Computational Methods for the Analysis of Molecular Dynamics Simulations

Noah C. Benson

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2010

Program Authorized to Offer Degree:
Medical Education and Biomedical Informatics

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Noah C. Benson

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

Valerie Daggett

Reading Committee:

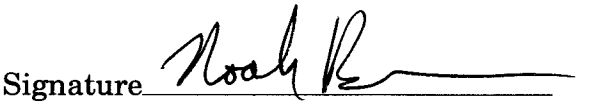
Valerie Daggett

James F. Brinkley

Peter Myler

Date: 8/5/10

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature 

Date 8-5-2010

University of Washington

Abstract

Computational Methods for the Analysis of Molecular Dynamics Simulations

Noah C. Benson

Chair of the Supervisory Committee:
Professor Valerie Daggett
Bioengineering

Proteins are macromolecules that are involved in virtually every biological process and structure. The three-dimensional structure of these molecules is extremely important as a window into how they work but is extremely difficult to predict, as direct observation of their motion and the folding pathway is possible only through very limited experimental techniques. Nonetheless, observing protein structure alone has proven insufficient for understanding how proteins fold or behave natively. Molecular dynamics (MD) is a computational technique by which protein dynamics can be examined at resolutions well beyond the capabilities of experiment. The decrease in cost of computer resources have lead biologists to turn to MD more frequently in recent years, yet MD simulations produce data in quantity and complexity well beyond the capabilities of conventional biological analysis techniques. We have curated a database of protein native-state and thermal unfolding simulations, which is the largest database of unfolding simulations to date. We examine this database using two existing and three novel analysis methods and demonstrate the utility of each for high-throughput analysis. Finally, we demonstrate that these methods can be used to generate and support novel hypotheses concerning protein motion.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to thank Dr. Valerie Daggett for mentorship throughout his graduate carreer, as well as Drs. James Brinkley, Peter Myler, and Walter Ruzzo for serving on his committee. The author would like to thank Drs. Amanda Jonsson, Dustin Schaeffer, Karen Rutherford, Erik Merkley, David Beck, Kathryn Scott, Darwin Alonso, Rudesh Toofany, Gene Hopping, Marc van der Kamp, Peter Anderson, and Ira Kalet as well as Andrew Simms, Michelle McCully, Alex Scouras, Denny Bromley, and Steve Rysavy for helpful discourse and logistical support. The author would also like to thank the Department of Medical Education and Biomedical Informatics and the Division of Biomedical and Health Informatics for administrative support, the National Library of Medicine and National Institute of Health for financial support, and the National Energy Research Council for computational resources.

DEDICATION

Dedicated to my parents, K. D. and David Benson, and to my friend and colleague, Ethan Romero-Severson, without any one of whom I would never have realized a carreer in research.

Chapter 1

PROTEIN DYNAMICS AND THE DYNAMEOMICS DATABASE

1.1 Summary

Studying protein dynamics at high resolution is only possible through computational techniques, the most detailed and accurate of which is molecular dynamics (MD) simulation. The dwindling cost of computational resources has lead to increased use of MD in the biological community, yet interpretation of the highly complex and data-rich trajectories produced by MD can require intense human resources. Current analysis techniques capable of rapidly and intuitively characterizing MD simulations lag behind the the community's needs and ability to generate data. In order to enable the development of such techniques, to provide a contextual reference for the behavior of proteins, and to study the unfolding behavior of proteins, we have performed MD simulations of the native state and thermal unfolding pathways of over 1000 proteins. These proteins were chosen to represent the majority of globular folds. These data have been organized using a novel database approach, which has been designed both to be easily comprehensible and to connect several disparate data types and sources. This database can be easily mined so as to maximally facilitate further research. In this chapter, we describe the organization of this database and provide examples of how it can be mined and what types of analysis needs it exposes that we propose to fill.

1.2 Introduction

The motion, or *dynamics*, of proteins is critical to understanding how proteins fold (Daggett and Fersht, 2003; Schaeffer et al., 2008), produce mis-folding disease states (Chiti and Dobson, 2006; Daggett, 2006), and function (Karplus and Kuriyan, 2005; Glazer et al., 2009). Though difficult to study experimentally, these dynamics are not random but rather are determined by the net sum of fundamental forces incident on a protein from itself and its environment. These forces, which act on each atom of the protein, are simple individually, but extremely complex in union, and, similar to the n-body problem, are not known to be concisely predictable. Determining the sub-states, conformations, and modes of a proteins is thus very difficult. Numerical integration using Newton's laws of motion and a set of potential functions is possible, however, and forms the basis of molecular dynamics (MD) (Karplus and McCammon, 2002; Beck and Daggett, 2004). MD is a well-developed technique that allows for very high temporal and spacial resolution; thus it often serves as a hypothesis generation engine as well as a means of studying phenomena not easily accessible to experiment (Fersht and Daggett, 2002; van der Kamp et al., 2008).

