfi* value of a specific atomic position in the

protein is expressed as %*dfi*, which is a percentile rank of that atom in a sorted array of all *dfi* values in a given protein. The *dfi* calculation is performed on each biological assembly, which is comprised of two or more chains. The calculation is then done on a single chain taken from the biological assembly (Figure 3.1).

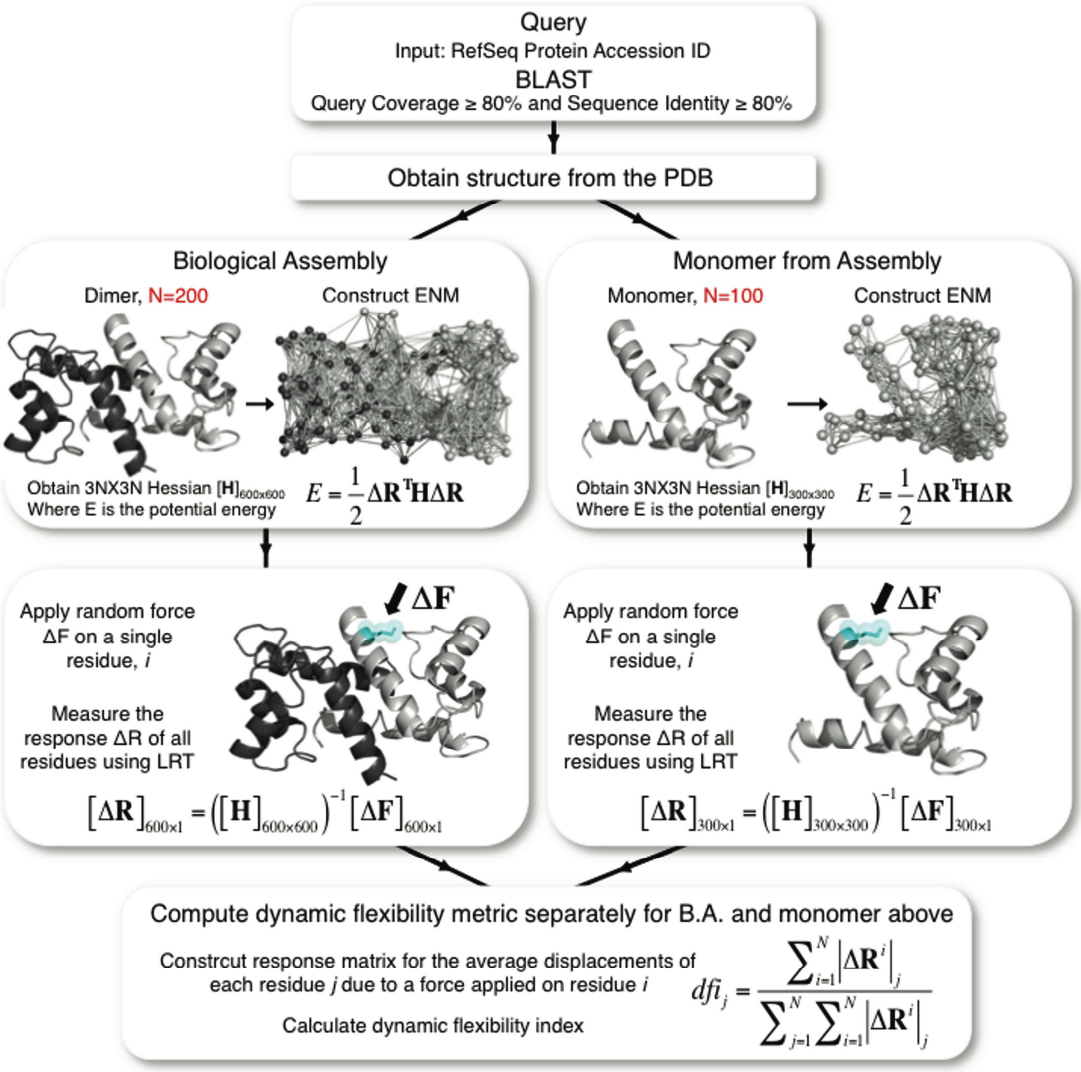

Figure 3.1: The schematic diagram of the method followed for structural dynamics analysis of each multimeric protein. We identify a three-dimensional (3-D) structure for each protein sequence through a BLAST search using protein data bank (PDB). In this search, the sequence coverage and the sequence identity between the reference sequence query and the known protein structures is set to >80% and >80%, respectively. The identified 3-D experimental structures from PDB are then used for the Perturbation Response Scanning (PRS) model to predict the dynamic flexibility index (%dfi) for each residue position.

3.2.3 Accessible Surface Area (ASA)

We compare the *dfi* metric with a static metric known as accessible surface area (ASA) and its capability to quantify phenotypes of nsSNVs. The ASA metric determines the amount of surface area in the crystal that is accessible (i.e. exposed to a solvent). We calculated ASA by using the DSSP program (Kabsch and Sander, 1983). Following the *dfi* procedure, we normalized ASA values for each residue position and expressed them as %ASA.

3.2.4 Prediction of Interface Sites

The prediction of molecular interface residues of BAs were determined using the *PISA* server (Krissinel and Henrick, 2004, 2007, 2005). *PISA* is a computational tool that predicts the strength of interaction between two monomers and the interfaces between them, resulting in the multimer that is likely the functional form of the BA.

3.2.5 Evolutionary Rates

We estimate the absolute evolutionary rate at each site by using a previously described method (S. Kumar *et al.*, 2009), which computes the number of amino acid substitutions in a given phylogeny following the parsimony algorithm for each site independently (Fitch, 1971). The evolutionary rate of amino acid changes across species is then the number of amino acid substitutions divided by the total time elapsed in the tree. Evolutionary rates are in the units of substitutions per amino acid per billion years (Byrs) and are based on protein sequence alignments of 46 species available from the University of California-Santa Cruz resource (UCSC Human Genome Browser) (Kent *et al.*, 2002).

3.3 Results and Discussion

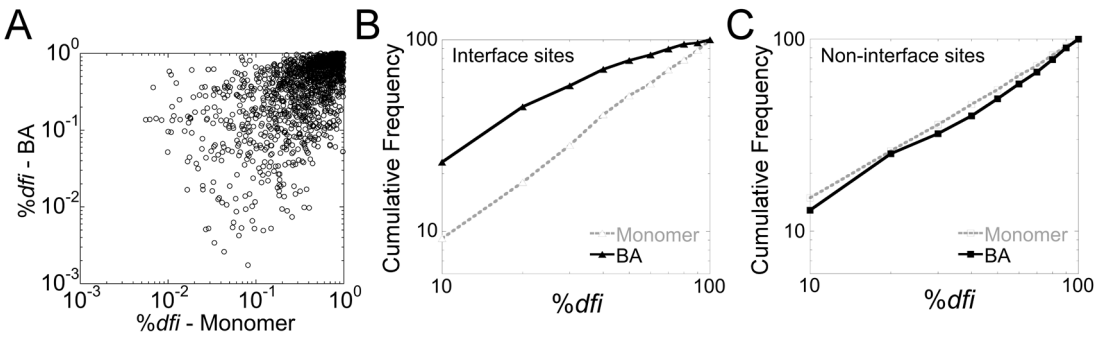

Figure 3.2: Distributions of interface and non-interfaces sites in the biological assembly proteins and their corresponding monomeric units. A scatter plot is shown in (A) of the %*dfi* values for all variants, disease-associated and neutral, using the biological assembly units (y-axis) their corresponding monomeric units (x-axis). Each axis is scaled logarithmically. Many sites exhibit low *dfi* in the BA but much higher *dfi* in their monomers, indicating that they are located at interfaces. Cumulative %*dfi* distributions of interface sites (B) and non-interface sites (C) for the BA units and their corresponding monomeric units.

To assess the effect of using biological assemblies (BAs) on the estimation of conformational dynamic parameters, we compared the *dfi* values of all 1,174 nsSNVs in 333 BAs with those obtained by using only the monomeric units. Many sites harboring sequence variants showed large differences in %*dfi* calculated from the BA and monomeric forms (Figure 3.2A). For example, many high %*dfi* sites in the monomeric calculations show rather low %*dfi* in the BA calculation. We found many of these residues to be located at interface sites in the BA, which seems reasonable since residues at interfaces exhibit a different fluctuation profile in assemblies. This is due to their interaction with the residues of another unit, unlike the monomeric forms where the same residues would interact with a solvent instead. When considering only the interface sites (357 of 1,174), we observe a large difference ($p < 0.0001$) in the cumulative %*dfi* distributions (Figure 3.2B) between the monomeric and multimeric forms with an average %*dfi* of 31% for the BA unit and 51% for the monomeric unit. The interface

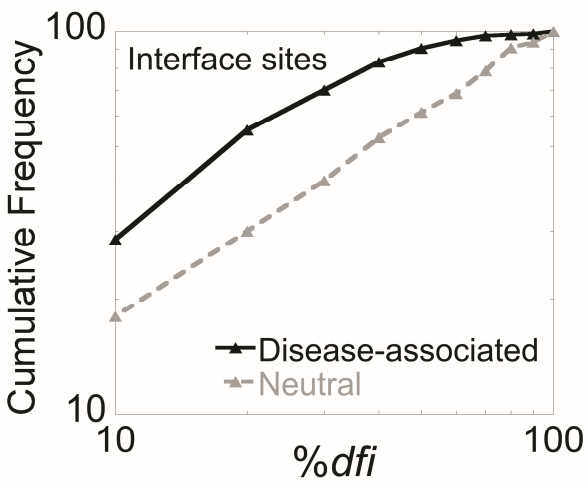

Figure 3.3: Cumulative %dfi distributions of protein interface sites for disease-associated variants (black line) and neutral variants (grey line) from the human population (compiled from HumVar and the 1000 genomes project). The average %dfi for disease-associated variants at interfaces is 23% while that for neutral variants is 42% ($p < 0.0001$).

variants had lower dynamic flexibility, with over 50% showing %*dfi* ≤ 25%. This tendency is expected since the interactions with other monomeric units in the BA lead to a decrease in flexibility. On the other hand, the cumulative %*dfi* distributions of monomeric and BA units are very similar for the nsSNVs at non-interface sites (817 of 1,174), as shown in Figure 3.2C. For these sites, the average %*dfi* for BA units was 50% and that for their monomeric units was 46%.

The above pattern prompted us to investigate whether considering the structural dynamics of the BA is more powerful in distinguishing disease-associated nsSNVs. We computed the cumulative distributions of 207 disease-associated nsSNVs from 62 proteins and 150 neutral nsSNVs from 71 proteins separately for interface sites (Figure 3.3). There is a distinct separation between the two cumulative distributions. At lower *dfi*, the separation of the two curves was pronounced, indicating that sites containing disease-associated variants have lower *dfi* than those containing neutral variants at interfaces. The

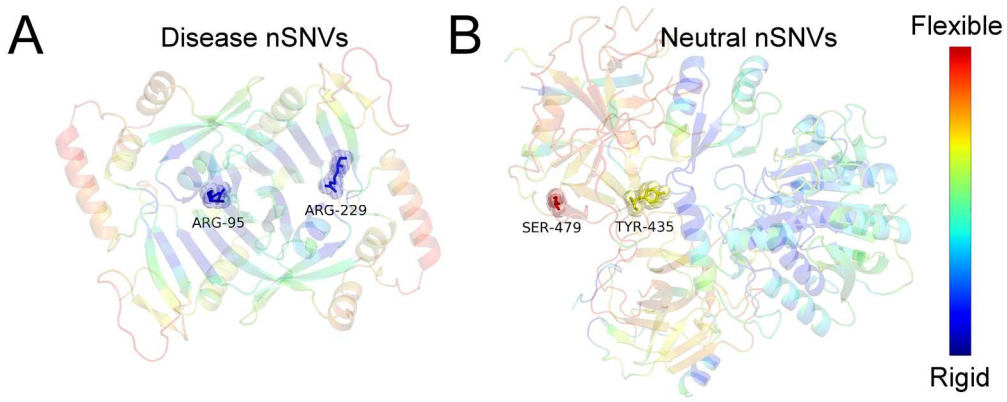

Figure 3.4: The ribbon diagrams of two proteins containing disease and neutral nSNVs given by their *dfi* profiles. (A) recombinant human pyridoxine-5'-phosphate oxidase (PDB code: 1NRG) and (B) human carboxypeptidase A1 (PDB code: 1PYT) with respect to dynamic flexibility index, %*dfi,* are shown. Each structure is colored within a spectrum of red–yellow–green–cyan–blue, where red shows the highest (flexible) and blue the lowest values (rigid) of %*dfi*. (A) Two disease-associated variants at interface sites are predicted to be rigid by *dfi*. (B) Two neutral variants at interface sites are predicted to be flexible by *dfi*. The colors of their sticks and spheres correspond to their %*dfi*.

average %*dfi* for disease-associated variants at interfaces is 23% while that for neutral variants is 42% ($p < 0.0001$).

We chose two case studies to shed light on the mechanistic differences for the analysis of individual proteins and BAs. Human pyridoxine-5'-phosphate oxidase (1NRG in the Protein Data Bank) is a homodimer that serves as an important enzyme to catalyze reactions in the vitamin B6 metabolism pathway. Two variants with known disease implications from HGMD were mapped onto the protein interface, as shown in Figure 3.4A. The structure is colored within a spectrum of red–yellow–green–cyan–blue, where red shows the highest and blue the lowest values of %*dfi*. Based on Figure 3.4A, it is clear that these two variants located at the interface have low dynamic flexibility (ARG-95 and ARG-229 have a %*dfi* of 0.07981 and 0.15962 respectively). With such low *dfi* values those sites are likely critical for function, thus a mutation there will likely lead to a

disruption in function. For instance, the site ARG-229 is mutated to TRP-229, which results in the potentially fatal disease, neonatal epileptic encephalopathy (NEE) (Stenson *et al.*, 2003; Mills *et al.*, 2005). For the second case, three neutral variants from the 1000 Genomes Project were mapped to the model structure of human carboxypeptidase A1 (homologous structure is 1PYT in the Protein Data Bank) with TYR-435 occurring at an interface site and the other two at non-interface sites (Figure 3.4B). From Figure 3.4B, it can be seen that these sites have noticeably higher dynamic flexibility. Interestingly, even TYR-435 had a high *dfi* score of 0.62084 despite its location at an interface. It is expected that interface sites generally have lower *dfi* values since they are interacting with residues of another protein, thus high *dfi* at an interface is surprising and could lend useful information relating to the phenotype. Fig. 3.4 shows how variants within an individual protein could lead to the general trend seen in Figure 3.3, which is based on the analysis of more than 100 proteins. Moreover, the trend exhibited in Figure 3.3 and the case study presented in Figure 3.4 together gives further indication to the notion that *dfi* may discriminate disease-associated from neutral variants.

For comparison, we also examined the performance of ASA, a metric based on the static form of the protein structure, which has also been utilized to differentiate disease-associated nsSNVs from neutral variants (Jordan *et al.*, 2010; Franzosa and Xia, 2009; Wei *et al.*, 2012). We found that the average %ASA showed only a small difference (45% for disease-associated and 66% for neutral population variants), as compared to a 2.5 times difference observed for average %*dfi* (21% for disease associated and 54% for neutral population variants). We found that there is a correlation between ASA and *dfi*, as sites with low ASA that are surrounded with other residues rather than

solvent would exhibit fewer fluctuations and cause lower *dfi* values. However, among these low ASA positions, certain positions can be more dynamically critical in translating or controlling the functionally related motion than others due to their residue interaction pattern within the protein structure. By utilizing *dfi*, we are able to capture these dynamically critical positions. Thus, the above result suggests that the interface residues that play an important role in the collective motion of the BA are more susceptible to damaging mutations.

We examine whether the predictive capabilities of *dfi* for the BA go beyond that afforded by evolutionary conservation of positions involved by estimating the evolutionary rate (r) for each nsSNV site (as described in the methods section). We divided the estimated evolutionary rate (r) into two different categories: ultra-conserved ($r = 0$) or less-conserved ($r > 0$). In our analysis, 37% of interface sites and 30% non-interface sites were ultra-conserved sites. Likewise, 63% of interface sites and 70% of non-interface sites were less-conserved sites. This difference in evolutionary rates is rather small, as compared to conformational dynamics where a higher fraction of interface sites have very low *dfi* (53% of interface sites and 29% non-interface have *dfi* ≤ 25%). This prompted us to consider the phenotypic prediction of nsSNVs at interface sites, as the ability to correctly identify disease associated variation at less-conserved sites is not high for many evolutionary rate based *in silico* prediction tools (S. Kumar *et al.*, 2009; Kumar *et al.*, 2011) and many interface sites are at less conserved positions. We surmised that *dfi* calculated using BAs might provide information beyond that afforded by evolutionary conservation at those sites. Thus, we explored the ability of *dfi* to discriminate disease-associated and neutral nsSNVs at less-conserved sites ($r > 0$).

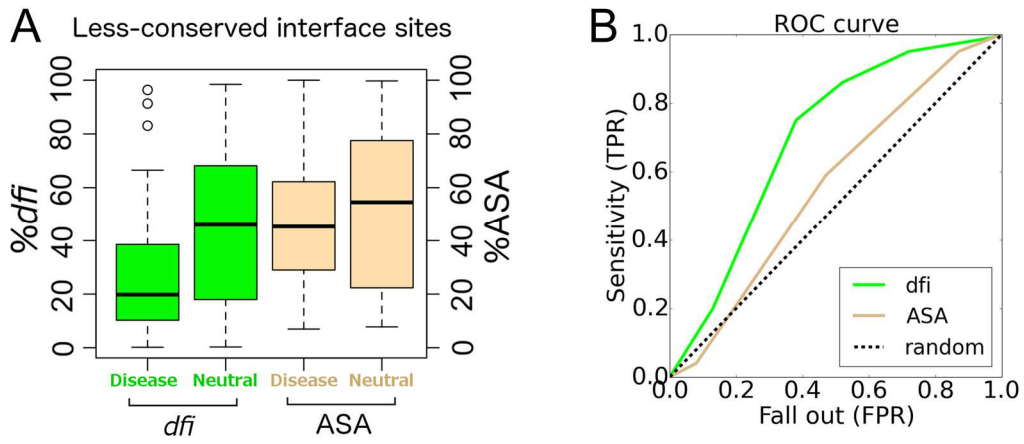

Figure 3.5: A box plot of %*dfi* (green) and %ASA (brown) distributions comparing disease-associated and neutral nsSNVs for less-conserved variants (evolutionary rate $r > 0$) occurring at protein interfaces. Box plots show median, upper, and lower quartiles, and whiskers represent maximum and minimum values. (B) A receiver operating characteristics (ROC) curve for *dfi* and ASA using a test set that was generated from 10% of the whole data set. The area under the curve (AUC) for *dfi* and ASA was 0.71 and 0.56 respectively. TPR and FPR are true and false positive rates in predicting disease associated nsSNVs to be identified as non-neutral, respectively.

We compared box plots of %*dfi* and %ASA for disease-associated and neutral variants at interface sites that were less-conserved (Figure 3.5A). Remarkably, the average %*dfi* of disease-associated nsSNVs is approximately 2.5 times lower than that of neutral nsSNVs gathered from human population statistics (Adzhubei *et al.*, 2010). The average %*dfi* for disease-associated variants was 25% at less-conserved sites at interfaces, whereas the average %*dfi* for neutral variants from the 1000 Genomes Project and HumVar was 45% ($p < 0.001$ when comparing both datasets). This suggests that *dfi* is likely a useful metric for predicting phenotypes of nsSNVs at less-conserved sites. In comparison, we did not see a suggestive difference in ASA between neutral and disease-associated variants, as the average %ASA for disease-associated sites was 47% at less-conserved interface sites, whereas the average %ASA for neutral sites was 52% ($p = 0.63$ for disease vs. 1000 Genomes Project and HumVar). We then conducted a receiver

operating characteristics (ROC) curve analysis for %*dfi* and %ASA to elucidate their ability to distinguish between disease and neutral phenotypes of nsSNVs. A randomly generated test set consisting of 10% of the entire data set (which only includes nsSNVs at interfaces) was used and the remaining 90% was used for training (Stone, 2014; Kumar *et al.*, 2012). The area under the curve (AUC) for *dfi* is 0.71 and 0.56 for ASA (Figure 3.5B). Therefore, the use of *dfi* appears to be advantageous for use in future diagnostic methods.

3.4 Conclusion

This work has provided evidence that non-synonymous variants observed at protein interface sites with low *dfi* are more likely to be disease-associated. This may be due to the fact that protein interface sites with low *dfi* play a critical role in modulating the functionally important inter-dynamics of biological assemblies. Indeed, evolutionary based metrics as well as proteins' static structure based metrics such as ASA have unique strengths in predicting the phenotypic impact, thus incorporating metrics based on structural dynamics (such as *dfi*) along with other metrics may increase the prediction accuracy of phenotypes of interface nsSNVs.

CHAPTER 4

4 DYNAMICS AND ALLOSTERY IN GENETIC ANALYSIS OF ENZYMES

As excerpted from:

Butler, B., Kumar, A., and Ozkan, S.B. "DARC spots: dynamic allosteric residue coupling reveals disease mechanism for Gaucher disease and nSNVs across the proteome," Nature Communications, Submitted.

4.1 Introduction

Advancements in genome sequencing have led to an exponential growth in the number of known non-synonymous single nucleotide variants (nSNVs). Each individual genome contains millions of variants, many of which are rare nSNVs (Kumar *et al.*, 2011). Further, variations in exomes (protein coding regions) have been associated with more than a thousand major diseases (Kumar, Butler, *et al.*, 2015; Nussinov and Tsai, 2013; S. Kumar *et al.*, 2009; Kumar *et al.*, 2011). The grand challenge has been transforming exome variation data into biomedically relevant information that informs actionable treatment recommendations. The amount of nSNV data found through genome-wide association studies, whole-genome sequencing, and exome sequencing has led to a vast array of computational techniques that leverage evolutionary, biophysical, structural, and dynamics information to assess the impact of nSNVs on protein function and, thus, phenotypic expression. However, despite these prodigious efforts, explicitly defining the relationships between disease and nSNVs, estimating the level of risk of a

given individual, and elucidating the underlying molecular mechanisms remains a hurdle yet to be surmounted (Kumar, Butler, *et al.*, 2015; Nussinov and Tsai, 2013; S. Kumar *et al.*, 2009; Kumar *et al.*, 2011).

Shedding light on the molecular mechanisms of missense nSNVs is particularly crucial; the manner in which missense variants impact protein function enables us to correctly identify a causal relationship between specific variations and disease contributions. Moreover, a better understanding of these mechanisms can potentially provide novel therapeutic strategies. It is known that disease-associated variants alter the stability of a protein (Guerois *et al.*, 2002; Alber, 1989; Yue *et al.*, 2005). Conversely, a recent study based on high-throughput functional assays of over 2000 variants revealed that only one-third of mutations led to a decrease in protein stability (Sahni *et al.*, 2015). Rather than affecting stability, disrupting binding or both, a large fraction of disease-associated variants impair protein-ligand function or enzymatic activity (Butler *et al.*, 2015; Kumar, Butler, *et al.*, 2015; Wang *et al.*, 2012). Additionally, disease-associated variants are not always located at highly conserved (i.e., functionally critical) positions. Since current state-of-the-art methods focus or even rely on assessing these conserved positions, they often fail to accurately diagnose variants at non-conserved positions (S. Kumar *et al.*, 2009). To confound the problem, studies that combine evolutionary approaches with biochemistry for protein design have also revealed disease-causing mutations at non-conserved sites can involve very complex and poorly understood mechanisms. The basic evolutionary principle that biochemically similar substitutions on non-conserved sites do not alter function does not necessarily hold. On the contrary, regardless of biochemical similarity, amino acid substitutions at non-conserved sites lead

to a wide-range of outcomes, increasing or decreasing functional activity at up to three orders of magnitude (i.e., rheostatic pattern of change) (Swint-Kruse, 2016).

Conformational dynamics of proteins are essential to explain the complexity of functional responses upon an amino acid substitution. In proteins, all positions are dynamically linked to each other within a network, where the strength of each link varies across the protein. These intrinsic dynamics are structure-encoded and govern protein function (Haliloglu and Bahar, 2015). The obsolete view of the single native structure has been long replaced by "an ensemble of substates" that accurately represent the native state (Tawfik and Tokuriki, 2009). In the ensemble model, a protein samples a variety of conformations through local changes such as loop motions, side-chain rotations, or global changes through domain rearrangement. Allostery, commonly known as regulation at a distance, is a widely used emergent property of this ensemble view. Rather than forming a new structure, a ligand binding to a remote site promotes a shift in dynamics, changing the intrinsic structure-encoded dynamics and dynamic linking (i.e., distribution of accessible conformational states in the ensemble), promoting easy access to certain conformers for allosteric regulations (Nussinov and Tsai, 2013; Guo and Zhou, 2016; Woldeyes *et al.*, 2014). Furthermore, the ensemble view also agrees with the evolutionary adaptability of a protein in which the same conserved 3D native fold can adopt new functions (Haliloglu and Bahar, 2015). Mutations throughout protein evolution alter conformational dynamics, shifting the distribution of the ensemble and lead to the emergence of new functions (Zou *et al.*, 2015; Kim *et al.*, 2015; Glembo *et al.*, 2012; Bhabha *et al.*, 2013; Campbell *et al.*, 2016) and adaption to different environments (Villy Isaksen *et al.*, 2016).

The importance of protein structure-encoded dynamics in allostery, evolution, and disease, prompted the recently developed position-specific metric, *dynamic flexibility index* (DFI) that can statistically measure the functional contribution and impact of each amino acid position on structural dynamics. DFI quantifies the resilience of a given position to the perturbations that occur at different parts of the protein using linear response theory, capturing the multi-dimensional effects when the protein structure is displaced out of equilibrium and identifies flexible and rigid positions in the structure (Nevin Gerek *et al.*, 2013; Butler *et al.*, 2015). DFI can be considered a measure of the local conformational entropy of a given position within the set of interactions governed by the 3D fold of the protein due to its ability to probe the conformational space of a protein at the residue level. A DFI analysis on the evolution of different protein families, including green fluorescence proteins (GFP) (Kim *et al.*, 2015) and beta-lactamase inhibitors (Zou *et al.*, 2015), has revealed that mutations of conserved regions observed during evolution alter the local flexibility/rigidity of different parts of the structure; this leads to changes in structure-encoded dynamics and the emergence of new biological functions. A proteome-wide conformational dynamics analysis of over 100 human proteins showed strong correlations between DFI profiles and corresponding evolutionary rates of individual positions (Nevin Gerek *et al.*, 2013). Another analysis of DFI profiles of the wild type light chain subunit of the human ferritin protein along with its neutral and disease forms revealed that neutral variants exhibit similar DFI profiles to the wild type, in which experimentally-determined critical functional sites act as hinges (i.e., sites with low flexibility) for controlling global motions. However, the disease mutations

caused these hinges to become loose (i.e., increased flexibility), impairing the structural dynamics and function of the protein (Kumar, Glembo, *et al.*, 2015).

The present study focuses on disease nSNVs occurring in human enzymes that are commonly misdiagnosed by machine learning approaches based on evolutionary principles. The misdiagnosed nSNVs are usually at non-conserved sites, distal to catalytic sites, yet they impair enzymatic function and lead to disease phenotypes. Our earlier studies suggested that conformational dynamics might provide insights for these cases (Kumar, Glembo, *et al.*, 2015; Butler *et al.*, 2015). We investigated the dynamic allosteric coupling of disease sites with the catalytic sites using our *dynamic coupling index* (DCI) metric. DCI can identify dynamic allosteric residue coupling sites (DARC spots), which are strongly coupled to active sites that are critical for function. A mutation at a DARC spot likely influences the conformational dynamics and allosteric regulation and thus, is highly susceptible to disease phenotype.

One of the signature human enzymes with over 200 disease-associated nSNVs is β-Glucocerebrosidase (GCase). Missense mutations in the GCase lead to Gaucher disease (GD) (Lieberman, 2011). GD is a human catabolic disorder that leads to a buildup of its substrate Glucocerebroside (GlcCer). The buildup of GlcCer leads to "Gaucher cells" which result in enlarged organs, splenomegaly, hepatomegaly and, in severe cases, central nervous system disorders (Hruska *et al.*, 2008). GD was first discovered by Philippe Gaucher in 1882 while treating a woman with an enlarged spleen, and is highly prevalent in the Ashkenazi Jewish population (Beutler *et al.*, 1993; Hruska *et al.*, 2008). Most of the mutations are far from functional catalytic sites but still impact enzymatic rates (Lieberman, 2011). We focused on four specific disease mutations—H255Q (Stone

et al., 2000), M123V (Finn *et al.*, 2000), V375L (Finn *et al.*, 2000) and N370S (Lieberman *et al.*, 2007)—where the common prediction servers (e.g., PolyPhen-2 and SIFT) misdiagnosed them as benign. These disease mutations are shown to be expressible in a stable, folded protein, yet the catalytic activity of the protein is significantly reduced. Previous GD studies have catalogued mutations that have genotypic and phenotypic correlations (Beutler *et al.*, 2005; Hruska *et al.*, 2008). However, few studies investigated the mechanistic impact of these mutations on conformational dynamics and allosteric regulation (Lieberman *et al.*, 2007).

Here we report a case study of GCase, which revealed that GD mutations disrupt allosteric regulation due to changes in dynamic flexibility around the catalytic sites, thereby impacting the function of the enzyme. Thus, the disease mutation sites manifest as key dynamic allosteric coupling sites (i.e., DARC spots). The results indicate that DFI can identify sites in GCase where mutations have a severe functional impact. To further study the role of dynamics and allostery in missense variants, we conducted a proteome-wide DFI analysis on a set of enzymes showed that DFI is a robust predictor of the impact of nSNVs and is complementary to established evolutionary metrics.

4.2 Results

4.2.1 Missense Variants of GCase

Over a century of research on Gaucher Disease (GD) has provided an extensive catalogue of missense mutations of GCase that lead to the disruption of the enzymatic function of the protein. GCase is a member of the family of glycoside hydrolases that uses catalytic glutamates for general acid/base hydrolysis. Specifically, GCase hydrolyzes its primary substrate, glucocerebrosidase, into glucose and ceramide, and its

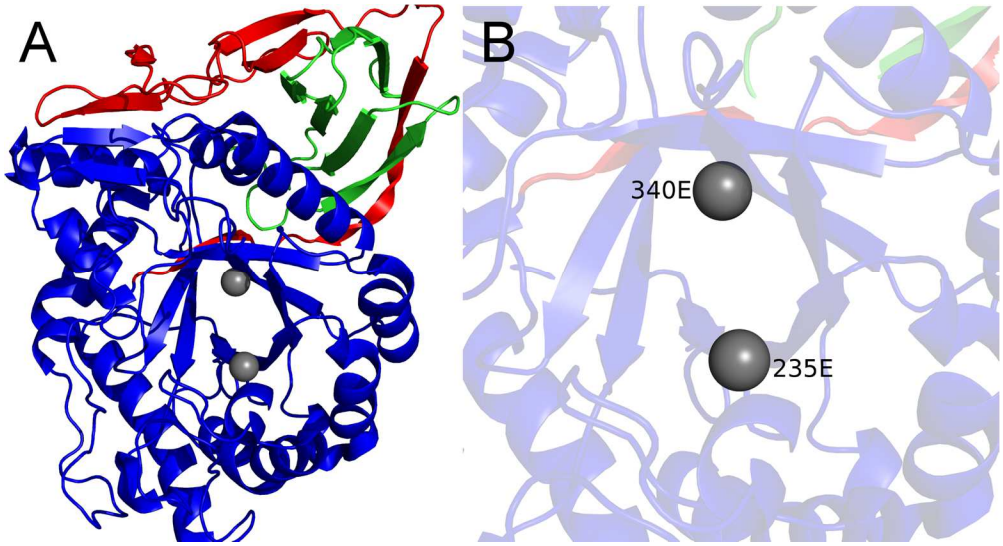

Figure 4.1: Structure and active region of GCase (A) Ribbon diagram of GCase (PDB code: 1OGS), an enzyme whose malfunction leads to Gaucher Disease (GD). GCase is a member of the family of glycoside hydrolases that use glutamates for hydrolyzing glucocerebrosidase into glucose and ceramide. It is comprised of three domains. The first domain (red) is an anti-parallel β□sheet; the second domain (blue) is a TIM barrel that contains the active site. The third domain (green) is a β□barrel. (B) A close-up view of the active region contained in the second domain including the two catalytic sites; the catalytic sites Glu235 and Glu340 are shown as grey spheres.

secondary substrate glucosphingosine, into glucose and sphingosine. GCase (PDB code: 1OGS) is a 497-residue protein that consists of three domains. As Figure 4.1 shows, the first domain (residues 1-85) is an anti-parallel β-sheet (red); the second domain (residues 86-424) is a TIM barrel that contains the active site (blue); the third domain (residues 425-497) is a β-barrel (green). The active site consists of two residues, Glu235 and Glu340 obtained from the Catalytic Site Atlas (Porter *et al.*, 2004). Residues that line the glucose-binding region are Arg120, Asp127, Phe128, Trp179, Asn234, Tyr244, Phe246, Try313, Cys342, Ser345, Trp381, Asn396, Phe397, and Val398. The aromatic residues are important for ligand recognition and the polar residues stabilize the substrate through the formation of hydrogen bonds (Henrissat, 1991).

Despite the identification of over 200 GD-associated missense mutations, there is little mechanistic insight as to how these mutations disrupt the function of GCase (Beutler *et al.*, 2005; Sidransky, 2004; Grabowski, 2004; Germain, 2004; Lieberman *et al.*, 2007). Determining the mechanism of GD is confounded by the anomaly that GCase with a GD-associated mutation is expressed as a stable protein, yet the catalytic activity is reduced. We address this by investigating the role of conformational dynamics in GCase by using the *dynamic flexibility index* (DFI). The DFI method utilizes the perturbation response scanning method (PRS) (Nevin Gerek *et al.*, 2013) that couples linear response theory (Ikeguchi *et al.*, 2005) along with the covariance matrix of the C-alpha atoms obtained from molecular dynamics simulations or an elastic network model. The method consists of applying a random Brownian kick as a mechanical perturbation to a single residue, and then computing the fluctuation response profile of all other residues in the network to this perturbation. Repeating this random perturbation sequentially for each site, we are able to compute the normalized response profile (i.e., DFI score) for every residue in the protein. The residues with low DFI indicate dynamic stability; they can absorb and transfer a perturbation throughout the chain in a cascade fashion. Low DFI positions will often be the hinge parts of the protein that control critical functional motions, similar to joints in a skeleton (i.e., the motion of a forearm is only possible by the elbow acting as a hinge). Conversely, sites with high DFI are more susceptible to perturbations in the amino acid chain. They are structurally flexible sites and important for biochemical function (e.g., anchoring sites during binding or signaling). DFI can, therefore, elucidate the mechanism of missense mutations by measuring changes in conformational dynamics and dynamic allosteric coupling.

The mutant forms of GCase include a neutral mutation, Q169R, and four disease mutations that are shown to produce structurally stable proteins (Beutler *et al.*, 2005; Sidransky, 2004). The four disease-associated mutations, H255Q (Stone *et al.*, 2000), M123V (Finn *et al.*, 2000),V375L (Finn *et al.*, 2000) and N370S (Lieberman *et al.*, 2007). M123V, V375L and N370S, lead to mild type I GD, which is the most frequent non-neuropathic form of the disease (Sidransky, 2004); H255G leads to severe type II GD which results in traumatic and rapid neurological devastation (Sidransky, 2004). The neutral mutation, Q169R, and the disease mutations—H255Q, M123V, V375L—were found in the HumVar database (Adzhubei *et al.*, 2010). These disease mutations provided an excellent test set, as evolution-based servers have misdiagnosed them as neutral. The mutant forms were modeled using the mutagenesis wizard in PyMol (Schrodinger, 2010) and the wild type crystal structure was obtained from the Protein Data Bank (PDB code: 1OGS). The N370S mutation was chosen for two reasons. First, this mutation is commonly found in approximately 70% of the Ashkenazi Jewish population and has been extensively studied (Lieberman *et al.*, 2007). Second, it is not necessary to model this mutant structure like the other variants since the crystal structure of the N370S mutant is available (PDB code: 3KE0). This allows us to compare the native equilibrium dynamics of N370S using the crystal structure as the initial structure for the simulation with the modeled mutant structures of the other mutations to verify the efficacy of the models.

The spatial distance of each disease mutation to a catalytic site was determined by computing the distances between their respective C-alpha atoms. All four mutations were found to be located remotely from the active site of GCase with distances ranging from

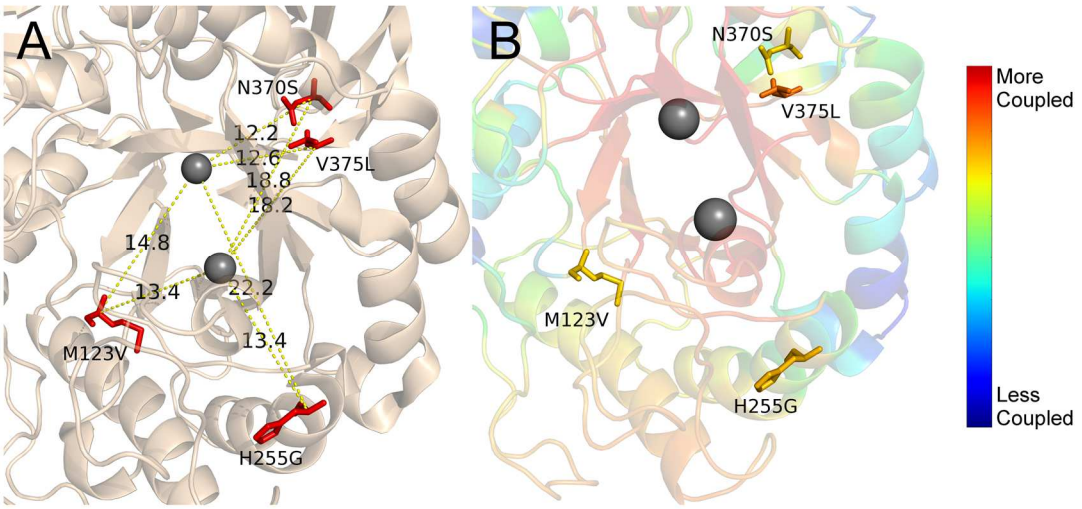

Figure 4.2: Disease mutations of GCase. (A) The positions of the four disease mutations (red sticks)—M123V, H255G, V375L, and N370S—are mapped onto the crystal structure of GCase (PDB code: 1OGS). The mutations M123V, V375L, and N370S are known to lead to mild type I Gaucher Disease while H255G leads to severe type II Gaucher Disease. The catalytic sites are shown as gray spheres. The distance between the C-alpha atom of each mutation site and the catalytic sites are labeled with a yellow dashed line. The large distances, ranging from 12.2–22.2Å, indicate that mutation sites are not in direct interaction (i.e., van der Walls interaction) with the catalytic sites. (B) A ribbon diagram of GCase is colored according to its DCI profile. The DCI metric measures the dynamic allosteric residue coupling to functional residues, which here are the catalytic sites (grey spheres). Positions in red are highly coupled to the catalytic sites, whereas positions in blue are weakly coupled to the catalytic sites. The disease mutations (yellow, orange, and red sticks) are strongly coupled to the active site.

12.2Å to 22.2Å (Figure 4.2A). This rules out the possibility that direct interactions with the active site disrupted the function upon mutation.

4.2.2 Disease Mutations Alter Conformational Dynamics

The decreased enzymatic activity upon mutation suggests that the mutation sites are allosterically coupled to the catalytic sites (Glu235 and Glu340). To verify this, we computed the *dynamic coupling index* (DCI) scores for each site, which is a site-specific metric that quantifies the degree of dynamic allosteric residue coupling with specified

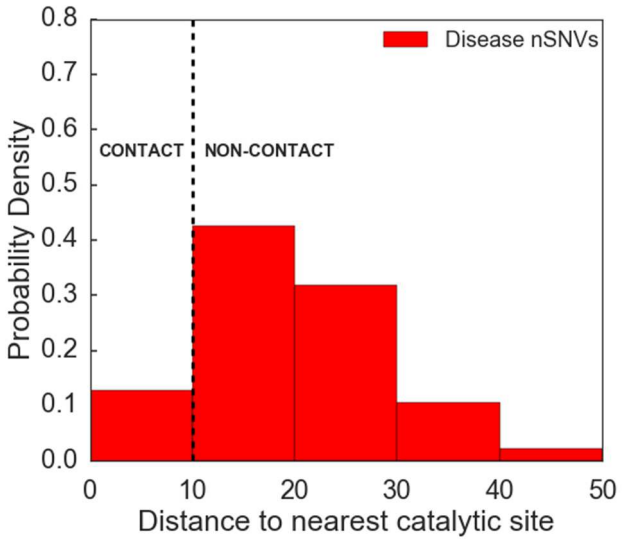

Figure 4.3: The probability distribution of the minimum distances to nearest catalytic sites (r_{min}) of the disease sites exhibiting %DCI >0.5 (i.e., highly coupled). The minimum distance, r_{min}, is the distance of a given nSNV site to the closest catalytic based on their relative C-alpha positions. The distribution of disease nSNVs shows that over 80% are distally located to catalytic sites, which indicates that they are dynamically coupled (DARC spots) important for allosteric regulation.

sites that are critical for function (e.g., active sites such as catalytic residues). Essentially, it measures the change in the response fluctuation profile of a site upon perturbing only functional sites as compared to the average response profile of perturbing all sites. Sites distal to the active sites that exhibit high DCI values are DARC spots (i.e., dynamic allosteric residue coupling spots), and a mutation there would likely also impact the active sites. Figure 4.2B presents a ribbon diagram of GCase colored according to its DCI profile from a spectrum of red-orange-yellow-green-cyan-blue, where red indicates sites that are strongly coupled to the catalytic sites (high DCI), whereas sites in blue are weakly coupled (low DCI). The disease mutations exhibit high DCI (yellow/orange), indicating they are dynamically linked to the catalytic sites. In addition to these four disease mutations, we also computed the distance of 84 other missense GD mutations

(categorized as severe impact (Beutler *et al.*, 2005)) from the catalytic sites and compared this to their respective DCI scores. Approximately 80% of the disease mutations that exhibited high DCI (>0.5) were not in direct contact with the catalytic sites (See Figure 4.3). This highlights the role of dynamic allosteric residue coupling between mutation sites and catalytic sites, which suggests that mutations at DARC spots alter conformational dynamics at catalytic sites, leading to suppression of enzymatic activity in GCase (Sinha and Nussinov, 2001).

To further elucidate the change in conformational dynamics upon mutation, we obtained DFI profiles of each mutant form and compared them with that of the wild type. The average DFI profile of the disease mutants differs from the wild type, particularly in the first and second domains (TIM barrel) where the catalytic site exhibits lower DFI values at specific locations (Figure 4.4A). Our earlier analysis on ancestral proteins and proteome-wide analysis of missense variants suggests that the distribution of rigid and flexible parts of the proteins and their communication through dynamic allosteric coupling channels underlies the function of the protein (Glembo *et al.*, 2012; Zou *et al.*, 2015; Kumar, Butler, *et al.*, 2015). The change in the average DFI profile of disease mutants compared to the wild type agrees with our earlier findings that the selected mutations allosterically affect conformational dynamics (Kumar, Glembo, *et al.*, 2015). Specifically, mutations not only affect conformational dynamics of sites within the vicinity of mutation, but also affect the dynamics of distal sites through dynamic allosteric coupling.