In recent years, large-scale simulations of proteins have become increasingly common. These simulations can generate massive amounts of data. Common time resolutions are on the picosecond or smaller scale, and it is not uncommon for simulations to exceed tens or hundreds of nanoseconds. Large protein systems can contain tens of thousands of atoms; thus uncompressed individual trajectories can be terabytes in size. Databases of protein simulations, such as the MoDEL project (Rueda et al., 2007), P-found (Silva et al., 2006), and BioSimGrid (Murdock et al., 2005; Ng et al., 2006), contain even larger quantities of data and demonstrate the demand for such simulation warehouses. These databases can be considered extensions of the Protein Data Bank (PDB)

(Berman et al., 2000), in that they provide similar structural information but add the dimension of time.

Here we describe our own MD simulation database, the Dynameomics project, which is the largest database of its kind. Dynameomics includes not only native-state simulations (proteins simulated at 298 K), but also seeks to provide a resource for the study of protein unfolding by simulating proteins at high temperature (498 K). As a research-enabling project as well as a research project itself, the Dynameomics database has been carefully organized and annotated. We will describe this organization as well as the work-flow and meta-data required to support and create it. Finally, we will discuss the implications of database and the research questions that it now poses.

1.3 Generation of the Dynameomics Database

1.3.1 Protein Selection

Proteins share similar structures in nature, and several classifications of fold families, called domain dictionaries, have been proposed, each of which has a unique approach to classifying proteins and protein subunits into related domains. One goal of the Dynameomics project is to cover as diverse a set of protein folds as possible so as to examine all possible native-state and unfolding dynamics. In order to do this, we combined the major fold domain dictionaries (Murzin et al., 1995; Cuff et al., 2009; Dietmann and Holm, 2001) into a single Consensus Domain Dictionary (CDD) (Day and Daggett, 2003; Schaeffer et al., 2010). This process consisted of two steps: (1) identification of all domains in each protein structure stored in the PDB, and (2) identifying 'metafolds' for which at least two of the domain dictionaries have at least an 80% sequence overlap. Filtering was performed to make sure that redundancies in sequence did not bias the CDD. Our CDD currently contains 1,695 metafolds, which en-

capsulate 80,062 domains.

Dynameomics aims to simulate at least one representative structure from each of these domains. These 'fold representatives' were chosen based on structure quality, size, biomedical relevance, availability of experimental data, autonomy of the structure, and absence of complex cofactors. A rank was given to metafolds based on their populations (i. e., those metafolds to which a larger number of structures were assigned received higher ranks). Preference was given to the highest ranking metafolds when choosing which proteins to simulate first. Those that were ultimately chosen for simulation are referred to as simulation 'targets'. To date, we have simulated at least one representative from 807 of our 1,695 metafolds; this represents 81% of the domains in the CDD. The remaining metafolds were almost exclusively of low-rank and did not have suitable simulation targets, generally due to problems associated with simulating them outside of the context in which their structures were determined. Figure 1.1 shows an example of one such domain.

In addition to the 807 targets chosen to represent the diversity among protein fold families, Dynameomics includes 29 proteins with disease-causing single nucleotide polymorphisms (SNPs). While these simulations are a subset of the folds represented in Dynameomics project, we ran additional simulations of variants of these proteins encompassing 200 single-point mutations (649 simulations total) for at least 30 ns. These proteins include p53 and 8-oxoguanine glycosylase (cancer), DJ-1 (Parkinson's disease), superoxide dismutase (amyotrophic lateral sclerosis), catechol O-methyltransferase (alcoholism and aggression in schizophrenia), transthyretin (amyloidosis), and thiopurine S-methyltransferase (drug-metabolism disorders). The goal of these SNP simulations is to enable research on the dynamic effects of point mutations.

1.3.2 Protein Preparation

Coordinates for the targets to be simulated were obtained from the PDB. In some cases, targets had missing atoms or residues due to experimental limitations; these were built into the PDB prior to additional preparation. During preparation, all His residues were changed to one of four states: Hie (Nϵ protonated), Hid (Nδ protonated), Hin (neither Nδ or Nϵ protonated) or Hip (both Nδ and Nϵ protonated). These states were chosen based on the proximity of residues that might lead to a preference of one of the four; for example, a Ser OH group near the Nϵ would indicate a preference for Hid due to the nearby positively charged hydrogen. A strong preference was given for Hie and Hid. During preparation of the unfolding trajectories, Cys residues were additionally reduced to Cyh. Proteins then underwent a brief energy minimization prior to solvation in a periodic water box using experimental density for the temperature of interest.