Figure 4.4A shows the average DFI profile of the four disease mutants (red) as compared to the wild type (black). There is an appreciable decrease in the DFI profile of

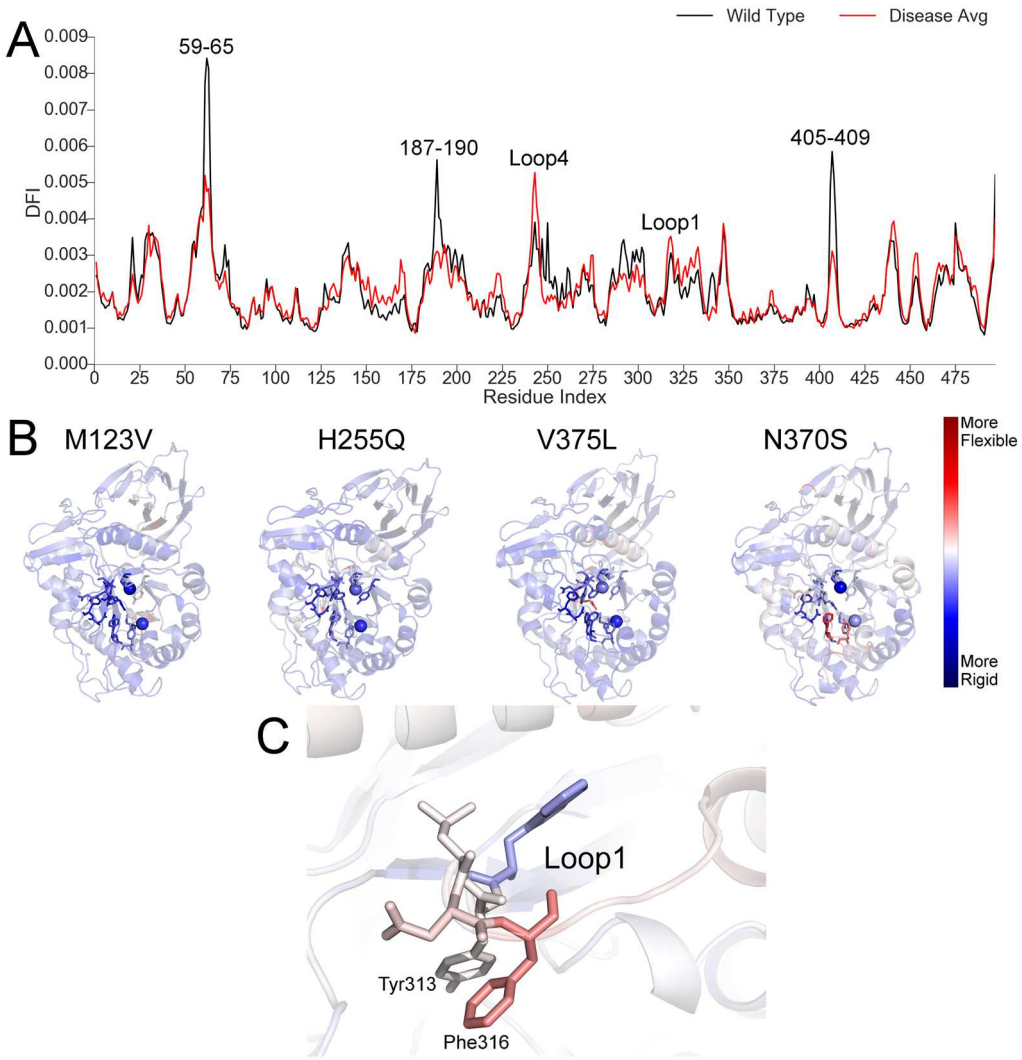

Figure 4.4: DFI profile of disease mutants. (A) The DFI profile of the wild type (black) and average DFI of disease mutants (red). Residues exhibiting high DFI are the most flexible parts of the protein, whereas residues with low DFI are the most rigid parts of the protein. The average disease profile is significantly different than the wild type indicating that the disease variants rigidified at specific regions of the protein. (B) The color-coded ribbon diagrams of GCase using □DFI ($[DFI_{disease} - DFI_{wt}] / DFI_{wt}$) profiles for each disease mutant, where red indicates an increase in the flexibility of the residue upon mutation; blue indicates a decrease in the flexibility. The catalytic residues at the active site are shown as spheres. The residues represented as sticks are key sites in binding recognition and ligand stability. Overall, all disease mutants show a rigidification around the active site suggesting alteration of ligand binding rates as a plausible mechanism. (C) The color-coded ribbon diagram of loop1 based on □DFI for the disease mutant N370S, which shows an increase in flexibility of loop1 (red).

mutants compared to wild type in three regions. Interestingly, the first region, residues 59-65, includes the glycosylation site N60 (Figure 4.4A) which is known to be critical for enzymatic function (Pol-Fachin *et al.*, 2016). It follows that a drastic decrease in the flexibility of this position as exhibited in disease mutants may interfere with glycosylation. The functional role of the other two regions, 187-190 and 405-409 (Figure 4.4A) are not known. However, the mutation D409H is a unique case where a an intracellular change in GCase activity generates a phenotypic response, including specific cardiovascular symptoms (Pasmanik-Chor *et al.*, 1996). Likewise, region 187-190, includes position N188 where the atypical mutation N188S is associated with myoclonic epilepsy (Kowarz *et al.*, 2005) due to decrease of functional activity in GCase (Tajima *et al.*, 2010).

We further quantified the change in flexibility per position upon a disease mutation by computing the fractional change in DFI as ΔDFI ($[DFI_{disease} - DFI_{wt}] / DFI_{wt}$) with particular interest in positions identified as binding and recognition sites near catalytic sites. In Figure 4.4B ribbon diagrams of the mutant structures are color-coded according to their ΔDFI profiles for each disease mutant. Regions that exhibit a large decrease in DFI as compared to the wild type become more rigid upon mutation (blue). Interestingly, Figure 4.4B shows that each mutant structure shows a consistent trend where disease mutations lead to the rigidification of the two catalytic sites. Moreover, most of the positions critical for ligand binding and recognition (shown as sticks in Figure 4.4B) exhibit a large decrease in DFI, suggesting that the decreased flexibility may impair catalytic turnover rates in enzymatic function.

Among the 5 loops surrounding the active site, we observe that loop1 (312-317) and loop4 (237-248) exhibit an increase in DFI (Figure 4.4A), suggesting that increased flexibility of these loops could contribute to the decrease in enzymatic activity by hindering the accessibility of the ligand to the active site as observed in previous work (Li *et al.*, 2015). In accordance with work of Liebermen et al (Lieberman *et al.*, 2007), the N370S mutant shows the most drastic increase in DFI of loop1 (Figure 4.4C). Based on crystal structures of IFG bound N370S, unbound N370S and wild type, it was proposed that binding of IFG to loop1 of the N370S mutant leads to increased enzymatic efficiency and trafficking due to stabilization of loop1, locking GCase into a substrate-bound conformation (Lieberman *et al.*, 2007). The increase flexibility of loop1 in N370S allows conformations where Tyr313 would hydrogen bind to Glu235, hindering the accessibility of substrate to the active site (Lieberman *et al.*, 2007). Indeed, Figure 4.4C shows a notable increase in flexibility of Tyr313 and Phe316 (i.e., the positions where IGF binds and stabilizes).

In summary, the change in DFI profiles of disease mutants provides an explanation as to why the mutant forms are expressed as stable proteins yet have decreased enzymatic activity. The disease mutations lead to a protein where certain regions (binding and recognition sites) become highly rigid while other regions (loop1 and loop4) exhibit enhanced flexibility. Importantly, the flexibility was restricted to only these localized changes, as a global increase in flexibility would likely lead to a drastic destabilization of the enzyme. Therefore, the disease mutants are still stable enough to be expressed. In addition to increased flexibility of loop1 and loop4, we also observed decreased flexibility of ligand recognition sites in disease mutants as compared to the

wild type, which suggests another possible mechanism: when the orientations of recognition and binding sites become restricted, they lose the required flexibility necessary to accommodate the ligand binding event. This may lead to decreased catalytic turnover rates of the enzyme and obstruct enzymatic function. Furthermore, these results show an allosteric mechanism in which mutation sites that are not in direct contact (i.e., Van der Waals interactions) with the critical functional sites could still alter their conformational dynamics.

4.2.3 Loss of Dynamic Allosteric Coupling in Missense Variants

In general, GCase is a large protein with three domains, where dynamic allosteric coupling between residues in different domains, beyond the TIM Barrel active region, likely plays a role in overall function for this enzyme. Indeed it has been shown that interaction with Saposin C (SapC) is critical for GCase activity by remodeling the lipid membrane, presumably by helping GCase access the short head group of the lipid bilayers (Qin, 1996). Thus, dynamic allosteric coupling between the catalytic domain and the other two domains of GCase should play a key role in regulation of enzymatic function. We further investigated how dynamic coupling of each position to the two catalytic sites (E235 and E340) changes upon disease mutation using the dynamic coupling index (DCI). Here, we use DCI to determine whether or not disease sites are DARC spots, and if mutations at DARC spots impair the long-range dynamic coupling of catalytic sites to the rest of the chain. Interestingly, the DCI profiles of all four disease mutants showed a drastic decrease in dynamic allosteric residue coupling as shown in Figure 4.5A. It is remarkable that all disease mutants lead to a global loss of dynamic coupling with the two active site positions due to changes in dynamic flexibility,

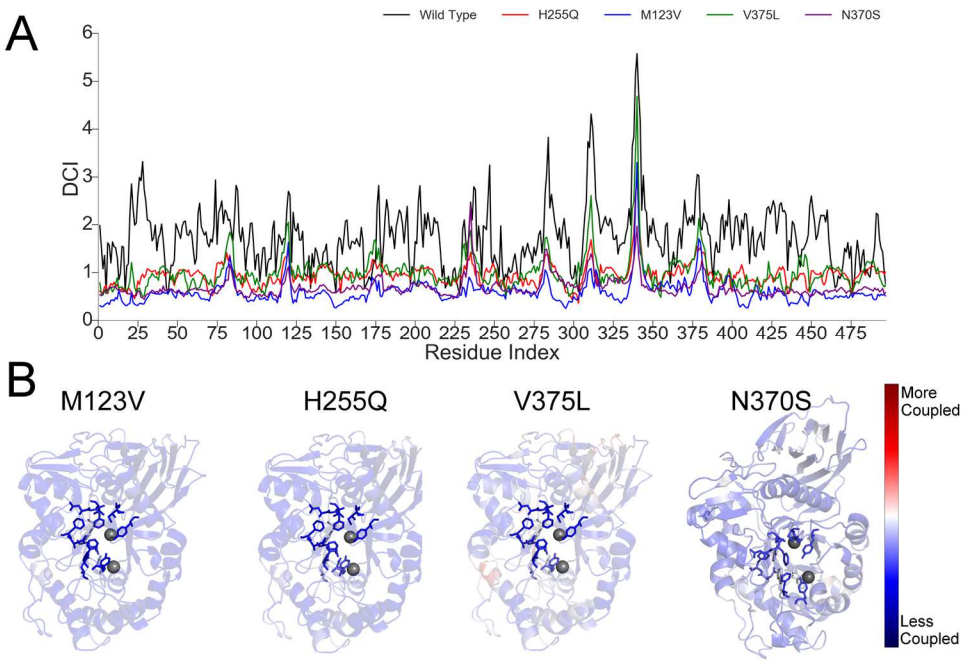

Figure 4.5: DCI profiles of disease mutants. (A) The DCI profile of the wild type (black) and disease mutants—H255Q (red), M123V (blue), V375L (green), N370S (purple). The DCI metric measures the dynamic allosteric residue coupling to functionally critical residues in a protein, which in this case are the catalytic residues Glu235 and Glu340. Residues with high DCI values are strongly coupled to the catalytic sites; residues with low DCI values are weakly coupled to the catalytic sites. All disease mutants show a global loss in dynamic coupling to the catalytic sites. This global loss of allosteric coupling disrupts any allosteric pathway involved in the regulation and function of the protein. (B) A ribbon diagram of the disease mutants colored according to ΔDCI ($[DCI_{disease} - DCI_{wt}] / DCI_{wt}$). The catalytic sites are colored as grey spheres. The positions in red indicate an increase in dynamic coupling upon mutation as compared to the wild type. The positions in blue indicate a decrease in dynamic coupling upon mutation as compared to the wild type. All residues involved in ligand binding have decreased ΔDCI profiles, indicating a loss in allosteric coupling with the catalytic sites and a loss in allosteric regulation for enzyme catalysis.

particularly with the decreased flexibility of critical positions (i.e., binding recognition sites) near the catalytic sites. We further analyzed the fractional change in dynamic coupling of the catalytic sites as upon disease mutation compared to the wild type as ΔDCI ($[DCI_{disease} - DCI_{wt}] / DCI_{wt}$). Figure 4.5B shows ribbon diagrams of the mutant structures color-coded according to their ΔDCI profiles for each disease mutant. Regions

that exhibit a large decrease in DCI as compared to the wild type become less coupled to the catalytic sites upon mutation (blue). The ΔDCI of all binding sites showed a severe loss in dynamic coupling to catalytic sites. These results suggest a plausible disease mechanism: a loss in dynamic coupling with catalytic sites, particularly binding recognition sites, can explain the drastic decrease in catalytic activity of GCase in disease mutants. It is worth noting that the modeled disease mutants exhibit similar DFI and DCI profiles with the N370S mutant in which the available crystal structure was used (PDB code: 3KE0), suggesting that the equilibrium dynamics of the modeled structures are reasonable and give consistent results with the simulation of the N370S crystal structure.

We did further analysis by comparing the DFI profiles of the neutral mutation, Q169R (found in the Humvar Dataset (Adzhubei *et al.*, 2010)), to the disease mutation M123V. The Q169R mutation is accepted as neutral putatively due to its prevalence in a large portion of the human population and thus has not been associated with the disease. However, this must be taken with caution since recent studies have reported that variants

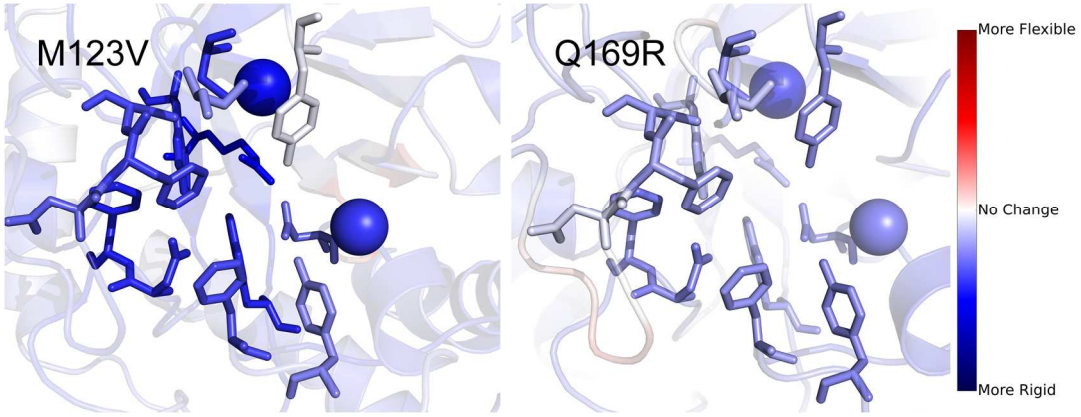

Figure 4.6: Ribbon diagram of the ΔDFI profile of the Q169R neutral mutation compared to the M123V disease mutation. The neutral mutation Q169R shows a slight decrease in the flexibility of the active site and residues involved in binding, however the degree of rigidification is significantly less than that of the disease variant M123V.

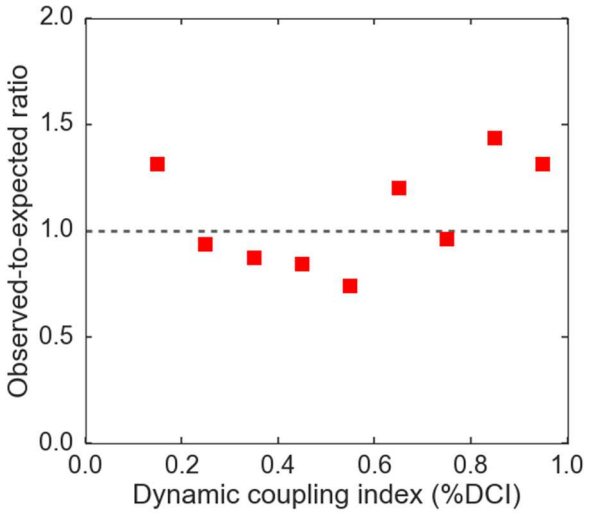

Figure 4.7: Observed-to-expected ratio of severe Gaucher Disease mutations. The expected set is the dynamic coupling index (%DCI) distribution of all residues in the protein. The observed set is the %DCI distribution of all residues associated with a severe type II Gaucher Disease missense mutation (Beutler et al., 2005). The observed-to-expected ratio of %DCI shows that severe mutations are abundantly found at high %DCI, which are sites allosterically coupled to catalytic residues and disproportionately not found at sites of low %DCI. A mutation at high %DCI sites will allosterically impact the dynamics of functional sites, leading to disruption in dynamic coupling and decreased enzymatic activity.

often observed in healthy individuals can still be weakly disruptive for molecular function (Bromberg *et al.*, 2013). The ΔDFI profile of Q169R indicates that the dynamics are only slightly altered as compared to a much more significant change exhibited in M123V (Figure 4.6). In particular, the sites critical for binding and recognition show a minimal change in dynamics in Q169R as compared to those in M123V, which exhibits a large decrease in flexibility. This evidence further substantiates our proposed mechanism that local changes in dynamic flexibility due to a mutation disrupts critical dynamic allosteric residue coupling with the catalytic sites, hampering catalytic activity.

4.2.4 A Majority of GD Variants at DARC Spots

Given the observation that the four disease mutations are allosterically coupled to the catalytic sites and lead to changes in DFI profiles of positions at distal sites in GCase, we speculated that this may be a general trend in 84 missense variants in GCase categorized as severe Type II GD (Beutler *et al.*, 2005). The DCI values were ranked into percentiles (%DCI) and sorted into bins of .10. The observed-to-expected ratio of %DCI values were computed, where the expected values were based on the %DCI distribution of all sites in GCase, and the observed values were the %DCI values of the 84 mutations. Under the null hypothesis of no effect, the ratio of the expected and observed numbers of sites hosting disease mutations should be close to 1.0 for each %DCI bin. A strong relationship between severe mutation sites and dynamic allosteric coupling with the active site (i.e., those exhibiting high DCI) would reject the null hypothesis that disease mutations are distributed uniformly in sites with low and high dynamic coupling. This null hypothesis was rejected in our analysis ($p < 0.046$). Figure 4.7 shows the observed-to-expected ratio of %DCI which indicates that severe disease mutations are overabundant at high %DCI sites (%DCI value of 0.8-1.0) with values greater than 1.0. This evinces that mutations at DARC spots likely impact function, leading to disease phenotypes.

4.2.5 Proteome-wide Analysis: Conformational Dynamics and DARC Spots

The role of site-specific structure-encoded dynamics was first demonstrated in a proteome-wide study using a dataset of Mendelian diseases, which revealed the correlation between the dynamic flexibility index (DFI) of residues in monomeric proteins and the biological phenotype of nSNVs (Nevin Gerek *et al.*, 2013). A subsequent

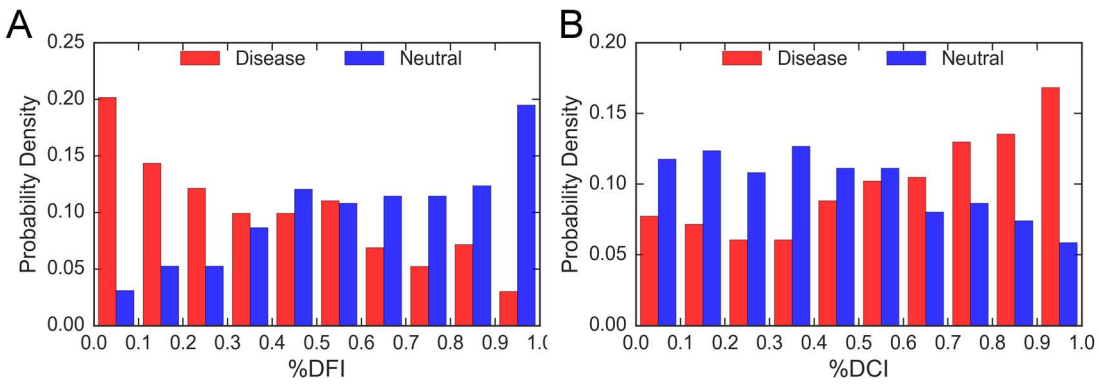

Figure 4.8: DFI and DCI distributions of nSNVs in a proteome-wide analysis of conformational dynamics of nSNVs. The %DFI distribution (A) and %DCI distribution (B) of 362 disease variants and 323 neutral variants expressed as a probability mass function (PMF). A student t-test comparing the disease and neutral distributions revealed a significant difference for both %DFI and %DCI ($p < 0.0001$).

study demonstrated the efficacy of DFI in analyzing the impact of nSNVs in biological assemblies (Butler *et al.*, 2015). These studies confirmed the utility of structural dynamics-based measures as a way to predict the functional impact of nSNVs across the proteome. Here we investigated the role of dynamic allosteric residue coupling with catalytic sites in human enzymes. We computed DFI and DCI profiles for 75 monomeric enzymes containing 685 missense mutations in which the phenotype was known (See Methods for the dataset). The DFI and DCI profile of each protein was converted to a percent ranking, %DFI and %DCI, so that the values could be compared across different proteins. As Figure 4.8A shows, the %DFI distribution for the 362 disease mutations exhibits a distinctly opposite trend as compared to the 323 neutral mutations ($p < 0.0001$), in agreement with our previous findings (Butler *et al.*, 2015; Kumar, Butler, *et al.*, 2015). The average %DFI for disease variants is 0.37 while that for neutral variants is 0.6, quantifying the validity of DFI as a measure of the functional and biological impact of mutations.

We also calculated dynamic coupling index (DCI) profiles for the enzymes in our data set to measure the dynamic allosteric residue coupling of mutation sites to catalytic sites. A given mutation may not occur at a hinge (i.e., low %DFI sites) but can be strongly coupled to, and thus dynamically alter, a catalytic site from distances of over 10Å. The DCI metric can, therefore, identify disease mutations that may have otherwise been overlooked by DFI. A mutation at a site that is highly coupled to an active site (%DCI >0.6) is a DARC spot that will likely disrupt function and lead to a disease phenotype. Figure 4.8B shows the %DCI distributions of 362 disease and 323 neutral mutations, which show opposite trends ($p < 0.0001$). As expected, the frequency of disease variants begins to increase sharply for %DCI >0.6, confirming that sites highly coupled to catalytic sites are likely to be disease-associated. Some neutral mutations (i.e., abundant in human population) were also found at DARC spots, which suggests the possibility that substitutions at these positions may still impact function, being mildly disruptive as implied by other approaches (Bromberg *et al.*, 2013; Reeb *et al.*, 2016). The DCI metric is particularly useful for two reasons: first, it can discriminate between disease and neutral phenotypes for mutations that are not spatially close to active sites. Second, it can be used in conjunction with DFI to ascertain likely disease mutations with increased prediction accuracy.

Computational methods for diagnosing phenotypic effects of mutations (e.g., EvoD (Kumar *et al.*, 2012), PolyPhen-2 (Adzhubei *et al.*, 2013), SIFT (Ng and Henikoff, 2003), PhD-SNP (Capriotti *et al.*, 2006), PANTHER (Thomas *et al.*, 2003), MutPred (Li *et al.*, 2009), SNPs&GO (Calabrese *et al.*, 2009), SNAP (Bromberg and Rost, 2007) and nsSNPAnalyzer (Bao *et al.*, 2005)) can achieve accuracies between 70-80% on

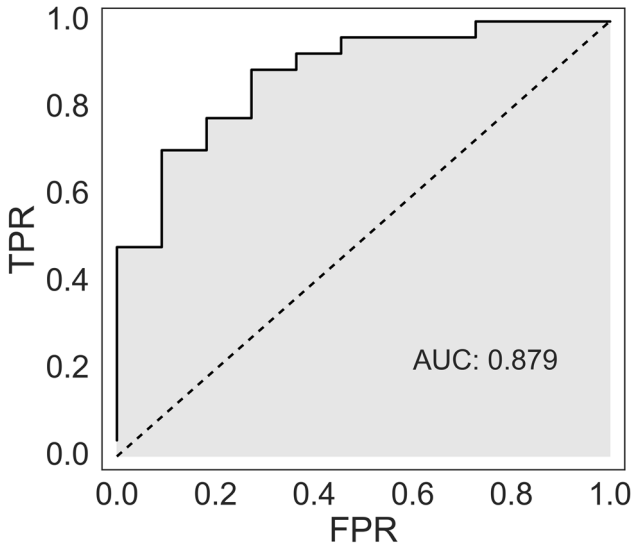

Figure 4.9: An ROC Curve for evolutionary misdiagnosed variants. From the dataset we selected all instances where at least one evolutionary method misdiagnosed nSNV phenotype (i.e., EvoD, PolyPhen-2, Sift). The dataset was divided into 90% training and 10% test sets and the features %DFI and %DCI were used to train a logistic regression model to predict whether an nSNV was deleterious or neutral. We evaluated the performance of our model by calculating an ROC curve, which yields an area under the curve (AUC) of 0.879, indicating that DFI and DCI are superior predictive metrics where evolutionary methods are deficient. Thus, dynamics can complement evolutionary methods in disease classification to obtain overall better results. A 5-fold and 10-fold cross-validation was also performed and shown in supplementary Figure 4.10.

independent, non-trained datasets. These metrics, which are largely based on evolutionary information such as conservation, tend to fail for disease variants at highly variable positions and benign variants at highly conserved positions (Kumar *et al.*, 2011; S. Kumar *et al.*, 2009). Thus, we further explored whether dynamics can be used as a complementary feature to evolution. From our original dataset we selected all instances where at least one evolutionary method misdiagnosed an nSNV phenotype (i.e., EvoD, PolyPhen-2, Sift). This subset of data was divided into 90% training and 10% test sets where the features %DFI and %DCI were used to train a logistic regression model to predict whether an nSNV was deleterious or neutral. We evaluated the performance of

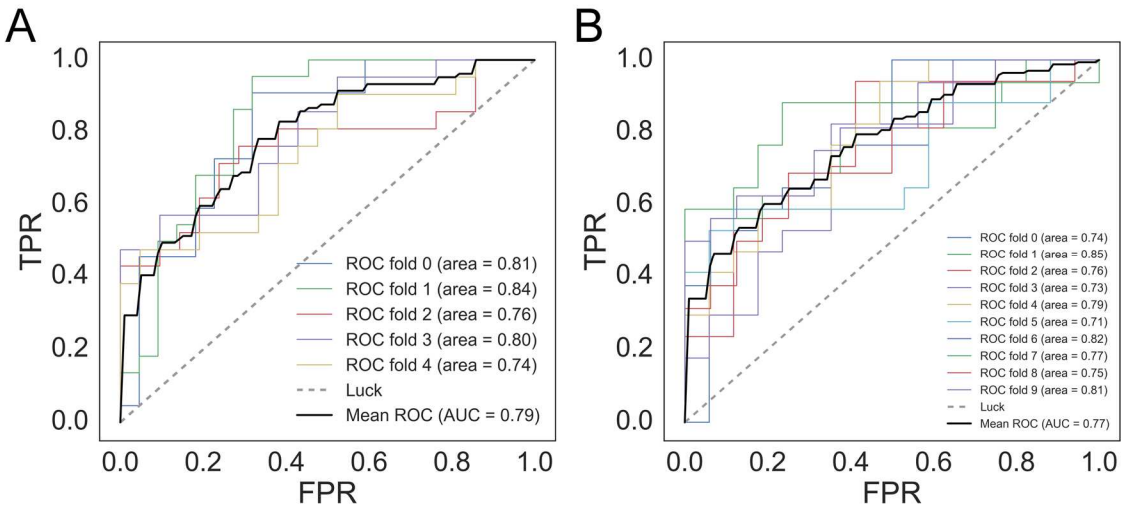

Figure 4.10: A K-fold stratified cross-validation plot. From the dataset we selected all instances where at least one evolutionary method misdiagnosed nSNV phenotype of a variant (i.e., EvoD, PolyPhen-2, Sift). A stratified K-fold cross-validation was performed in which our dynamics-based features DFI and DCI were used to train a logistic regression model to predict whether an nSNV was deleterious or neutral. (A) In a 5-fold cross-validation the AUC ranged from 0.74 to 0.84 with a mean AUC of 0.79. (B) In a 10-fold cross-validation the AUC ranged from 0.71 to 0.85 with a mean AUC of 0.77. This indicates that DFI and DCI are robust predictors for cases where the evolutionary methods are lacking in predictive power.

our model by calculating an ROC curve (Figure 4.9), which yielded an area under the curve (AUC) of 0.879, indicating that DFI and DCI are predictive metrics and can complement evolutionary methods in classifying nSNVs. Furthermore, a 5-fold and 10-fold cross-validation analysis resulted in a mean AUC of 0.79 and 0.77 respectively (Figure 4.10). These results suggest that dynamics-based metrics such as DFI and DCI can provide complementary information that can assist evolution-based models, particularly in the regime where they lack predictive ability.

4.3 Discussion

Previous analysis of our work and that of others has demonstrated the important role of conformational dynamics in the emergence of new functions from protein

evolution (Zou *et al.*, 2015; Kim *et al.*, 2015; Glembo *et al.*, 2012; Bhabha *et al.*, 2013; Campbell *et al.*, 2016; Kumar, Glembo, *et al.*, 2015; Kumar, Butler, *et al.*, 2015; Butler *et al.*, 2015; Nevin Gerek *et al.*, 2013). Our site-specific DFI analysis revealed that amino acid substitutions alter the flexibility/rigidity of various regions within a given protein, leading to the emergence of new functions. Additionally, DFI analysis of protein evolution showed that sites subject to drastic flexibility change during evolution are not the actual mutational sites but at sites distal to these mutations, indicating allosteric regulations play a role in protein evolution. For instance, the emergence of red-emitting proteins is achieved by 13 mutations of the least-evolved green fluorescence ancestor protein. While DFI profiles of these mutations do not significantly change, this distributed set of mutations on GFP causes enhanced flexibility of several positions near the chromophore that propagates throughout the protein, ultimately converging on a distant location for compensatory stiffening (Kim *et al.*, 2015).

If nature uses dynamic allosteric residue coupling for the emergence of new functions it follows that disease mutations could exploit the same mechanism for disease pathogeny. Here we explored the mechanistic role conformational dynamics in human disease. In particular, we focused on how disease mutations influence protein dynamics, whether they impair critical allosteric residue couplings to catalytic sites and, subsequently, their impact on enzymatic function. We first analyzed the missense variants associated with GD. There are over 200 missense variants frequently observed in human populations that drastically decrease the enzymatic function of GCase and lead to different GD phenotypes (Hruska *et al.*, 2008). Interestingly, we have found a majority of

disease mutations occur at DARC spots, meaning they have long-range dynamic coupling to the catalytic sites.

We did a rigorous evaluation of four disease variants that are often misdiagnosed as neutral in many evolution-based prediction methods. DFI profiles of GD mutations show a significant change in flexibility/rigidity profiles as compared to the wild type. In agreement with the earlier work of Liebermen et al (Lieberman *et al.*, 2007) disease variants, particularly N370S, exhibited an increased flexibility of both loop1 and loop4 and a decrease in flexibility of the sites involved in binding; both changes could interfere with enzymatic function. Overall, this analysis shows that mutations remotely alter the flexibility of the regions critical for the enzymatic function, a phenomenon we also observed in the emergence of new functions in protein evolution. Furthermore, disease mutations also lead to loss in dynamic allosteric coupling of the catalytic site with the rest of protein. This results in a global affect on the protein's allosteric network and may significantly impact enzymatic function. As discussed in a review (Gunasekaran *et al.*, 2004), allostery is a common property among all proteins and is necessary for enzymatic function. In GCase, this network disruption could interfere with communication between the enzymatic domain and SapC interacting domains. In previous work we also observed how disease mutations altered the dynamic coupling of the functional loop in human ferritin (Kumar, Glembo, *et al.*, 2015; Kumar, Butler, *et al.*, 2015).

Lastly, we conducted a proteome-wide analysis of 75 monomeric human enzymes to investigate the role of conformational dynamics and dynamic allosteric residue coupling. To our knowledge, this is the first proteome-wide analysis of disease mutations in enzymes. We observed the same trend as seen in missense disease variants of GCase.

Compared to neutral mutations, most of the disease mutations occurred at low DFI positions and/or have high DCI values indicating they are DARC spots. Interestingly, when we focused on cases that are challenging to correctly diagnose by evolution-based prediction methods, we observed that DFI and DCI was able to complement these methods and correctly predict at least 70% of misdiagnosed missense variants.

To summarize, we are in the era of rapid development of next-generation methods for whole-genome, whole-exome, and targeted sequencing that has generated an unprecedented amount of data. Among all the variation data, the most commonly observed variants are nSNVs, and identifying the nSNVs with pathogenic effects that contribute to disease or drug sensitivities is the primary goal of 21st century genomic analysis and phylomedicine. As stated in a review of allostery by Liu and Nussinov (Liu and Nussinov, 2016), uniting the genetic code, which constitutes "the first secret of life," and allostery, "the second secret of life," could reveal a generalized disease mechanism and allow for discovery of novel drugs, as well as the blueprints for deeply innovative personalized treatment methods.

4.4 Methods

4.4.1 Dataset

A total of 75 individual monomeric protein structures from the Protein Data Bank (PDB) (Berman *et al.*, 2000) were collected from a BLAST search of sequences with requirements of ≥80% sequence identity and ≥80% query coverage to ensure only structures that could be accurately mapped to human variation data were included. Human genetic variations were obtained from the HumVar and HumDiv databases

(Adzhubei *et al.*, 2010) with a total of 685 non-synonymous single nucleotide variants (nSNVs), where 323 were neutral and 362 were deleterious.

4.4.2 Determining Catalytic Sites

The catalytic sites were gathered from the Catalytic Site Atlas (CSA) database (Furnham *et al.*, 2014) which identifies the residues that are directly involved in catalyzing the reactions of enzymes. Since these residues are critical for protein function, they were used as input into our dynamic coupling index (DCI) metric. The entries in the CSA were either "original entries" derived from the literature itself or "homology entries" based on sequence comparison with the literature-based original entries. In either case, the catalytic sites purported by the CSA should accurately represent functional sites on the protein. Our dataset contained 75 enzymatic proteins that mapped to entries in the CSA database.

4.4.3 Calculating Functional-dynamics Profiles

The method for obtaining the dynamic flexibility index (DFI) is based on the perturbation response scanning (PRS) method (Atilgan and Atilgan, 2009), in which the C-alpha atom of each residue in the protein is modeled as a node in an elastic network model (ENM). The interaction between each node is modeled by a harmonic potential with a distance-dependent spring constant (Atilgan and Atilgan, 2009; Atilgan *et al.*, 2001). A small perturbation in the form of an external random force (i.e., Brownian kick) is sequentially applied on each node in the network and the perturbation response of all nodes is recorded according to linear response theory as

$$[\Delta \boldsymbol{R}]_{3N\times 1} = [\boldsymbol{H}]^{-1}_{3N\times 3N}[\Delta \boldsymbol{F}]_{3N\times 1} \tag{4.1}$$

Where $\boldsymbol{F}$ is the external random force, $\boldsymbol{H^{-1}}$ is the inverse Hessian, and $\Delta\boldsymbol{R}$ is the positional displacement of all *N* nodes in three dimensions. Each perturbation is performed in ten different directions to ensure an isotropic response. The perturbation is repeated for every node in the network, and the positional displacements $\Delta\boldsymbol{R}$ of each node are stored in a perturbation matrix $\boldsymbol{A}$ given by

$$[\boldsymbol{A}]_{N\times N} = \begin{bmatrix} \Delta|R^1|_1 & \Delta|R^2|_1 & \cdots & \Delta|R^N|_1 \\ \Delta|R^1|_2 & \Delta|R^2|_2 & \cdots & \Delta|R^N|_2 \\ \vdots & \vdots & \ddots & \vdots \\ \Delta|R^1|_{N-1} & \Delta|R^2|_{N-1} & \cdots & \Delta|R^N|_{N-1} \\ \Delta|R^1|_N & \Delta|R^2|_N & \cdots & \Delta|R^N|_N \end{bmatrix} \tag{4.2}$$

Where $\left|\Delta R^j\right|_i = \sqrt{\langle \Delta R^2 \rangle}$ is the magnitude of the positional displacement of each residue *i* in response to a perturbation at residue *j*. The DFI score of residue *i* is defined as the sum of the total displacement of residue *i* induced by a perturbation on all residues, which is computed by taking the sum of the *i*th row of the perturbation matrix **A**,

$$DFI_i = \frac{\sum_{j=1}^{N}\left|\Delta R^j\right|_i}{\sum_{i=1}^{N}\sum_{j=1}^{N}|\Delta R^j|_i} \tag{4.3}$$

Where the denominator is the total displacement of all residues, used as a normalizing factor.

Recently, we have extended this method to identify allosteric links or dynamic coupling between any given residue and functionally important residues by introducing a new metric called the *dynamic coupling index* (DCI) (Kumar, Glembo, *et al.*, 2015). The DCI metric can identify DARC spots, which are sites that are distal to functional sites but control them through dynamic allosteric coupling. This type of allosteric coupling is important; sites with strong dynamic allosteric coupling to functionally critical residues (DARC spots), regardless of separation distance, likely contribute to the function as well.

Thus, a mutation at such a site can disrupt the allosteric dynamic coupling or regulation, leading to functional degradation. As defined, DCI is the ratio of the sum of the mean square fluctuation response of the residue *i* upon functional site perturbations (i.e., catalytic residues) to the response of residue *i* upon perturbations on all residues. DCI enables us to identify DARC spot residues, which are more sensitive to perturbations exerted on residues critical for function. This index can be utilized to identify the residues involved in allosteric regulation. It is expressed as

$$DCI_i = \frac{\sum_{j=N_{functional}}^{N_{functional}} \left|\Delta R^j\right|_i \Big/ N_{functional}}{\sum_{i=1}^{N} |\Delta R^j|_i \Big/ N} \tag{4.4}$$

Where $\left|\Delta R^j\right|_i$ is the response fluctuation profile of residue *i* upon perturbation of residue *j*. The numerator is the average mean square fluctuation response obtained over the perturbation of the functionally critical residues $N_{\text{functional}}$ and the denominator is the average mean square fluctuation response over all residues. As discussed below, the DFI and DCI profiles can also be computed using the covariance matrix obtained from Molecular Dynamics simulations rather than the inverse Hessian of the elastic network model.

4.4.4 Molecular Dynamics Simulations

Molecular Dynamics (MD) simulations were performed using the AMBER 14 MD package (Pearlman *et al.*, 1995). Simulations were run using the Amber14SB forcefield. The TIP3P (Sun and Kollman, 1995) water model was used for solvation. The pmemd.cuda (Salomon-Ferrer *et al.*, 2013) executable of the AMBER14 package was used for GPU acceleration. All simulations were run for 50ns.

To obtain dynamics for each variant, each trajectory was divided into 5ns windows. The covariance matrix **G** for each window was extracted, instead of the inverse hessian in the ENM (as in Equation 1), and used to calculate the corresponding DFI and DCI profiles as

$$[\Delta \boldsymbol{R}]_{3N\times 1} = [\boldsymbol{G}]_{3N\times 3N}[\Delta \boldsymbol{F}]_{3N\times 1} \tag{4.5}$$

The DFI and DCI profiles calculated from the last four windows (the final 20ns) were averaged to calculate an average DFI and DCI profile. We further investigated the change in dynamics upon mutation compared to the wild type structure using ΔDFI and ΔDCI. The delta-DFI (ΔDFI) profile was calculated as

$$\Delta DFI_i = \frac{DFI_{disease} - DFI_{wt}}{DFI_{wt}} \tag{4.6}$$

Where $DFI_{disease}$ is the dynamics profile for the mutated protein structure and DFI_{wt} is the dynamics profile for the wild type structure. Similarly, the delta-DCI (ΔDCI) profile was calculated as

$$\Delta DCI_i = \frac{DCI_{disease} - DCI_{wt}}{DCI_{wt}} \tag{4.7}$$

CHAPTER 5

5 ESTIMATING DYNAMICS FROM PROTEIN SEQUENCES

We describe a novel approach that estimates the dynamics profile of a protein from its amino acid sequence. This *de novo* approach leverages the evolutionary principle of coevolution and the Gaussian network model (GNM). We demonstrate that our sequence-based GNM approach produces values in good agreement with crystallographic B-factors as well as theoretical values predicted from the original GNM that uses the structure with a distance cutoff. Remarkably, the results also suggest the ability of our sequence-based approach to classify the phenotypes of genomic variants across the proteome.

5.1 Introduction

The advent of high-throughput genomic sequencing has led to a burgeoning of sequences, providing an unprecedented amount of data for genomic analysis. Furthermore, this has also driven the rapid classification of novel genetic variations through genome-wide association studies (M. J. Li *et al.*, 2012; Xie *et al.*, 2014). Genetic variations usually manifest as non-synonymous single nucleotide variants (nSNVs) that can severely impact protein function and lead to disease. Evolutionary approaches based on positional amino acid conservation are the most common way to diagnose nSNVs. Protein dynamics can also be used to elucidate the functional impact of nSNVs and mechanisms of disease, and some recent studies have evinced that a site-specific conformational dynamics was capable of diagnosing nSNVs irrespective of evolutionary

conservation (Nevin Gerek *et al.*, 2013; Butler *et al.*, 2015; Kumar, Butler, *et al.*, 2015). This was the first implementation of site-specific conformational dynamics into proeome-wide analysis to predict the functional impacts of nSNVs. Although the importance of protein dynamics in genetic variation analysis is undeniable, the 3D structure from the Protein Data Bank (Berman *et al.*, 2000) is still required to calculate protein dynamics. This drastically limits the range of applicability in genomic analysis, since there are exponentially more sequences than experimental structures.

Recently, coevolution has become a popularized tool for its ability to predict structural contacts of 3D structures from sequence information (Marks *et al.*, 2011; Hopf *et al.*, 2014; Morcos *et al.*, 2011; Marks *et al.*, 2012; Hopf *et al.*, 2012). Coevolving residues are inferred from a multiple sequence alignment (MSA) of a given protein family, whereby if two given amino acids exhibit concordant patterns of evolution throughout the MSA then they are assumed to be close in spatial proximity in the folded 3D structure. For a given protein family, the conservation of certain mutations over different homologs serves as a restraint on function, which is why there can be many different sequences in a family that lead to a protein with the same function. Thus, this evolutionary principle allows us to leverage sequence information to describe protein topology, making *de novo* structure predictions possible (Marks *et al.*, 2012). It has been reported that only one correct contact for every 12 residues in a protein is necessary for accurate topology-level modeling (Kim *et al.*, 2014). A study by Marks et al. used coevolution to predict 15 structures from different fold classes (ranging between 50-260 residues), including a G-protein coupled receptor (a class of membrane proteins that are notoriously difficult to predict) with minimal RMSD error between 2.7–4.8 Å relative to

the known structure (Marks *et al.*, 2011). In addition to structure prediction, coevolution analysis has also been used to identify critical interactions between protein complexes (Hopf *et al.*, 2014) important functional sites (Marks *et al.*, 2012) and allosteric response (Smock *et al.*, 2010). The use coevolution in structure prediction is largely possible for two reasons. First, the amount of sequence data for different protein families is sufficient to be leveraged by this technique to make predictions. Second, the methods for inferring coevolving residues from an MSA are becoming increasingly superior and accurate (de Juan *et al.*, 2013).