1.3.3 Protein Simulation

Simulations were performed using our in-house developed MD simulation package, *in lucem* molecular mechanics (*il*mm) (Beck et al., 2008; Beck and Daggett, 2004). All atoms were explicitly simulated with parameters defined in our force field (Levitt et al., 1995). Waters were also explicitly represented and used a flexible 3-center water model (Levitt et al., 1997). Further details of the simulation and preparation protocol have been presented elsewhere (Beck et al., 2008). Each target underwent at least one native-state (298 K) simulated for at least 31 ns as well as at least 3 long (at least 31 ns) and 2 short (at least 2 ns) unfolding (498 K) simulations. Coordinates were saved every 0.2 ps for the short unfolding simulations and every picosecond for the other simulations.

1.3.4 Analysis and Quality Control

Each complete simulation was analyzed with a set of analyses that are standard in MD research. These include measurements of the root mean square deviation (RMSD), root mean square fluctuation (RMSF), solvent accessible surface area (SASA), (Φ, Ψ) angle propensities, radius of gyration, number of native and non-native contacts, and secondary structure content using the dictionary of secondary structure of proteins (DSSP) (Kabsch and Sander, 1983). Examples of these analysis can be found in Figure 1.2. Further details about standard analysis are outlined by Beck et al. (2008) and given in Section 7.2.

An important part of analysis was determining if our simulations were realistic. Due to the many variables involved in preparation and execution of a simulation, that certain proteins will simply not be stable in our force field. This is can be due, for example, to the protein being simulated in an unusual context (e. g., without a binding partner), having a poorly parameterized co-factor, or having a poor quality structure. Due to the fact that the simulation conditions are slightly different than the experimental conditions in which proteins are characterized, we expect some moderate changes in the structure during simulation. Often this includes a slight expansion of the structure and small rearrangements of secondary structure. In addition, fluctuations are expected because proteins are not static entities but rather are in constant motion. The goal of our initial quality control was to determine when a target's rearrangements and fluctuations went beyond what was appropriate for a protein in solution. We determined that this occurred when one of three instability criteria was met: (1) the protein's hydrophobic core became exposed to solvent, (2) the protein lost entire elements of secondary structure, or (3) the protein does not find a stable rearrangement by the end of the simulation.

Searching each simulation for these events by hand is a daunting task, so

we developed a method for automatically placing simulations in high-risk and low-risk categories. To do this, we first calculated the 'canonical' or median structure using the protein's core only. The protein core was defined as the set of residues whose SASA was fewer than 40 $Å^2$ and was used because loop and tail rearrangements are common and generally of little consequence in terms of stability. The median structure is the single structure in the simulation that has the lowest mean RMSD to every other structure in the simulation; alternately, it can be thought of as the physical structure that is closest to the mean structure. We then calculated two values: the median RMSD (mRMSD) and the median RMSF (mRMSF), again using only the protein core. The mRMSD is the RMSD between the starting structure and the median structure; it measures the amount of rearrangement that has occurred over the course of the simulation from the experimental structure. The mRMSF is the mean RMSF using the median structure as a reference. In other words, this is the average RMSD from the median structure to every other structure. The mRMSF measures the overall stability of the rearranged state. Conceptually, these two values can be thought of as the distance from the experimental structure to the solution structure (mRMSD) and the overall stability of the solution ensemble (mRMSF). An unusually high mRMSD indicates that the protein has undergone an unusually large amount of rearrangement from its experimental structure while an unusually high mRMSF indicates that the protein's rearranged state is unusually unstable. High values of one or the other can be tolerated in certain circumstances (e. g., for largely unstructured proteins), but they often indicate that something is wrong. A plot of the proteins in Dynameomics according to mRMSD and mRMSF is shown in Figure 1.3.

In general, most simulations had low values of mRMSD and mRMSF (Fig. 1.3). To categorize a simulation as high-risk, we calculated the Euclidean norm of the (mRMSD, mRMSF) vector, which we call the stability norm.