Inferring evolutionary couplings from an MSA are based on three primary approaches: maximum entropy models, Bayesian network models, and machine learning. A problem in coevolution analysis is that many residue pairs predicted to be correlated in are not close in spatial proximity, and thus are not true structural contacts. This is mainly due to transitive correlations in the MSA of the protein family that leads statistical "noise" and incorrectly predicted couplings (indirect couplings). For instance, if residue B is correlated with A and C, then we may find that A and C are correlated even though they may not be structural contacts–A and C is a transitive correlation. Thus, the challenge is discerning coevolving pairs that are true spatial contacts (direct couplings) from the background of noise created by weakly correlated mutations due to transitive effects that are not true spatial contacts (indirect couplings). Mutual information (MI) is a statistical model that has been used to infer coevolving residue pairs in an MSA (Gouveia-Oliveira and Pedersen, 2007; Dunn *et al.*, 2008). It considers the frequency of occurrence of amino acids in each column *i,j* in a given MSA and uses Shannon information entropy to determine which pairs are most likely correlated. For a given

column *i* in an MSA, $f_i(A)$ is the frequency count of a particular amino acid, *A*. Likewise, for a pair of columns *i,j* in an MSA, $f_{ij}(A,B)$ is the frequency of amino acids, *A* and *B*, appearing simultaneously in the two columns. Then the MI score indicates the degree of correlation between two residues *i,j* (columns in the MSA) as $MI_{ij} = \sum_{A,B} f_{ij}(A,B) \ln \frac{f_{ij}(A,B)}{f_i(A) f_j(B)}$ (Cover and Thomas, 2006). The MI method is a local measure of correlations (i.e., it only considers the correlation of one residue pair at a time), thus it is inherently limited by the transitivity effect and cannot discern direct from indirect couplings (Marks *et al.*, 2011). For this reason, MI is not sufficient to describe spatial proximity of contacts in the 3D structure. Direct coupling analysis (DCA) is a global approach that can disentangle direct couplings from indirect couplings and captures true spatial contacts (Morcos *et al.*, 2011). This statistical approach is based on the maximum entropy method, which gives the maximal probability function of residue pair correlations over the whole sequence of length *L* as $P(A_1, \ldots, A_L) = \frac{1}{Z} \exp\left[\sum_{i<j} e_{ij}(A,B) + \sum_i h_i(A)\right]$, with *Z* being a normalizing factor, $e_{ij}(A,B)$ the pairwise couplings, and $h_i(A)$ a local bias field. This formulation is analogous to the Ising model describing neighboring spin interactions in ferromagnetic materials. The pairwise couplings $e_{ij}(A,B)$ and $h_i(A)$ are solved by various numerical technqies (e.g., mean field approximation). Then the strength of coupling between residue pairs is given by their direct information (DI) score as $DI_{ij} = \sum_{A,B} P_{ij}^{Dir}(A,B) \ln\left(\frac{P_{ij}^{Dir}(A,B)}{f_i(A) f_j(A)}\right)$, where $P_{ij}^{Dir}(A,B)$ is the effective pair probability and $f_i(A)$ and $f_j(B)$ are the single residue frequencies. A notable difference between DI and MI is that the local frequency count

$f_{ij}(A, B)$ in MI is replaced by the pair probability $P_{ij}^{Dir}(A, B)$ which captures global coupling effects over all pairs *i,j* in the sequence. Thus, the DCA approach eliminates the problem of transitive correlations that occurs with MI, and can be surmised as a "noise filter" that can identify the most causative correlations that represent spatial contacts in the tertiary structure.

Several web servers have been developed that use DI in the DCA framework such as EVfold (Marks *et al.*, 2011) and PSICOV (Jones *et al.*, 2012). Another approach uses pseudo-likelihood maximization (PLM) instead of DI to infer direct couplings (Ekeberg *et al.*, 2013), which is implemented in the servers GREMLIN (Kamisetty *et al.*, 2013) and CCMpred (Seemayer *et al.*, 2014). Bayesian network models can also be used to predict evolutionary couplings and improve contact map prediction (Burger and Van Nimwegen, 2010). Regardless of the method, the accuracy of detecting coevolving residues that correspond to structural contacts is contingent on the number of sequence homologs in the MSA (a sufficient number is on the order of 5 ×[length of protein sequence]). Most of the current methods use only the sequence homologs of the protein family of the target sequence, which is often less than the optimal number of homologs to produce accurate statistical inferences. Dunn et al. previously addressed this concern by integrating multiple protein families that were orthologs (i.e., they share similar phylogeny and retain similar functions) to increase the number of homologs (Dunn *et al.*, 2008). A recent approach, RaptorX, leveraged this joint family approach and used a supervised machine learning method along with coevolution information to infer 3D contacts, and has proven to obtain higher accuracy than the other methods (Ma *et al.*, 2015; Wang *et al.*, 2016).

We propose a novel sequence-based approach to estimate the dynamics profile of a protein, with no *a priori* knowledge of its 3D structure. This *de novo* approach based on a Gaussian network model (GNM) enables the prediction of the magnitude of mean-square fluctuations of residues, which are proportional to the B-factors determined by X-ray crystallography experiments (as described in detail in Chapter 2). However, instead of using a cutoff distance to determine 3D contacts as in the original GNM, we use coevolving residues (evolutionary couplings (ECs)) in our model. We show that the theoretical predictions from our sequence GNM are in good agreement with experimental crystallographic B-factors as well as values obtained from the original GNM. We also extend this analysis to determine the capacity of our model to assess the functional impact of nSNVs. We will demonstrate that the dynamics as predicted from the sequence GNM can classify disease and benign nSNVs across the proteome.

5.2 Methods

5.2.1 Dataset

A curated set of 139 structures was procured from a previous dataset used by Butler et al. in a proteomic study of enzymes (presented in Chapter 4). These structures were selected for several reasons. First, they have high query coverage (>80%) and sequence identity (>80%) as found from a BLAST search, and the structures had already been modeled using the Modeller software package (Eswar *et al.*, 2006) to account for any missing residues. Second, genetic variants were previously mapped onto these structures, such that the positions containing known nSNVs were already determined, enabling us to easily compare our results using sequence coevolution with the genetic variation data. A total of 738 genetic variants were obtained from the HumVar database

(Adzhubei *et al.*, 2010), which was comprised of 436 disease and 302 neutral nSNVs. Finally, the structures were either monomers or the single-chain unit of a multimer with <600 residues, allowing for tractable calculations of residue coevolution using the RaptorX web server. A table summarizing the dataset is presented in Table 1.

Table 1: Proteins used to compute theoretical B-factors based on sequence GNM and structural GNM approaches.

						Correlation Coefficient of B-factors		
PDB	Length	Resolution (Å)	Biological Unit	Disease nSNVs	Neutral nSNVs	SequenceGNM–Experiment	StructureGNM–Experiment	SequenceGNM–StructureGNM
12ca	255	2.4	MONOMERIC	2	1	0.489	0.605	0.479
12gs	208	2.1	DIMERIC		1	0.332	0.664	0.374
1bix	275	2.2	MONOMERIC		1	0.247	0.731	0.420
1c7p	133	2.4	MONOMERIC	1	1	0.671	0.696	0.715
1crm	256	2	MONOMERIC	3	16	0.530	0.640	0.487
1d4a	273	1.7	DIMERIC		2	0.432	0.673	0.191
1d5r	307	2.1	MONOMERIC	8		0.500	0.621	0.469
1d6n	214	2.7	DIMERIC	12	7	0.085	0.359	0.283
1dhf	182	2.3	DIMERIC	2		0.616	0.649	0.750
1eai	240	2.4	TETRAMERIC		1	0.636	0.770	0.734
1ege	387	2.75	TETRAMERIC	3	4	0.244	0.248	0.765
1eh5	279	2.5	MONOMERIC	5		0.464	0.646	0.552
1f7z	213	1.55	DIMERIC	2		0.678	0.551	0.820
1fb5	320	3.5	TRIMERIC	37	3	0.349	0.738	0.523
1fj2	229	1.5	MONOMERIC		1	0.827	0.560	0.573
1fro	176	2.2	DIMERIC		1	0.566	0.404	0.921
1gv7	123	2.1	MONOMERIC	5	1	0.337	0.569	0.692
1h0n	288	2.4	DIMERIC	2		0.462	0.561	0.574
1hdr	236	2.5	DIMERIC	5	2	0.352	0.622	0.561
1hne	218	1.84	HEXAMERIC	11		0.660	0.779	0.767
1hrk	359	2	DIMERIC	2		0.411	0.581	0.530
1hrn	335	1.8	DIMERIC	2		0.397	0.443	0.691
1hyo	416	1.3	DIMERIC	6		0.636	0.586	0.467
1i0z	332	2.1	TETRAMERIC		9	0.729	0.611	0.721
1iat	556	1.62	DIMERIC	5	3	0.624	0.485	0.556
1ib0	272	2.3	MONOMERIC	7	2	0.559	0.644	0.704
1ihg	364	1.8	MONOMERIC		2	0.255	0.399	0.704
1is7	194	2.8	EICOSAMERIC	7		0.541	0.534	0.719
1itq	369	2.3	DIMERIC		1	0.517	0.729	0.811
1ivh	387	2.6	TETRAMERIC	1		0.366	0.426	0.739
1j9w	257	2.6	MONOMERIC		3	0.298	0.522	0.584

1jph	357	2.1	DIMERIC	7	2	0.532	0.556	0.781
1jqi	384	2.25	TETRAMERIC		1	0.561	0.573	0.774
1jxv	149	2.2	HEXAMERIC	1	2	0.201	0.407	0.809
1k62	450	2.65	TETRAMERIC	1		0.549	0.573	0.696
1l1f	496	2.7	HEXAMERIC	3	16	0.420	0.267	0.799
1li4	430	2.01	TETRAMERIC		1	0.240	0.306	0.657
1ls6	288	1.9	MONOMERIC		2	0.391	0.637	0.296
1m9n	590	1.93	DIMERIC	1	1	0.402	0.233	0.571
1mc5	373	2.6	DIMERIC		1	0.335	0.364	0.625
1mir	313	2.8	DIMERIC		1	0.552	0.704	0.638
1og2	462	2.6	MONOMERIC		1	0.549	0.594	0.704
1ogs	497	2	MONOMERIC	53	1	0.362	0.688	0.576
1ohv	461	2.3	DIMERIC	1	10	0.435	0.518	0.265
1ore	179	2.1	DIMERIC	1		0.657	0.744	0.794
1pf7	288	2.6	TRIMERIC	2		0.490	0.576	0.721
1q6x	600	2.5	MONOMERIC	1	3	0.515	0.684	0.673
1qo5	360	2.5	TETRAMERIC		7	0.582	0.445	0.319
1r46	390	3.25	DIMERIC	23		0.490	0.554	0.774
1rx0	384	1.77	TETRAMERIC	2		0.415	0.497	0.692
1ry0	319	1.69	MONOMERIC		1	0.607	0.698	0.400
1s8o	545	2.6	DIMERIC		1	0.439	0.641	0.638
1sir	390	2.6	TETRAMERIC	13	4	0.311	0.363	0.684
1spj	236	1.7	MONOMERIC		1	0.767	0.826	0.809
1sw0	247	1.71	DIMERIC	2	2	0.442	0.303	0.695
1t1u	597	1.55	MONOMERIC	10		0.478	0.731	0.670
1tdi	218	2.4	DIMERIC		1	0.743	0.596	0.446
1u7t	255	2	TETRAMERIC	1		0.374	0.411	0.759
1umk	271	1.75	MONOMERIC	1	1	0.778	0.788	0.682
1v9e	259	1.95	MONOMERIC	3		0.620	0.733	0.644
1vrp	370	2.1	DIMERIC		1	0.401	0.550	0.804
1waw	366	1.75	DIMERIC		2	0.615	0.671	0.785
1wsr	371	2	DIMERIC	2		0.622	0.655	0.760
1wva	309	1.94	TRIMERIC	1		0.305	0.365	0.808
1xfb	342	3	TETRAMERIC	1	4	0.680	0.566	0.778
1yuw	554	2.6	MONOMERIC		3	0.453	0.637	0.722
1z10	465	1.9	MONOMERIC		8	0.418	0.678	0.698
1zmc	472	2.53	DIMERIC	1	18	0.436	0.528	0.728
2aaq	461	2.6	DIMERIC		1	0.290	0.415	0.596
2ag5	256	1.84	TETRAMERIC		1	0.568	0.459	0.769
2bh9	489	2.5	DIMERIC	34		0.403	0.422	0.811
2boa	404	2.2	MONOMERIC		1	0.677	0.382	0.388
2c9y	218	2.1	MONOMERIC		1	0.265	0.343	0.788
2cga	245	1.8	DIMERIC		1	0.167	0.636	0.337
2esl	181	1.9	DIMERIC		1	0.543	0.616	0.775
2etl	223	2.4	MONOMERIC	1		0.435	0.596	0.586
2f3b	326	1.8	TETRAMERIC		1	0.680	0.552	0.708
2fpg	305	2.96	DIMERIC	1		0.254	0.498	0.402
2fvl	323	2.4	MONOMERIC		1	0.288	0.425	0.417
2gao	186	2	DIMERIC	1		0.160	0.677	0.472
2h57	165	2	MONOMERIC	4	10	0.433	0.621	0.739
2he3	185	2.1	TETRAMERIC		1	-0.060	0.180	0.533
2hgs	472	2.1	DIMERIC	1	4	0.600	0.600	0.627
2i3y	188	2	TETRAMERIC		1	0.227	0.413	0.469
2ib7	391	2.05	TETRAMERIC	1	5	0.540	0.709	0.768
2iw2	478	1.82	DIMERIC	3		0.388	0.586	0.445
2j0f	446	2.31	DIMERIC	2		0.148	0.464	0.418
2j6l	497	1.3	TETRAMERIC	1	1	0.614	0.365	0.325
2jao	196	2	DIMERIC		1	0.833	0.625	0.577
2jbm	290	2	HEXAMERIC		1	0.500	0.204	0.458
2jif	381	2	TETRAMERIC		1	0.482	0.398	0.838
2o48	331	2.59	DIMERIC		2	0.645	0.537	0.651
2ozl	362	1.9	TETRAMERIC	6	8	0.454	0.350	0.604
2p02	380	1.21	DIMERIC	5		0.641	0.705	0.600
2p9q	416	2.7	MONOMERIC	7	17	0.391	0.490	0.704
2pla	343	2.51	DIMERIC	2	6	0.140	0.323	0.675

2vcy	332	2.41	DIMERIC		1	0.488	0.636	0.819
2w2j	268	1.6	MONOMERIC		4	0.810	0.746	0.736
2w8n	480	2	TETRAMERIC	2	1	0.234	0.157	0.519
2wd9	533	2.6	MONOMERIC		1	0.105	0.249	0.416
2wid	527	2.3	MONOMERIC	11	9	0.518	0.704	0.771
2wkl	497	2.7	MONOMERIC	3	3	0.317	0.655	0.530
2xe6	413	1.74	MONOMERIC	1	8	0.568	0.567	0.781
3bur	325	1.62	MONOMERIC	2		0.375	0.751	0.393
3cmq	405	2.2	MONOMERIC		2	0.496	0.631	0.513
3ecr	339	2.18	MONOMERIC	1		0.552	0.656	0.572
3fw1	228	1.75	DIMERIC		2	0.385	0.580	0.596
3gro	277	2.53	DIMERIC	1		0.472	0.618	0.521
3h53	387	2.01	DIMERIC	1		0.459	0.557	0.753
3hlm	401	2.5	DIMERIC		1	0.369	0.265	0.948
3i2b	138	2.3	TRIMERIC	9		0.410	0.480	0.643
3iar	360	1.52	MONOMERIC	6	1	0.384	0.552	0.788
3ibd	465	2	MONOMERIC		1	0.419	0.519	0.663
3k9v	464	2.5	MONOMERIC	4		0.426	0.490	0.692
3o9m	531	2.98	TETRAMERIC	6	4	0.454	0.630	0.777
3pm0	463	2.7	MONOMERIC	2		0.528	0.509	0.643
3qic	453	2.2	MONOMERIC	1	7	0.107	0.676	0.415
3ruk	473	2.6	MONOMERIC	2	16	0.511	0.329	0.383
3sza	447	1.48	DIMERIC		2	0.627	0.795	0.790
3t1g	349	2.35	MONOMERIC	12		0.658	0.786	0.753
3u2o	366	2.18	MONOMERIC	1		0.567	0.553	0.507
3v9g	541	2.5	DIMERIC		1	0.414	0.443	0.796
3vn9	291	2.6	MONOMERIC		3	0.580	0.705	0.669
4ah6	589	3.7	DIMERIC	2		0.269	0.420	0.680
4aj4	331	1.9	TETRAMERIC		1	0.560	0.428	0.878
4ald	363	2.8	TETRAMERIC	1	1	0.745	0.500	0.404
4aoh	123	1.04	MONOMERIC	10		0.581	0.616	0.729
4awn	260	1.95	MONOMERIC		2	0.590	0.739	0.743
4b3e	153	2.15	DIMERIC	24		0.762	0.471	0.484
4fdi	494	2.2	DIMERIC	3		0.471	0.715	0.371
4g1c	267	1.94	DIMERIC		1	0.699	0.526	0.542
4gab	316	1.6	MONOMERIC		1	0.374	0.707	0.342
4h2i	524	2	DIMERIC		1	0.429	0.625	0.736
4hva	238	2.07	TETRAMERIC		1	0.631	0.665	0.735
5pnt	157	2.2	MONOMERIC		1	0.775	0.684	0.747
7pck	314	3.2	TETRAMERIC	2		0.450	0.548	0.713
1dch	104	3	TETRAMERIC	1		–	–	0.477
1pbh	317	3.2	MONOMERIC		1	–	–	0.656
1xwn	174	–	MONOMERIC		1	–	–	0.932

5.2.2 Obtaining Coevolved Residues

The amino acid sequence from each of the 139 structures was used as input for the evolutionary coupling (EC) analysis. The choice of taking the amino acid sequence from the structure was done so that the predicted EC contacts could be compared directly to the experimentally observed structure contacts as verification that the model was producing realistic contact maps. Moreover, the theoretical B-factors predicted by our sequence-based model could be directly compared to the experimental B-factors for each

protein. If the structure was unknown, however, sequence databases (e.g. UniProt, PFAM, etc.) as discussed in Chapter 1 could be used. The PDB sequences were given to the RaptorX web server (Wang *et al.*, 2016; Ma *et al.*, 2015), which computed the relative probability of each residue pair *i, j* of being in 3D contact based on their coevolution strength. In order to ensure consistency between different proteins of varying lengths, we converted the raw scores into percentile ranks. We then used a threshold value, taking only the top scoring evolutionary couplings (i.e., the strongest couplings are more likely to be in spatial contact). An optimized threshold value was systematically evaluated and is discussed in the Results (Section 5.3).

5.2.3 Sequence-based GNM Model

The Gaussian network model (GNM) is an isotropic approach based on the contact topology of a crystal protein structure to obtain the equilibrium fluctuations of residues due to thermal motion. It uses a specified cutoff distance to define interacting pairs that are connected by springs with a single-parameter harmonic potential (detailed review in Chapter 2). In this structure-based GNM, the interacting residue pairs within the cutoff range are represented as contacts in the Kirchhoff (connectivity matrix).

In the proposed sequence-based GNM approach we will instead use coevolving residue pairs (evolutionary couplings) as contacts in the Kirchhoff. In this way, the 3D structure is no longer a prerequisite to form a GNM. To construct the Kirchhoff, a threshold is defined where any evolutionary coupling scores above that threshold are sufficiently coupled such that they are spatially close in 3D structure. If a given evolutionary coupling pair meets the threshold criteria, it is assigned a value in the Kirchhoff for non-bonded contacts of –1 multiplied by its evolutionary coupling score

(i.e., $-1\times EC_{score}$). This will permit that the strength of each connection will attenuate proportionally to the evolutionary coupling strength. The Kirchhoff can be decomposed into the individual contributions from the bonded contacts representing the chain connectivity (Rouse chain) and that from the non-bonded contacts (Bahar *et al.*, 1997). In the sequence GNM the contribution of non-bonded contacts to the Kirchhoff is constructed according to

$$\Gamma_{ij}^{nb} = \begin{cases} -1 \times \mathrm{EC}_{\mathrm{score}}, & i \neq j \quad \text{evolutionary coupling} \\ 0, & \mathrm{i} \neq j \quad \text{no coupling} \\ -\sum_{i,i\neq j} \Gamma_{ij}\,, & i = j \end{cases} \tag{5.1}$$

Similarly, the chain connectivity (Rouse chain) matrix was constructed such that every residue pair *i, i* ± 1 to *i, i* ± 3 are in contact as

$$\Gamma_{ij}^{cc} = \begin{cases} -1, & i \neq j \quad \text{and} \quad \sum_{i|k=1,2,3}^{L} i, i \pm k \\ 0\,, & i \neq j \quad \text{else} \\ -\sum_{i,i\neq j} \Gamma_{ij}\,, & i = j \end{cases} \tag{5.2}$$

Then the overall Kirchhoff is the combination of the two contributions ($\Gamma_{ij} = \Gamma_{ij}^{cc} + \Gamma_{ij}^{nb}$). The vibrational dynamics due to thermal fluctuations can then be evaluated in the same way as the original GNM by inverting the Kirchhoff matrix. The magnitude of mean-square fluctuations is then written in terms of the inverse Kirchhoff as

$$\langle(\Delta \boldsymbol{R}_i)^2\rangle \cong [\Gamma^{-1}]_{ii} \tag{5.3}$$

This is proportional to the Debye-Waller temperature factors or B-factors, which describe the attenuation of X-ray scattering due to the thermal motions of atoms ($B_i = 8\pi^2\langle(\Delta \boldsymbol{R}_i)^2\rangle/3$). Here there is no single-parameter force constant as in the structure GNM (Tirion, 1996), and the pair-wise interactions are simply the strength of the

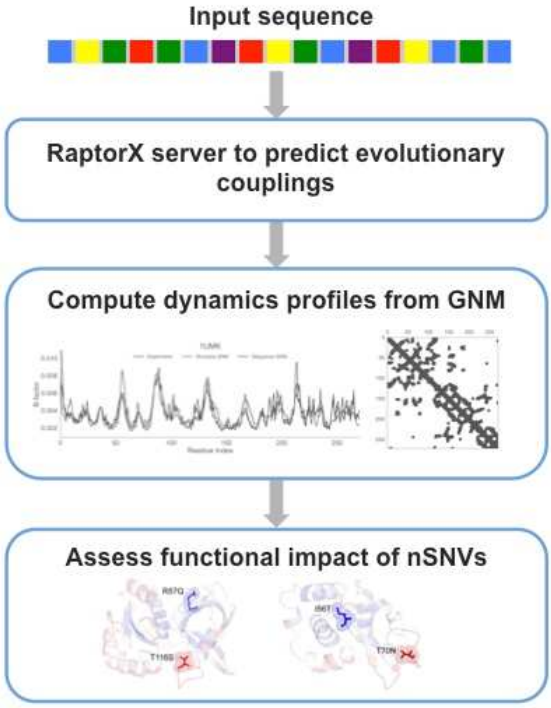

Figure 5.1: A workflow of our method to use predicted evolutionary couplings to determine protein dynamics and assess the functional impact of nSNVs. The initial input is an amino acid sequence, which is given to the RaptorX webserver to predict evolutionary coupling pairs that are used as contacts in the GNM model. Using a GNM, we compute the dynamics profile of each protein, which can give insight into the functional impact of nSNVs. This was done for a curated set of 139 structures.

evolutionary couplings as given by their ranked scores. The theoretical predictions of our sequence GNM can be compared to the predictions of the original structural GNM as well as observed crystallographic B-factors. A general workflow of our method is presented as a flow diagram in Figure 5.1.