Equation 1.1 describes the formula for the stability norm ρ for a simulation with structures $(s_1, s_2, ...s_T)$ and median structure s_m in terms of the function $D_{RMS}(a, b)$, which is the RMSD between a structure a and a structure b. For any simulation, if its stability norm was $\geq$ 3.5 Å, it was categorized as high-risk. These simulations were then examined by hand to determine if any of the instability criteria were met.

$$\rho = \sqrt{\left(\frac{1}{T}\sum_{t=1}^{T} D_{RMS}(s_m, s_t)\right)^2 + (D_{RMS}(s_m, s_1))^2} \tag{1.1}$$

Of our initial 821 simulation, 19 of them (2.3%) were deemed unstable; 5 of these had viable replacements, leaving us with 807. Notably, all of the structures that were deemed unstable were experimentally determined by NMR. The simulations that were rejected as well as the reasons for rejection are given in Table 1.1. The highest rank of those that could not be replaced was 633.

In addition to these quality control metrics, we also compared our simulations to experimental data when available. This involved comparison to nuclear Overhauser effect (NOE) crosspeaks during native-state simulations, in which we find good agreement (Beck et al., 2008). Although experimental data for unfolding simulations is very limited, we have found good agreement with experiment in past unfolding studies (Daggett et al., 1998; Ladurner et al., 1998; Mayor et al., 2003; Daggett, 2006).

1.3.5 Database Organization

One of the foremost challenges of the Dynameomics project was organizing such a diverse set of data in a way that was both internally consistent and externally comprehensible. Biology is a field that is often defined by exceptions to the established rules, and proteins are no exception to this. Representing even as simple a concept as protein structure in a logical format can be chal-

lenging. The Dynameomics database desired not only to represent this fundamental data, but to merge it seamlessly with data from many different kinds of analyses and a variety of external databases. Critically, whatever schema was adopted also needed to enable external users, who may not be well-versed in computer and database technologies, to examine the data. Finally, data access had to be flexible and fast in order to enable research of both a biological and informatic nature.

The Dynameomics database was implemented in Microsoft SQL Server 2008 (Microsoft, 2008). SQL Server was chosen because it supports a wide variety of interfaces including several that enable custom user-developed tools. Our simulation package, *il*mm (Beck et al., 2008), in fact, can transfer data directly to the Dynameomics database. Additionally, we have developed a variety of tools that allow users to navigate the database without having to learn the details of the database schema. One such tool, Dynamanal, is described in detail in Appendix A, while a Mathematica (Wolfram Research, 2008) library for communication with the Dynameomics database is described in Appendix B. The database is organized into several sections, each of which represents a unique piece of the overall project but which are connected through a complex schema. These parts are the Prep database, the Simulation database, and the Directory database.

The Directory database is the central organizing repository. This database is highly curated and connects the many types of data in a coherent schema. Because the Dynameomics database is so large, it had to be fragmented into many smaller databases and distributed across several servers (Simms et al., 2008), each if which hosts a unique subset of the data. The Directory database allows one to immediately determine which server, database, and table contains a particular piece of the overall database.

The Prep database is also highly curated and contains all the information

related to the work-flow and process behind Dynameomics. This includes all the relevant data about target selection, preparation, simulation parameters, and quality control. Even fold members that were not simulated are represented in this database. Conceptually, should the Simulation database be wiped out for some reason, the Prep database would contain all the information required to regenerate it.

The Simulation database is by far the largest database and contains all of the 3D coordinates and analyses of the simulations for the 807 targets. In addition to containing all of the time-course data encapsulated by Dynameomics, the Simulation database stores structural definitions of each simulation. These structure definitions are an extension to the standard PDB structures, as they link traditional PDB concepts such as residue number and icode, to concepts that are easily understood by a computer, database, or SQL query.

The Dynameomics database is 4 magnitudes of order larger than the PDB and stores $> 10^8$ structures in over 53 TB. Proteins range from 29 to 417 residues in size with an average size of 137. The total length of simulation time is over 180 μs. The number of proteins and protein domains simulated that are fold representatives or other metafold members is over 1000. The length of these soluble proteins ranges between 29 and 417 residues, with an average size of 137. The majority of targets do not contain co-factors: 22 domains contain zinc, 9 heme and 2 calcium. Over 70% of the starting structures were determined by X-ray diffraction (average resolution: 2.06 Å), the others were obtained by protein NMR. Additional statistics regarding the data, including distribution of source organisms and enzyme classifications, are reported by van der Kamp et al. (2010).

1.4 Implications of Dynameomics

The Dynameomics project was designed not only to explore native-state and unfolding dynamics but also to serve as a scientific repository from which additional research, both biological and informatic in nature, could be engendered. Although Dynameomics has already enabled several interesting discoveries about protein dynamics, these discoveries fall far short of its potential. Here, we will discuss some examples of the data observed from the database as well as some of the challenges that prevent further research from moving forward.