5.3 Results and Discussion

5.3.1 Optimizing Threshold Value for EC Scores

Not all of the predicted EC contacts are true 3D contacts, largely due to noisy artifacts in the MSA such as the transitivity of correlations and phylogeny. With this in mind, we decided to accept only the top scoring contact scores predicted by RaptorX. To

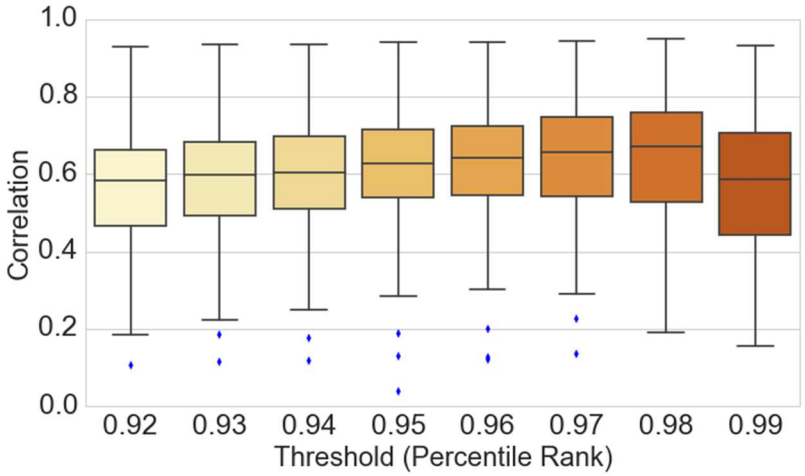

Figure 5.2: A boxplot comparing the correlations of predicted B-factors by the sequence GNM with that of the structural GNM for all 139 structures using a constant threshold for EC contacts. In order to determine the optimal threshold value the GNM analysis was conducted 8 times using a constant threshold (between 0.92 and 0.99) each time. The best correlations were produced when the constant threshold value of 0.98 was used. In this context the average correlation coefficient was 0.63 for all 139 cases.

ensure consistency when analyzing different proteins with varying lengths, we converted the raw scores into a percentile rank. We then computed the sequence GNM for all 139 structures using a constant threshold value, and measured the correlation between the B-factors predicted by our sequence GNM to the original GNM. To determine the optimal threshold value, this procedure was done 8 times using a range of threshold values from 0.92 to 0.99. A threshold value $\leq$0.92 yields superfluous contacts leading to a noisy contact map, and thus, a lower overall correlation (Figure 5.2). Conversely, a threshold value $\geq$0.99 gives a deficient number of contacts, which yields an excessively sparse contact map and a lower overall correlation. As Figure 5.2 shows, a threshold value of 0.98 produced the best overall correlation with the original GNM and, thus, was taken to be the optimal threshold value used in the analysis.

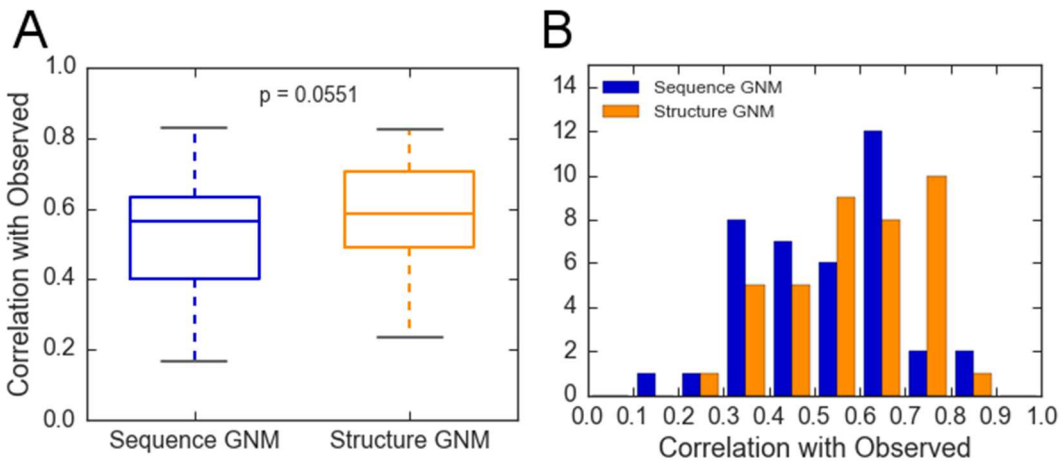

Figure 5.3: Comparing the distribution of correlation coefficients of experimental B-factors with the theoretical B-factors from the sequence GNM and the structure GNM. (A) A boxplot showing the correlation of predicted B-factors by the sequence GNM with experimentally observed B-factors (blue) in comparison to that of the structural GNM (orange) for a subset of 39 structures with resolution better than 2.0 Å. (B) A distribution plot of the same correlations binned in into 10 bins with sizes of 0.1. A student t-test revealed no significant difference between the two distributions (p=0.05508) indicating that the sequence GNM is producing competitive results compared to the original structural GNM. The mean correlation of the sequence GNM is 0.53 while that of the structural GNM is 0.58.

5.3.2 B-factor Correlations: Sequence, Structure, and Experimental

We compared our theoretical predictions using the sequence GNM with the original structural GNM and crystallographic B-factors. Unreliable B-factors are common for many PDB structures. For instance, poor B-factors could be due to low-resolution X-ray diffraction data, the fact that crystal contacts are formed during the experiments that impede motion and do not accurately reflect the protein in the cellular environment, or the temperature of the experiment was not at physiological temperature, which will affect the thermal motions. Thus, for the experimental comparison we extracted a subset of 39 structures that had a resolution better than 2.0 Å to obtain more realistic crystallographic B-factors. The same cutoff of 2.0 Å was used by Kundu et al. to compare GNM predicted B-factors with those determined by crystallography (Kundu *et al.*, 2002). For all 39

structures, the sequence GNM and structure GNM was computed and their estimated B-factors were compared with the observed B-factors, yielding a correlation for each protein. As shown in Figure 5.3A, the boxplot distributions of correlations are similar between the sequence and structure GNM ($p = 0.055$ in a student t-test), with the structure GNM appearing to be slightly superior to the sequence GNM. Figure 5.3B shows the same distribution separated into 10 individual bins of size 0.1. The overall shapes of the two distributions are similar, except for the exaggerated peak of the sequence GNM at 0.4. Interestingly, the sequence GNM has a higher count of correlations between 0.8–0.9 compared to structure GNM by a factor of 2. It should also be noted that for these cases where sequence GNM had low correlations, the EC threshold could be adjusted to yield much higher correlations. If this were done on a case-by-case basis, the overall correlation distribution would be almost identical. Using the threshold as an adjustable parameter to tune the correlation coefficient such that it is maximal is entirely possible. For instance, we could compute the sequence GNM using 5 different threshold values and accept the value that yields the best correlation with experiment. Although this technique would likely produce as good or better results as the structure GNM, the procedure would no longer be *de novo* since it would rely on knowledge of the structure. The mean correlation coefficient for the sequence GNM was 0.53 while the mean correlation coefficient for the structure GNM was 0.58. This is consistent with the findings of Kundu et al. that computed the GNM for 113 high-resolution structures (resolution <2.0 Å) in which the mean correlation coefficient with observed B-factors was also 0.59 (Kundu *et al.*, 2002). The fact that the sequence GNM produces comparable correlation coefficients to the structure GNM, and is superior in the

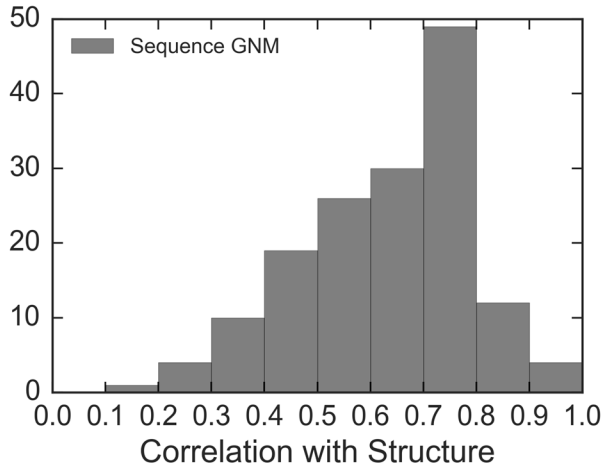

Figure 5.4: The distribution of correlation coefficients between the sequence and structure GNM predicted mean-square fluctuations as computed from 139 structures (listed in table 5.1). The threshold for evolutionary couplings was 0.98 and the average correlation coefficient was 0.63.

range of coefficients between 0.8–0.9, is an impressive result, particularly in that it is a rather crude model to approximate protein dynamics using only a protein sequence.

Even when using high-resolution X-ray structures, there is still some uncertainty about the realistic nature of crystallographic B-factors. For this reason, a more plausible way to determine the efficacy of the sequence GNM is to compare it directly with the structure GNM. The structure GNM is a powerful tool to describe thermal fluctuations in a protein, and in many cases it performs as good or better than the ANM or MD (Doruker *et al.*, 2000; Kundu *et al.*, 2002). We systematically evaluated the sequence and structure GNM for the entire set of 139 structures and obtained the correlation coefficients for each protein (Figure 5.4). The average correlation between the two models is 0.63. As seen in Figure 5.4, the distribution of correlation coefficients increases monotonically until 0.8, and then subsequently decreases. Interestingly, there are still an appreciable number of exceedingly high correlations from 0.8 to 1.0. A distinguishing feature of the distribution

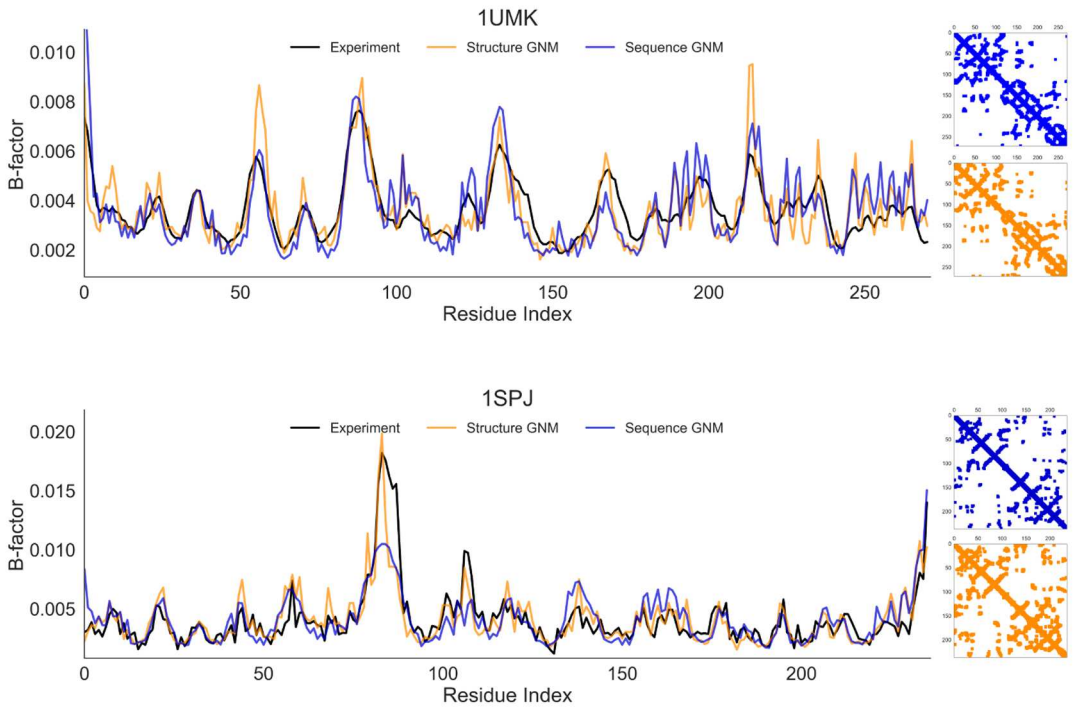

Figure 5.5: A plot of theoretical B-factors as calculated by our sequence GNM (blue), the original GNM (orange), and observed experimental B-factors (black) for two proteins, Human Erythrocyte NADH-cytochrome b5 Reductase (PDB code: 1UMK) and Human Kallikrein 1 (PDB code: 1SPJ). The respective contact maps are shown to the right. These two proteins 1UMK and 1SPJ are adequate for comparing experimental B-factors since they have resolution of 1.75 Å and 1.7 Å respectively, thus their B-factors are likely more reliable. In both cases our sequence GNM produced exceptional correlations with both experimental and structural GNM B-factors. For 1UMK the sequence GNM produced a correlation of 0.78 with experiment, and 0.68 with the structural GNM. For 1SPJ the sequence GNM produced a correlation of 0.77 with experiment, and 0.81 with the structural GNM. Moreover their contact maps reveal the predicted contacts between the sequence and structural GNM approaches are remarkably similar.

is the pronounced peak in the bin from 0.7 to 0.8. The saturation of high correlations between 0.7–0.8 provides evidence that the sequence GNM is clearly capturing protein dynamics. This is indicative of the efficacy of using evolutionary couplings as contacts in the GNM in place of the structure contacts.

We also considered two high-resolution proteins (<2.0 Å), 1UMK and 1SPJ, to examine their B-factor profiles and predicted contact maps. Coevolution analysis using

DCA has been shown to recapitulate accurate structural contact maps for many proteins (Morcos *et al.*, 2011; Marks *et al.*, 2012, 2011). As expected, in both cases the contact maps between sequence and structure GNM were remarkably similar (Figure 5.5). In looking at their B-factor profiles, both sequence and structure GNM exhibit good agreement with observed B-factors, capturing each of the peaks. The correlation between the sequence GNM and observed B-factors is 0.78 for 1UMK and 0.77 for 1SPJ, whereas the correlation between the structure GNM and observed B-factors is 0.79 for 1UMK and 0.83 for 1SPJ. We surmise that the sequence GNM produces exceptional correlations with crystallographic B-factors that are very close to those produced by the structure GNM. Moreover, both theoretical B-factor profiles identified the catalytic sites on these proteins (i.e., low mobility sites).

As a further test to the efficacy of the sequence GNM, we superimposed the predicted B-factors on the structures of three diverse proteins–2JAO, 1FJ2, and 1UMK–to visually contrast the predicted B-factors with that of experiment. Figure 5.6 shows each protein color-coded according to their B-factor profile on a spectrum of blue–white–red, where blue represents the lowest B-factors (less mobility) and red represents the highest B-factors (more mobility). The left panel shows the experimental B-factors for each protein, while the right panel shows the theoretical values predicted by the sequence GNM. We investigated whether secondary structure was a factor in how the B-factors were distributed across the protein, and if certain secondary structure domains would exhibit less agreement with experiment. In this context, the proteins were selected so that they had a variety of secondary structure components–2JAO contains primarily alpha helices, 1UMK is mainly composed of beta-sheets, and 1F2J is a combination of alpha

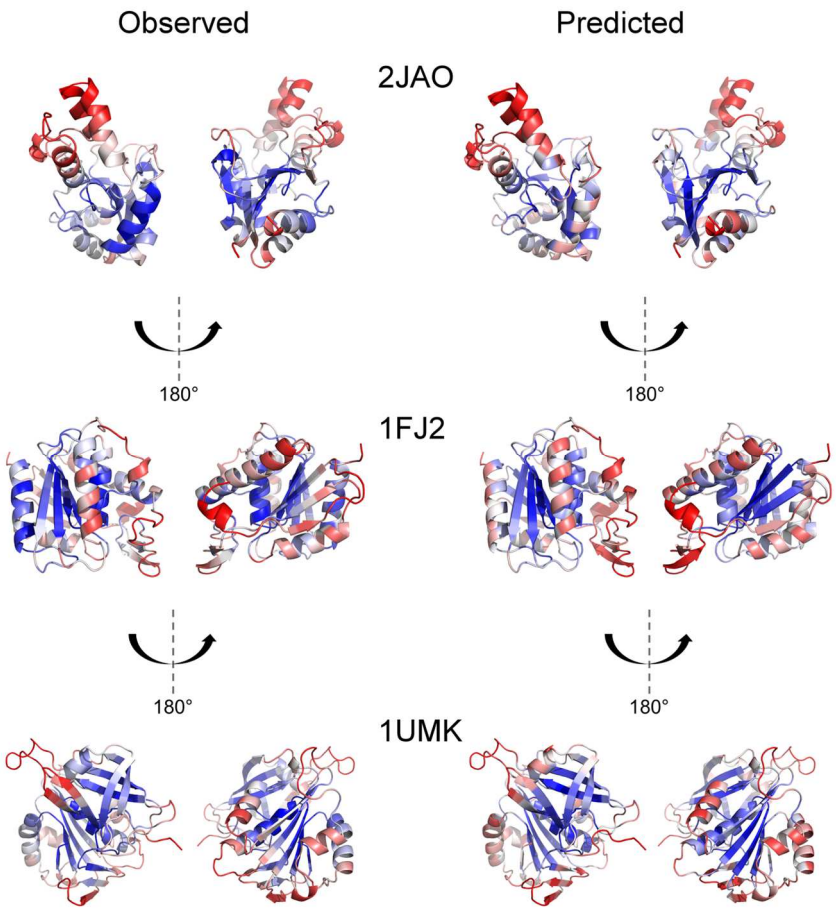

Figure 5.6: The observed crystallographic B-factors (left) and the predicted B-factors from the sequence GNM superimposed on the structure. The three proteins selected–2JAO, 1F2J, and 1UMK–were high-resolution structures better than 2.0 Å. The B-factors are color-coded according to their B-factor profile on a spectrum of blue–white–red where blue represents the lowest B-factors (less mobility) and red represents the highest B-factors (more mobility). The B-factor scores were converted to a percentile rank so that they could be compared across different proteins. Each protein was also rotated 180° so that both sides could be visualized and compared. Moreover, the proteins were selected so that they had a variety of secondary structure components–2JAO contains primarily alpha helices, 1UMK is mainly composed of beta-sheets, and 1F2J is a combination of alpha helices and beta-sheets.

helices and beta-sheets. For 2JAO, the exterior helices that are flexible (red) in the observed structure are all reproduced in the predicted structure. The one highly rigid (blue) helix in the observed structure was more flexible in the predicted structure, but was

still in overall agreement. 1FJ2 shows a remarkable agreement between observed and predicted structures, particularly considering that it is a combination of alpha helix and beta-sheet elements. Similarly, 1UMK showed excellent agreement aside from some very miniscule differences. Overall, this gives further credence to the efficacy of our sequence GNM model, as it is capable of recapitulating crystallographic B-factor profiles for many cases.

5.3.3 Assessing nSNV Phenotypes Using the Sequence GNM

Crystallographic B-factors have previously been used to assess the impact of nSNVs on protein function (Chasman and Adams, 2001; Adzhubei *et al.*, 2010; Alber *et al.*, 1987; Nevin Gerek *et al.*, 2013). A study by Alber et al. found that mutations on lysozyme that impaired function exhibited lower than average temperature factors, suggesting that rigid sites on the protein are more susceptible to destabilizing nSNVs than flexible sites (Alber *et al.*, 1987). Another study also revealed a relationship between crystallographic B-factors and the impact of nSNVs on protein function (Chasman and Adams, 2001). A commonly used tool to diagnose nSNVs, PolyPhen-2, that uses machine learning coupled with evolutionary information and structural information uses crystallographic B-factors in their predictions (Adzhubei *et al.*, 2010). In essence, these studies indicate that crystallographic B-factors can be used to predict the tolerance of a given residue to an nSNV (i.e., whether or not the occurrence of an nSNV would impact function).

We investigated whether B-factors predicted by the sequence GNM was indicative of biological phenotype for nSNVs in the human population. A total of 738 nSNVs were mapped to the 139 enzymes, in which 436 are associated with disease and

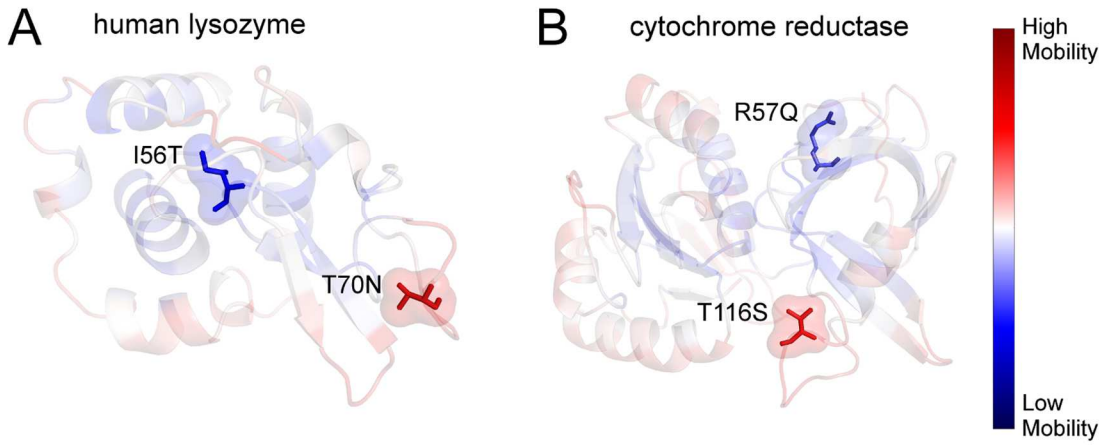

Figure 5.7: A ribbon diagram for two human enzymes, human lysozyme (A) and cytochrome reductase (B) colored according to their predicted B-factors by the sequence GNM. Red indicates high mobility sites, and blue indicates low mobility sites. Each protein contains two known nSNVs. I56T and R57Q are disease-associated, and they occur on low mobility (rigid) sites. Conversely, the neutral nSNVs T116S and T70N occur on high mobility sites.

302 are neutral. Table 5.1 shows the number of disease and neutral nSNVs that occur on each protein. The sequence GNM was computed systematically for all 139 enzymes to obtain their dynamics profiles. The theoretical B-factors scores were converted into a percentile rank so that the values could be compared across different proteins.

We initially looked at two human enzymes, human lysozyme (PDB: 1C7P) and human cytochrome reductase (PDB: 1UMK). They were chosen because they were short proteins with each containing a disease and neutral nSNV. Human lysozyme is a glycoside hydrolase that functions in the immune system by causing damage to cell walls of bacteria. Human cytochrome b5 reductase is involved in many oxidation/reduction reactions including converting methemoglobin to hemoglobin (Elahian *et al.*, 2012). Each structure is color-coded according to its theoretical B-factor profile on a spectrum of blue–white–red. Sites that exhibit high mobility (flexible) are red, and sites that have low

mobility (rigid) are blue. Regions that are characterized by low mobility are usually important for maintaining stability and function, thus a mutation could act to destabilize the protein and impair its function. Figure 5.7A shows the disease mutation I56T occurring on a rigid site with a B-factor of 0.0075. The neutral mutation T70N has a B-factor of 0.96 indicating that it is a highly mobile site. Both I56T and T70N occur on loop regions. Although loops are generally more flexible, three alpha helical domains encompass the loop containing I56T, which implies that it may be involved in interactions that contribute to stabilizing the functional conformation. Thus, the I56T mutation may disrupt these critical interactions and impair the enzymatic function. In the case of cytochrome reductase (Figure 5.7B), the disease mutation R57Q is also on a rigid site with a B-factor of 0.14. Instead of being located near the core, R57Q is highly exposed protruding outwardly from a beta-barrel. However, since beta-barrels often harbor functional residues, the R57Q mutation may disrupt certain interactions critical for modulating function. The neutral mutation T116S is located on a loop and has a B-factor of 0.96, indicating that is it has a high mobility. Sites that are highly flexible (e.g., loop regions, or superficial sites) are more robust to mutations. Conversely, rigid sites are more susceptible to mutations that significantly impact function. For these two cases, the theoretical B-factors produced by the sequence GNM convincingly discriminated between the disease and neutral nSNVs.

The findings based on the two enzymes encouraged us to look at a proteome-wide set of 139 enzymes to determine whether the distribution of B-factors for all 436 disease and 302 neutral nSNVs was predictive. The B-factor scores were ranked into percentiles (%B-factor) and then binned into 5 bins of size 0.2. Then we computed the observed-to-

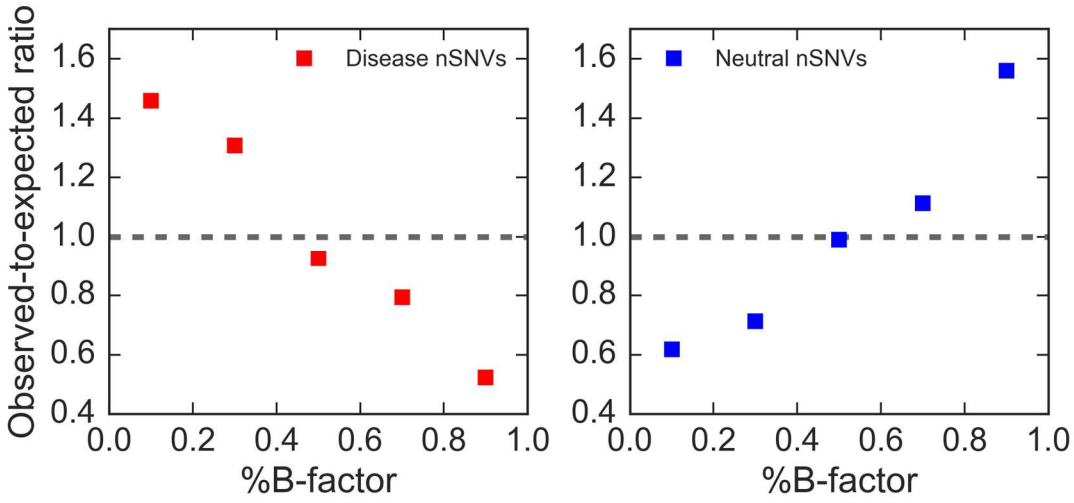

Figure 5.8: The relationship of observed-to-expected numbers between 436 disease nSNVs (red) and 302 neutral nSNVs (blue) from 139 human enzymes. The %B-factor scores derived from the sequence GNM are binned into 5 bins of size 0.2.

expected ratio of B-factors, where the expected values were based on the B-factor distribution of all 51,618 sites across all 139 proteins, and the observed values were based on the B-factors of the 436 disease sites. The same process was done for the 302 neutral nSNVs. Then under the null hypothesis of no effect, the ratio of expected and observed sites harboring disease mutations for each %B-factor bin should be close to 1. A strong relationship between disease sites and the %B-factor score would reject the null hypothesis that disease sites are distributed uniformly between sites with low and high mobility. For the 436 disease nSNVs, this null hypothesis was rejected ($p < 0.001$). Figure 5.8 shows the observed-to-expected ratio plot of disease and neutral nSNVs, which indicates that disease nSNVs are overabundant at low %B-factor sites (<0.4) and underabundant at high %B-factor sites. Conversely, neutral nSNVs are overabundant at high %B-factor sites (>0.6) and underabundant at low %B-factor sites. This strong evidence purports that the occurrence of an nSNV on a site with a low B-factor is likely damaging, whereas an nSNV on a high B-factor site is likely to be benign. Low B-factors

usually signify a residue that is crucial for modulating functional motions (e.g., a hinge) and high B-factor sites are more flexible sites such as loops that are more robust to mutations. Figure 5.8 reveals the ability of predicted B-factors to discriminate between disease and neutral nSNVs using only sequence as input. The distributions in Figure 5.8 are remarkably similar to an observed-to-expected ratio plot in a previous DFI analysis of nSNV phenotype (Nevin Gerek *et al.*, 2013).

CHAPTER 6

6 CONCLUSION

6.1 Summary of Current Work

This thesis has presented research on a novel method to explore the role of protein conformational dynamics in genomic variation. Some recent studies have evinced that molecular evolution utilizes dynamics as a means to control protein function (Haliloglu and Bahar, 2015; Tawfik and Tokuriki, 2009; Zou *et al.*, 2015; Kim *et al.*, 2015; Glembo *et al.*, 2012; Bhabha *et al.*, 2013; Campbell *et al.*, 2016; Nevin Gerek *et al.*, 2013). Over the course of evolution, changes in the conformational landscapes of a protein allow for development of new functions or optimization of existing functions (e.g., enzyme function), which suggests that molecular evolution and protein dynamics are intertwined (Campbell *et al.*, 2016). One evolutionary mechanism that induces changes in protein dynamics is mutation. With advanced sequencing technologies it is now known that each exome contains many thousands of single nucleotide variants (nSNVs) that can be harmful to function. Detecting which variants will lead to human disease is of paramount importance, and most current methods involve using evolutionary information. The evolution-based methods, which depend largely on amino acid conservation, are able to diagnose nSNVs for highly conserved positions but often fail to predict disease-associated nSNVs at less conserved positions or neutral nSNVs at highly conserved positions. Efforts to include structural information resulted in only a minimal increase in predictive power, principally because of the use of a static structure. We demonstrated

the importance of including protein conformational dynamics, which plays a significant role in nSNV phenotype.

Proteins usually perform their function by forming complexes biomolecules that act in conjunction. The interactions between these complexes at protein-protein interfaces is very crucial to their ability to function, thus nSNVs that occur at specific interface sites are highly susceptible to disease. We used a novel metric called the dynamic flexibility index (DFI) that measures the contribution of site-specific conformational dynamics to functional importance. It predicts which sites are flexible and which are more rigid (i.e., act as hinges). We demonstrated in a proteome-wide analysis using the DFI metric that interface sites are less flexible than non-interface sites, and that particularly low flexibility sites (low DFI) at interfaces are more likely to harbor disease nSNVs. Indeed, we observed that disease nSNVs at interface sites had appreciably low DFI, whereas neutral nSNVs had higher DFI. The DFI metric can quantify which sites contribute the most to function (i.e., low DFI sites are more susceptible to damaging nSNVs), and it can also identify sites that are less important for function (i.e., high DFI sites are more robust to damaging nSNVs). Thus, our dynamics-based metric was able to discriminate disease and neutral nSNVs at interface sites. Moreover, we observed that accessible surface area (ASA), a metric based on static structure commonly used to study nSNV phenotypes, was inferior to DFI in discriminating disease and neutral nSNVs. This emphasizes the importance of considering protein dynamics in these types of analysis. Finally, we showed that even for the regime of fast-evolving amino acid positions, where evolutionary methods tend to be inadequate, DFI was still able to discriminate nSNV

phenotypes. These findings point toward the potential efficacy of dynamics-based metrics like DFI in genomic analysis of nSNVs.

The concept of allostery (i.e., action at a distance) has been studied extensively recently for its role in protein function regulation and application in drug design. Certain sites on a protein are distal to crucial functional sites (e.g., active sites or binding sites) but can still have an impact on those sites and, thus, affect the functional activity of the protein. We developed a new dynamics-based metric called the dynamic coupling index (DCI), which can identify such sites (called DARC spots) that are distal to active sites but are dynamically coupled to them. These DARC spots are sites that are not located in close proximity to functional sites but are still important for allosteric regulation and function of the protein. Through DCI, we analyzed the enzyme GCase, which has over 200 mutations that lead to Gaucher disease (GD) in humans. The disease associated mutations lead to a drastic decrease in catalytic efficiency that disrupts the function of the enzyme. Despite the hundred years of research (i.e., since it was discovered in the late 19th century), GD has neither a cure or an accepted mechanism for the disease (Lieberman, 2011; Hruska *et al.*, 2008; Beutler *et al.*, 1993; Lieberman *et al.*, 2007). Through DFI we analyzed four common GD mutations that evolutionary-based methods usually misdiagnose as neutral. We observed that each mutation leads to a significant loss in flexibility at two highly functional catalytic sites and ligand binding recognition sites, both responsible for enzymatic function. This loss in flexibility (i.e., rigidification) restricts these sites from modifying their structural conformations to attain the required orientations for accommodating a substrate for catalytic function, which results in lower catalytic efficiency. Moreover, a DCI analysis of the four mutations revealed that each

mutation resulted in a global loss in allosteric coupling between remote sites and the active sites. The breaking of crucial allosteric regulation networks between remote sites and the active sites is catastrophic for enzymatic function. Thus, these findings provide a plausible mechanism for GD based on conformational dynamics. In an analysis of all other ~200 GD-related mutations, the DCI metric revealed that the majority of all GD occur at high DCI sites (i.e., DARC spots). Since this appeared to be a general trend, we conducted a proteome-wide analysis of enzymes, in which we demonstrated that disease nSNVs were largely located at DARC spots. Thus, the DCI metric is shown to be able to identify important sites (DARC spots) that are more susceptible to damaging nSNVs. We also showed that DFI and DCI as a combined metric exhibited remarkable predictive power for cases where evolutionary metrics were lacking predictive power in nSNV phenotypes. This further illustrates the obvious utility of conformational dynamics-based metrics (e.g., DFI and DCI) as *in silico* tools in genomic analysis to systematically quantify and assess nSNVs in terms of their disease risk.

Investigating protein conformational dynamics is contingent on the knowledge of a 3D native structure. While there is a large body of available experimental structures in the Protein Data Bank (PDB), there is still a disproportionate amount of readily available sequence data compared to known structures. This wealth of data is a result of advanced, high-throughput sequencing technologies and is projected to further increase at an exponential rate, continuing to outpace the much slower determination of experimental structures. Since our overall goal is to integrate conformational dynamics into genomic analysis, the inherent limitations of this must be addressed. The ability to obtain site-specific conformational dynamics is dependent on the known 3D structure. This begs the

question: how can protein dynamics be used in genome-wide analysis to predict functional impacts of nSNVs? In particular, the evolutionary methods that use sequence information are able to predict the consequences of nSNVs where the structure is unknown. For this reason, there is a need to be able to obtain protein dynamics by leveraging only sequence information, without *a priori* knowledge of a 3D structure. This ability would be a tremendous achievement considering the abundance of sequence data. We have developed a novel method to estimate the dynamics profile of a protein, using only a sequence as input. The method uses the coevolution of amino acids in an alignment of sequences (which tend to be spatially close in the 3D tertiary structure) and a simple Gaussian network model (GNM) to obtain dynamics. The original GNM based on the 3D structure is well-known for its intrinsic ability to describe residue dynamics profiles due to thermal motions in proteins (i.e., B-factors). We showed that our sequence-based GNM model was able to recapitulate the mean-square fluctuations (B-factors) produced by the original GNM. Our estimates of B-factors for a proteome-wide set of proteins exhibited an impressive correlation with the structure GNM. Moreover, our estimated B-factors were in good agreement with crystallographic B-factors for many cases. To address the issue of how protein dynamics can assess the impact of nSNVs occurring across the genome where there are no known 3D structures, we tested the ability of our predicted dynamics from the sequence GNM to assess nSNV phenotypes. For a large set of 738 nSNVs, the predicted B-factors using the sequence GNM was able to discriminate between disease and neutral nSNVs. A plot of the observed-to-expected ratio of the predicted B-factors revealed distributions of disease and neutral nSNVs that are remarkably similar to those in a previous DFI analysis (Nevin Gerek *et al.*, 2013).

This preliminary analysis shows that the sequence GNM approach makes it now possible to obtain estimates of dynamics without the use of a 3D structure, which allows for the plausible integration of conformational dynamics into large-scale analysis of genomic variants.

6.2 Future Directions

The research on protein conformational dynamics presented in this thesis can be exploited in the field of genomic medicine. Many of the current approaches to developing novel therapeutic drugs to treat diseases employ the concept of allostery. The ability to observe changes in conformational dynamics of specific regions of a protein in response to an event at a distance location (e.g. the binding of a substrate) is valuable for the development of novel drugs that act on proteins using allostery. Our dynamics-based metrics (DFI and DCI) make it possible to monitor changes in site-specific conformational dynamics due to external perturbations such as mutations or approaching ligands. In addition, they allow for identification of crucial allosteric regulation networks that are critical for maintaining protein function. Thus, site-specific conformational dynamics can be leveraged to investigate potential drug targets for novel therapeutic strategies. It has applications to genomics to systematically quantify and assess sites on a protein that are more, or less susceptible to harboring a harmful nSNV that may lead to disease. Moreover, the ability to estimate dynamics without a structure expands the breadth of its utility across the entire genome, such that it can compliment evolutionary methods regardless of whether the structure is known. As the cost of sequencing individual genomes is rapidly decreasing, complimentary metrics to assess human variations are especially enticing. Overall, the *in silico* tools proposed in this thesis can be

used with current approaches in genomic medicine to increase the efficacy of systematic assessments of an individual's disease risk based on their genome, as well as determine specialized treatment plans in personalized medicine that are tailored to each individual.

The characterization of 3D folded structures using through experimental methods (e.g., X-ray crystallography) remains essential for mechanistic understanding biological function of proteins (Shi, 2014). A problem in this area is that many crystal structures are low-resolution (e.g., worse than 4 Å) as they exhibit weak diffraction patterns, resulting in poor quality electron density maps. Methods to refine these low-resolution structures have been proposed, including using known homologous structures as a constraint to enhance the quality of electron density maps (Schröder *et al.*, 2010). Moreover, recent advances in single-particle electron cryo-microscopy (cryo-EM) is showing promise for obtaining higher resolution structures and complementing crystallography (Cheng, 2015). Our sequence-GNM method, which only requires an amino acid sequence to estimate the thermal motions of C-alpha atoms, may be used as an additional model to fit X-ray diffraction data to enhance the otherwise poor quality electron density maps. Thus, this could be applied to further complement current refinement methods for low-resolution diffraction data.

REFERENCES

Abdizadeh,H. *et al.* (2015) Perturbation response scanning specifies key regions in subtilisin serine protease for both function and stability. *J. Enzyme Inhib. Med. Chem.*, **30**, 867–873.

Abecasis,G.R. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Adzhubei,I. *et al.* (2013) Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr Protoc Hum Genet*, **2**.

Adzhubei,I. a *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–9.

Alber,T. (1989) Mutational effects on protein stability. *Annu. Rev. Biochem.*, **58**, 765–798.

Alber,T. *et al.* (1987) Temperature-sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. *Biochemistry*, **26**, 3754–8.

Atilgan,A.R. *et al.* (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, **80**, 505–15.

Atilgan,C. *et al.* (2010) Manipulation of conformational change in proteins by single-residue perturbations. *Biophys. J.*, **99**, 933–43.

Atilgan,C. and Atilgan,A.R. (2009) Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS Comput. Biol.*, **5**, e1000544.

Bagci,Z. *et al.* (2002) Residue packing in proteins: Uniform distribution on a coarse-grained scale. *J. Chem. Phys.*, **116**, 2269–2276.

Bahar,I. *et al.* (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, **2**, 173–81.

Bahar,I. *et al.* (2010) Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.*, **39**, 23–42.

Bahar,I. and Rader, a J. (2005) Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.*, **15**, 586–92.

Bao,L. *et al.* (2005) nsSNPAnalyzer: Identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.*, **33**, 480–482.

Bendl,J. *et al.* (2014) PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. *PLoS Comput. Biol.*, **10**, 1–11.

Berliner,N. *et al.* (2014) Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLoS One*, **9**.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bernstein,F.C. *et al.* (1977) The Protein Data Bank : A Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.*, **112**, 535–542.

Beutler,E. *et al.* (1993) Gaucher disease: gene frequencies in the Ashkenazi Jewish population. *Am. J. Hum. Genet.*, **52**, 85–88.

Beutler,E. *et al.* (2005) Hematologically important mutations: Gaucher disease. *Blood Cells. Mol. Dis.*, **35**, 355–364.

Bhabha,G. *et al.* (2011) A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis. *Science*, **332**, 234–238.

Bhabha,G. *et al.* (2013) Divergent evolution of protein conformational dynamics in dihydrofolate reductase. *Nat. Struct. Mol. Biol.*, **20**, 1243–1249.

Bogan,A.A. and Thorn,K.S. (1998) Anatomy of Hot Spots in Protein Interfaces. *J. Mol. Biol.*, **280**, 1–9.

Bolia,A. and Ozkan,S.B. (2016) Adaptive BP-Dock: An Induced Fit Docking Approach for Full Receptor Flexibility. *J. Chem. Inf. Model.*, acs.jcim.5b00587.

Boulton,S. and Melacini,G. (2016) Advances in NMR Methods To Map Allosteric Sites: From Models to Translation. *Chem. Rev.*, **116**, 6267−6304.

Bromberg,Y. *et al.* (2013) Neutral and weakly nonneutral sequence variants may define individuality. *Proc. Natl. Acad. Sci.*, **110**, 14255–14260.

Bromberg,Y. and Rost,B. (2007) SNAP: Predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.*, **35**, 3823–3835.

Brooks,B. *et al.* (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci U S A*, **80**, 6571–6575.

Burger,L. and Van Nimwegen,E. (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.*, **6**.

Butler,B.M. *et al.* (2015) Conformational dynamics of nonsynonymous variants at protein

interfaces reveals disease association. *Proteins*, **83**, 428–435.

Calabrese,R. *et al.* (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.*, **30**, 1237–1244.

Campbell,E. *et al.* (2016) The role of protein dynamics in the evolution of new enzyme function. *Nat. Chem. Biol.*

Capriotti,E. *et al.* (2004) A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics*, **20**, 63–68.

Capriotti,E., Fariselli,P., and Casadio,R. (2005) I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, 306–310.

Capriotti,E., Fariselli,P., Calabrese,R., *et al.* (2005) Predicting protein stability changes from sequences using support vector machines. *Bioinformatics*, **21**, 54–58.

Capriotti,E. *et al.* (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, **22**, 2729–2734.

Capriotti,E. and Altman,R.B. (2011) Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics*, **12**, S3.

Chamary,J. V *et al.* (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.*, **7**, 98–108.

Chang,J.C. and Kan,Y.W. (1979) Beta 0 Thalassemia, a Nonsense Mutation in Man. *Proc. Natl. Acad. Sci. U. S. A.*, **76**, 2886–2889.

Chang,K. *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

Chasman,D. and Adams,R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.

Cheng,J. *et al.* (2006) Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, **62**, 1125–1132.

Cheng,T. *et al.* (2008) Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS Comput. Biol.*, **4**.

Cheng,Y. (2015) Single-particle Cryo-EM at crystallographic resolution. *Cell*, **161**, 450–

457.

Cline,M. and Karchin,R. (2010) Using bioinformatics to predict the functional impact of SNVs. *Bioinformatics*, **27**, 441–448.

Cover,T. and Thomas,J. (2006) Elements of information theory Wiley, New York.

Cui,Q. and Bahar,I. (2005) Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems Chapman and Hall/CRC.

David,A. *et al.* (2012) Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum. Mutat.*, **33**, 359–63.

Davydov,E. V. *et al.* (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**.

Dehouck,Y. *et al.* (2013) BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Res.*, **41**, 333–339.

Dehouck,Y. *et al.* (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*, **12**, 151.

Dobson,R.J. *et al.* (2006) Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. *BMC Bioinformatics*, **7**, 217.

Doruker,P. *et al.* (2000) Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins*, **40**, 512–524.

Dudley,J.T. *et al.* (2012) Human genomic disease variants: a neutral evolutionary explanation. *Genome Res.*, **22**, 1383–94.

Dunn,S.D. *et al.* (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Echave,J. (2008) Evolutionary divergence of protein structure: The linearly forced elastic network model. *Chem. Phys. Lett.*, **457**, 4–6.

Echave,J. and Fernandez,F. (2009) A perturbative view of protein structural variation. *Proteins*, **78**, 173–80.

Eisenmesser,E. *et al.* (2002) Enzyme dynamics during catalysis. *Science (80-.).*, **295**, 1520–3.

Eisenmesser,E. *et al.* (2005) Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, **438**, 117–121.

Ekeberg,M. *et al.* (2013) Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, **87**, 1–16.

Elahian,F. *et al.* (2012) Human cytochrome b5 reductase: structure, function, and potential applications. *Crit. Rev. Biotechnol.*, **8551**, 1–11.

Eswar,N. *et al.* (2006) Comparative protein structure modeling using Modeller.

Eyal,E. *et al.* (2011) Cooperative dynamics of proteins unraveled by network models. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **1**, 426–439.

Finn,L.S. *et al.* (2000) Severe type II Gaucher disease with ichthyosis, arthrogryposis and neuronal apoptosis: molecular and pathological analyses. *Am. J. Med. Genet.*, **91**, 222–226.

Fitch,W.M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406–416.

Flanigan,K.M. *et al.* (2011) Nonsense mutation-associated Becker muscular dystrophy: Interplay between exon definition and splicing regulatory elements within the DMD gene. *Hum. Mutat.*, **32**, 299–308.

Flory,P.J. (1976) Statistical thermodynamics of random networks. *Proc. R. Soc. London*, **351**, 351–380.

Forbes,S.A. *et al.* (2011) COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **39**, 945–950.

Franzosa,E. a and Xia,Y. (2009) Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.*, **26**, 2387–95.

Furnham,N. *et al.* (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, D485–D489.

Gerek,Z.N. and Ozkan,S.B. (2011) Change in Allosteric Network Affects Binding Affinities of PDZ Domains: Analysis through Perturbation Response Scanning. *PLoS Comput. Biol.*, **7**, e1002154.

Germain,D. (2004) Gaucher's disease: a paradigm for interventional genetics: Gaucher's disease. *Clin. Genet.*, **65**, 77–86.

Glembo,T. *et al.* (2012) Collective Dynamics Differentiates Functional Divergence in

Protein Evolution. *PLoS Comput. Biol.*, **8**.

Go,N. *et al.* (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci.*, **80**, 3696–3700.

Gouveia-Oliveira,R. and Pedersen,A.G. (2007) Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms Mol. Biol.*, **2**, 12.

Goymer,P. (2007) Synonymous mutations break their silence. *Nat. Rev. Genet.*, **8**, 92–92.

Grabowski,G.A. (2004) Gaucher disease: lessons from a decade of therapy. *J. Pediatr.*, **144**, S15–S19.

Grant,M. a *et al.* (2007) Familial Alzheimer's disease mutations alter the stability of the amyloid beta-protein monomer folding nucleus. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 16522–7.

Grantham,R. (1974) Amino Acid Difference Formula to Help Explain Protein Evolution. *Science (80-.).*, **185**, 862–864.

Green,E.D. and Guyer,M.S. (2011) Charting a course for genomic medicine from base pairs to bedside. *Nature*, **470**, 204–213.

Guerois,R. *et al.* (2002) Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.

Gunasekaran,K. *et al.* (2004) Is allostery an intrinsic property of all dynamic proteins? *Proteins*, **57**, 433–43.

Guo,J. and Zhou,H.-X. (2016) Protein Allostery and Conformational Dynamics. *Chem. Rev.*, **116**, 6503−6515.

Haliloglu,T. *et al.* (1997) Gaussian Dynamics of Folded Proteins. *Phys. Rev. Lett.*, **79**, 3090–3093.

Haliloglu,T. and Bahar,I. (2015) Adaptability of protein structures to enable functional interactions and evolutionary implications. *Curr. Opin. Struct. Biol.*, **35**, 17–23.

Hamosh,A. *et al.* (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, 514–517.

Henrissat,B. (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.*, **280**, 309–316.

Hinsen,K. (1998) Analysis of domain motions by approximate normal mode calculations. *Proteins*, **33**, 417–429.

Hinsen,K. *et al.* (1999) Analysis of domain motions in large proteins. *Proteins Struct. Funct. Genet.*, **34**, 369–382.

Hinsen,K. *et al.* (2000) Harmonicity in slow protein dynamics. *Chem. Phys.*, **261**, 25–37.

Hopf,T. a *et al.* (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**, 1607–21.

Hopf,T.A. *et al.* (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*, **3**, e03430.

Hruska,K.S. *et al.* (2008) Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). *Hum. Mutat.*, **29**, 567–583.

Huang,T., Shi,X.-H., *et al.* (2010) Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. *PLoS One*, **5**.

Huang,T., Wang,P., *et al.* (2010) Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. *PLoS One*, **5**.

Ikeguchi,M. *et al.* (2005) Protein structural change upon ligand binding: Linear response theory. *Phys. Rev. Lett.*, **94**, 1–4.

Jackson,C. *et al.* (2009) Conformational sampling, catalysis, and evolution of the bacterial phosphotriesterase. *Proc. Natl. Acad. Sci.*, **106**, 21631–21636.

Jana,B. *et al.* (2014) From structure to function: the convergence of structure based models and co-evolutionary information. *Phys. Chem. Chem. Phys.*, **16**, 6496–6507.

Jones,D.T. *et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–90.

Jordan,D.M. *et al.* (2010) Human allelic variation: perspective from protein function, structure, and evolution. *Curr. Opin. Struct. Biol.*, **20**, 342–350.

de Juan,D. *et al.* (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–61.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kalodimos,C. (2012) Protein function and allostery: a dynamic relationship. *Ann. N. Y. Acad. Sci.*, **1260**, 81–86.

Kamisetty,H. *et al.* (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 15674–9.

Katsonis,P. *et al.* (2014) Single nucleotide variations : Biological impact and theoretical interpretation. *Protein Sci.*, **23**, 1650–1666.

Kent,W.J. *et al.* (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

Keskin,O. *et al.* (2008) Principles of Protein−Protein Interactions: What are the Preferred Ways For Proteins To Interact? **108**, 1225–1244.

Khan,S. and Vihinen,M. (2010) Performance of protein stability predictors. *Hum. Mutat.*, **31**, 675–684.

Kim,D.E. *et al.* (2014) One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins Struct. Funct. Bioinforma.*, **82**, 208–218.

Kim,H. *et al.* (2015) A hinge migration mechanism unlocks the evolution of green-to-red photoconversion in GFP-like proteins. *Structure*, **23**, 34–43.

Kimchi-sarfaty,A.C. *et al.* (2007) A 'Silent' Polymorphism in the MDR1 Gene Changes Substrate Specificity. *Science (80-.).*, **315**, 525–528.

Kloczkowski,A. *et al.* (1989) Chain Dimensions and Fluctuations in Random Elastomeric Networks. 1. Phantom Gaussian Networks in the Undeformed State. *Macromolecules*, **22**, 1423–1432.

Kowarz,L. *et al.* (2005) Gaucher mutation N188S is associated with myoclonic epilepsy. *Hum. Mutat.*, **26**, 271–275.

Krissinel,E. and Henrick,K. (2005) Detection of Protein Assemblies in Crystals. *Comput. Life Sci.*, **LNBI 3695**, 163–174.

Krissinel,E. and Henrick,K. (2007) Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.*, **372**, 774–797.

Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D. Biol. Crystallogr.*, **60**, 2256–68.

Kumar,A., Butler,B.M., *et al.* (2015) Integration of structural dynamics and molecular evolution via protein interaction networks: a new era in genomic medicine. *Curr. Opin. Struct. Biol.*, **35**, 135–142.

Kumar,A., Glembo,T.J., *et al.* (2015) The Role of Conformational Dynamics and Allostery in the Disease Development of Human Ferritin. *Biophys. J.*, **109**, 1273–1281.

Kumar,P. *et al.* (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–81.

Kumar,S. *et al.* (2012) Evolutionary diagnosis method for variants in personal exomes. *Nat. Methods*, **9**, 855–6.

Kumar,S. *et al.* (2011) Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends Genet.*, **27**, 377–386.

Kumar,S. *et al.* (2009) Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. *Genome Res.*, **19**, 1562–9.

Kundu,S. *et al.* (2002) Dynamics of Proteins in Crystals: Comparison of Experiment with Simple Models. *Biophys. J.*, **83**, 723–732.

Levitt,M. (1983) Molecular dynamics of native protein: I. Computer simulation of trajectories. *J. Mol. Biol.*, **168**, 595–617.

Li,B. *et al.* (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, **25**, 2744–2750.

Li,M. *et al.* (2014) Predicting the Impact of Missense Mutations on Protein−Protein Binding Affinity. *J. Chem. Theory Comput.*, **10**, 1770–1780.

Li,M.J. *et al.* (2012) GWASdb: A database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **40**, 1047–1054.

Li,M.X. *et al.* (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res.*, **40**.

Li,Y. *et al.* (2011) Predicting disease-associated substitution of a single amino acid by analyzing residue interactions. *BMC Bioinformatics*, **12**.

Li,Z. *et al.* (2015) A Rigid Hinge Region Is Necessary for High-Affinity Binding of Dimannose to Cyanovirin and Associated Constructs. *Biochemistry*, **54**, 6951–6960.

Liberles,D. *et al.* (2012) The interface of protein structure, protein biophysics, and

molecular evolution. *Protein Sci.*, **21**, 769–785.

Lieberman,R.L. (2011) A Guided Tour of the Structural Biology of Gaucher Disease: Acid- β -Glucosidase and Saposin C. *Enzyme Res.*, **2011**, 1–15.

Lieberman,R.L. *et al.* (2007) Structure of acid β-glucosidase with pharmacological chaperone provides insight into Gaucher disease. *Nat. Chem. Biol.*, **3**, 101–107.

Lisi,G.P. and Loria,J.P. (2016) Solution NMR Spectroscopy for the Study of Enzyme Allostery. *Chem. Rev.*, **116**.

Liu,J. and Nussinov,R. (2016) Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLOS Comput. Biol.*, **12**, e1004966.

Liu,L. *et al.* (2015) A Molecular Evolutionary Reference for the Human Variome. *Mol. Biol. Evol.*, **33**, msv198.

Liu,Y. *et al.* (2010) Role of Hsp70 ATPase domain intrinsic dynamics and sequence evolution in enabling its functional interactions with NEFs. *PLoS Comput. Biol.*, **6**.

Liu,Y. and Bahar,I. (2012) Sequence evolution correlates with structural dynamics. *Mol. Biol. Evol.*, **29**, 2253–2263.

Lori,C. *et al.* (2013) Effect of Single Amino Acid Substitution Observed in Cancer on Pim-1 Kinase Thermodynamic Stability and Structure. *PLoS One*, **8**.

Luu,T.D. *et al.* (2012) MSV3d: Database of human MisSense variants mapped to 3D protein structure. *Database*, **2012**, 1–8.

Ma,J. *et al.* (2015) Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics*, **31**, 3506–3513.

Marks,D.S. *et al.* (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation. *Structure*, **6**.

Marks,D.S. *et al.* (2012) Protein structure prediction from sequence variation. *Nat. Biotechnol.*, **30**, 1072–80.

Marth,G.T. *et al.* (2011) The functional spectrum of low-frequency coding variation. *Genome Biol.*, **12**, R84.

Martin,D. *et al.* (2012) Dissipative electro-elastic network model of protein electrostatics. *Phys. Biol.*, **9**.

Masso,M. and Vaisman,I.I. (2010) AUTO-MUTE: Web-based tools for predicting stability changes in proteins due to single amino acid replacements. *Protein Eng.*

Des. Sel., **23**, 683–687.

McCammon,A.J. *et al.* (1977) Dynamics of folded proteins. *Nature*, **265**, 585–590.

Miller,M.P. and Kumar,S. (2001) Understanding human disease mutations through the use of interspecific genetic variation. *Hum.Mol.Genet.*, **10**, 2319–2328.

Mills,P. *et al.* (2005) Neonatal epileptic encephalopathy caused by mutations in the PNPO gene encoding pyridox(am)ine 5′-phosphate oxidase. *Hum. Mol. Genet.*, **14**, 1077–1086.

Moore,D.J. *et al.* (2003) A missense mutation (L166P) in DJ-1, linked to familial Parkinson's disease, confers reduced protein stability and impairs homo-oligomerization. *J. Neurochem.*, **87**, 1558–1567.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, E1293–301.

Mosca,R. *et al.* (2012) Interactome3D: adding structural details to protein networks. *Nat. Methods*, **10**, 47–53.

Mullaney,J.M. *et al.* (2010) Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.*, **19**, 131–136.

Nei,M. *et al.* (2010) The neutral theory of molecular evolution in the genomic era. *Annu. Rev. Genomics Hum. Genet.*, **11**, 265–289.

Nevin Gerek,Z. *et al.* (2013) Structural dynamics flexibility informs function and evolution at a proteome scale. *Evol. Appl.*

Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.

Nussinov,R. (2016) Introduction to Protein Ensembles and Allostery. *Chem. Rev.*, **116**, 6263–6266.

Nussinov,R. *et al.* (2013) The underappreciated role of allostery in the cellular network. *Annu. Rev. Biophys.*, **42**, 169–89.

Nussinov,R. and Tsai,C.-J. (2013) Allostery in disease and in drug discovery. *Cell*, **153**, 293–305.

Nussinov,R. and Tsai,C.J. (2015) 'Latent drivers' expand the cancer mutational landscape. *Curr. Opin. Struct. Biol.*, **32**, 25–32.

Parthiban,V. *et al.* (2006) CUPSAT: Prediction of protein stability upon point mutations. *Nucleic Acids Res.*, **34**, 239–242.

Pasmanik-Chor,M. *et al.* (1996) The glucocerebrosidase D409H mutation in Gaucher disease. *Biochem. Mol. Med.*, **59**, 125–133.

Pearlman,D.A. *et al.* (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.*, **91**, 1–41.

Pearson,D.S. (1977) Scattered Intensity from a Chain in a Rubber Network. *Macromolecules*, **10**, 696–701.

Pires,D.E. V *et al.* (2014) DUET: A server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.*, **42**, 1–6.

Pol-Fachin,L. *et al.* (2016) Glycosylation is crucial for a proper catalytic site organization in human glucocerebrosidase. *Glycoconj J*, 237–244.

Porter,C.T. *et al.* (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–133.

Potapov,V. *et al.* (2009) Assessing computational methods for predicting protein stability upon mutation: Good on average but not in the details. *Protein Eng. Des. Sel.*, **22**, 553–560.

Prabhakar,S. *et al.* (2006) Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res.*, **16**, 855–863.

Qin,W. (1996) Functional Organization of Saposin C. *J. Biol. Chem.*, **271**, 6874–6880.

Reeb,J. *et al.* (2016) Predicted Molecular Effects of Sequence Variants Link to System Level of Disease. *PLOS Comput. Biol.*, **12**, e1005047.

Rouse,P.E.J. (1953) A Theory of the Linear Viscoelastic Properties of Dilute Solutions of Coiling Polymers. *J. Chem. Phys.*, **21**, 1272.

Sahni,N. *et al.* (2015) Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. *Cell*, **161**, 647–660.

Salomon-Ferrer,R. *et al.* (2013) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.*, **9**, 3878–3888.

Schröder,G.F. *et al.* (2010) Super-resolution biomolecular crystallography with low-resolution data. *Nature*, **464**, 1218–1222.

Schrodinger (2010) The PyMOL Molecular Graphics System, Version 1.3r1.

Schuster-Böckler,B. and Bateman,A. (2008) Protein interactions in human genetic diseases. *Genome Biol.*, **9**, R9.

Schwarz,J.M. *et al.* (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.

Schymkowitz,J. *et al.* (2005) The FoldX web server: An online force field. *Nucleic Acids Res.*, **33**, 382–388.

Seemayer,S. *et al.* (2014) CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics*, **30**, 3128–3130.

Shan,Q. *et al.* (2012) Function of hyperekplexia-causing alpha1R271Q/L glycine receptors is restored by shifting the affected residue out of the allosteric signalling pathway. *Br. J. Pharmacol.*, **165**, 2113–2123.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–11.

Shi,Y. (2014) A glimpse of structural biology through X-ray crystallography. *Cell*, **159**, 995–1014.

Shi,Z. and Moult,J. (2011) Structural and functional impact of cancer-related missense somatic mutations. *J. Mol. Biol.*, **413**, 495–512.

Sidore,C. *et al.* (2015) Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.*, **47**, 1272–1281.

Sidransky,E. (2004) Gaucher disease: complexity in a 'simple' disorder. *Mol. Genet. Metab.*, **83**, 6–15.

Sinha,N. and Nussinov,R. (2001) Point mutations and sequence variability in proteins: redistributions of preexisting populations. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 3139–3144.

Smock,R.G. *et al.* (2010) An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Mol Syst Biol*, **6**, 414.

Stefl,S. *et al.* (2013) Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.*, **425**, 3919–3936.

Stehr,H. *et al.* (2011) The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Mol. Cancer*, **10**, 54.

Stenson,P.D. *et al.* (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–81.

Stenson,P.D. *et al.* (2014) The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **133**, 1–9.

Steward,R.E. *et al.* (2003) Molecular basis of inherited diseases: A structural perspective. *Trends Genet.*, **19**, 505–513.

Stone,D.L. *et al.* (2000) Glucocerebrosidase gene mutations in patients with type 2 Gaucher disease. *Hum. Mutat.*, **15**, 181–188.

Stone,E. a (2014) Predictor performance with stratified data and imbalanced classes. *Nat. Methods*, **11**, 782–3.

Stone,E. a and Sidow,A. (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.*, **15**, 978–986.

Sun,Y. and Kollman,P.A. (1995) Hydrophobic solvation of methane and nonbond parameters of the TIP3P water model. *J. Comput. Chem.*, **16**, 1164–1169.

Swint-Kruse,L. (2016) Using Evolution to Guide Protein Engineering: The Devil IS in the Details. *Biophys. J.*, **111**, 10–18.

Tajima,A. *et al.* (2010) Gaucher disease patient with myoclonus epilepsy and a novel mutation. *Pediatr. Neurol.*, **42**, 65–8.

Tama,F. and Sanejouand,Y.H. (2001) Conformational change of proteins arising from normal modecalculations. *Protein Eng. ...*, **14**, 1–6.

Tavtigian,S. V (2005) Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.*, **43**, 295–305.

Tawfik,D.S. and Tokuriki,N. (2009) Protein Dynamism and Evolvability. *Science (80-.).*, **324**, 203–207.

Teng,S. *et al.* (2009) Modeling effects of human single nucleotide polymorphisms on protein-protein interactions. *Biophys. J.*, **96**, 2178–88.

Tennessen,J.A. *et al.* (2012) Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. **337**, 64–69.

Thomas,P.D. *et al.* (2003) PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129–2141.

Tian,J. *et al.* (2007) Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. *BMC Bioinformatics*, **8**, 450.

Tirion,M. (1996) Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.*, **77**, 1905–1908.

Topham,C.M. *et al.* (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng.*, **10**, 7–21.

Tsodikov,O. V. *et al.* (2002) A novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J. Comput. Chem.*, **23**, 600–609.

Tuncbag,N. *et al.* (2010) HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res.*, **38**, W402–6.

Tzeng,S. and Kalodimos,C. (2011) Protein dynamics and allostery: an NMR view. *Curr. Opin. Struct. Biol.*, **21**, 62–67.

Velazquez-Muriel,J. *et al.* (2009) Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC Struct. Biol.*, **9**.

Villy Isaksen,G. *et al.* (2016) Enzyme surface rigidity tunes the temperature dependence of catalytic rates. *Proc. Natl. Acad. Sci.*, **113**.

Vitkup,D. *et al.* (2003) The amino-acid mutational spectrum of human genetic disease. *Genome Biol.*, **4**, R72. Epub 2003 Oct 30.

Wagner,J.R. *et al.* (2016) Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chem. Rev.*, **116**, 6370−6390.

Wang,S. *et al.* (2016) CoinFold: a web server for protein contact prediction and contact-assisted protein folding. *Nucleic Acids Res.*

Wang,X. *et al.* (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.*, **30**, 159–64.

Wang,Y. *et al.* (2004) Global ribosome motions revealed with elastic network model. *J.*

Struct. Biol., **147**, 303–314.

Wang,Z. and Moult,J. (2001) SNPs, protein structure, and disease. *Hum. Mutat.*, **17**, 263–270.

Wei,Q. *et al.* (2012) Prediction of phenotypes of missense mutations in human proteins from biological assemblies. *Proteins Struct. Funct. Bioinforma.*, **81**, 199–213.

Woldeyes,R.A. *et al.* (2014) E pluribus unum, no more: From one crystal, many conformations. *Curr. Opin. Struct. Biol.*, **28**, 56–62.

Worth,C.L. *et al.* (2011) SDM - A server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res.*, **39**, 215–222.

Xie,L. *et al.* (2014) Towards Structural Systems Pharmacology to Study Complex Diseases and Personalized Medicine. *PLoS Comput. Biol.*, **10**.

Yang,L. *et al.* (2009) Protein elastic network models and the ranges of cooperativity. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 12347–52.

Yang,L.-W. *et al.* (2014) Ligand-induced protein responses and mechanical signal propagation described by linear response theories. *Biophys. J.*, **107**, 1415–25.

Yates,C.M. and Sternberg,M.J.E. (2013) The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *J. Mol. Biol.*, **425**, 3949–3963.

Ye,Z.Q. *et al.* (2007) Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). *Bioinformatics*, **23**, 1444–1450.

Yue,P. *et al.* (2005) Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.*, **353**, 459–473.

Yue,P. *et al.* (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**.

Zhao,N. *et al.* (2014) Determining effects of non-synonymous SNPs on protein-protein interactions using supervised and semi-supervised learning. *PLoS Comput. Biol.*, **10**, e1003592.

Zheng,W. *et al.* (2007) Allosteric transitions in the chaperonin GroEL are captured by a dominant normal mode that is most robust to sequence variations. *Biophys. J.*, **93**, 2289–2299.

Zheng,W. and Brooks,B.R. (2005) Probing the local dynamics of nucleotide-binding

pocket coupled to the global dynamics: myosin versus kinesin. *Biophys. J.*, **89**, 167–178.

Zheng,W. and Tekpinar,M. (2009) Large-scale evaluation of dynamically important residues in proteins predicted by the perturbation analysis of a coarse-grained elastic model. *BMC Struct. Biol.*, **9**, 45.

Zou,T. *et al.* (2015) Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. *Mol. Biol. Evol.*, **32**, 132–143.