1.4.1 Native-state Dynamics

A single native-state simulation, let alone an entire database of native-state simulations, contains a great deal of information. Although traditional analysis methods are common and established ways of characterizing a protein's dynamics, they suffer from a few shortcomings. Primary among these is the fact that, because they were originally developed for use on single structures, they tend to view a dynamic simulation as a set of single frames rather than as a dynamic evolving trajectory. SASA, for example, gives a measurement of solvent accessibility for each frame, but this measurement is independent of every other frame. Beyond this, traditional analyses do not tend to summarize data but to expand on data. While this can be desirable, large sets of simulations can make this kind of analysis difficult or intractable and may require techniques that, at least initially, summarize data and relate it to physical properties of a protein's trajectory. In chapter 2 we explore a technique for summarizing and visualizing the dynamics of a proteins and demonstrate that it can be used to facilitate biochemical discovery.

Native-state dynamics tend to be very subtle in nature compared to unfolding simulations. Often, when significant events occur in a native-state

protein simulation, they are difficult to locate due to the very small changes that comprise them. A perfect example of this phenomenon can be found in a set of methyltransferases whose SNP variants are part of the Dynameomics project. The proteins catechol O-methyltransferase (COMT), L-isoaspartate O-methyltransferase (PIMT), thiopurine S-methyltransferase, and histamine N-methyltransferase all have very similar folds, each with a mutation in a congruent spot at least 16 Å from the active site. These SNPs are known to cause significant effects on the protein's behavior and are associated with disease, but the mechanism by which this occurred was not previously understood.

Although explanations for these phenomena were developed using simulations in the Dynameomics database (Rutherford et al., 2006; Rutherford and Daggett, 2008; Rutherford et al., 2008b; Rutherford and Daggett, 2009a,b, 2010), the individual effects of each mutation were extremely elusive. The research made it clear that computational methods and tools for comparing similar simulations and for searching simulations for unusual or interesting events were severely lacking. Our research on wavelet analysis, which develops and tests methods designed to solve these problems, is discussed in chapters 3 (searching simulations) and 4 (comparing simulations). Additional research using graph theoretic methods to index and compare individual residues based on their chemical environments is discussed in chapter 5.

1.4.2 Unfolding Dynamics

Unfolding simulations tend to be very dramatic in comparison to native-state simulations thus have very different analysis needs. For example, Jonsson et al. (2009) used a conformational clustering method Li and Daggett (1994) to identify transition states in 183 high ranking targets in Dynameomics and used these data to characterize several features of the protein unfolding transition state. This method, however, is considerably less useful in native-state

simulations, in which all conformations tend to be extremely similar.

Because unfolding is a fast process that is characterized largely by changes to a protein's topology and hydrophobic core, we hypothesized that metrics analyzing the relationship between a part of the protein to the whole of the protein might be more useful in unfolding studies. Chapter 6 discusses our research into one such metric, the betweenness centrality of graph theory, that has shed considerable light on how protein's move through their transition states to their denatured states and back.

1.5 Conclusions

The Dynameomics project is an immense project around which considerable research has already been and will continue to be precipitated. The careful organization of the project facilitates the exploration of novel topics while the raw diversity and size of the database allows hypotheses to be formed and tested efficiently. Nonetheless, the project has exposed critical holes in the existing methodology for MD analysis and research. In this thesis, we explore our solutions to these holes and present the novel findings they have allowed us to discover.

Table 1.1: Simulations rejected by quality control.

PDB Code	Reason for Rejection
1j1h	Opening of hydrophobic core
1jyt	Opening of hydrophobic core
1t23	Opening of hydrophobic core
1e0g	Continuous rearrangements/no stable state
1iba	Continuous rearrangements/no stable state
1k0h	Continuous rearrangements/no stable state
1uw2	Continuous rearrangements/no stable state
1iyr	Continuous rearrangements/no stable state
2afp	Continuous rearrangements/no stable state
7hsc	Continuous rearrangements/no stable state
1b9g	Nontrivial secondary structure rearrangements
1fu9	Nontrivial secondary structure rearrangements
1hyw	Nontrivial secondary structure rearrangements
1i42	Nontrivial secondary structure rearrangements
1kkx	Nontrivial secondary structure rearrangements
1nr3	Nontrivial secondary structure rearrangements
1q3j	Nontrivial secondary structure rearrangements
1qu6	Nontrivial secondary structure rearrangements
1vpu	Nontrivial secondary structure rearrangements

Figure 1.1: The human P1/Mahoney Poliovirus (*1al2*) with chain 4 highlighted. Chain 4 is a fold representative in the Dynameomics CDD, but is clearly not self-contained and would be an inappropriate simulation target.

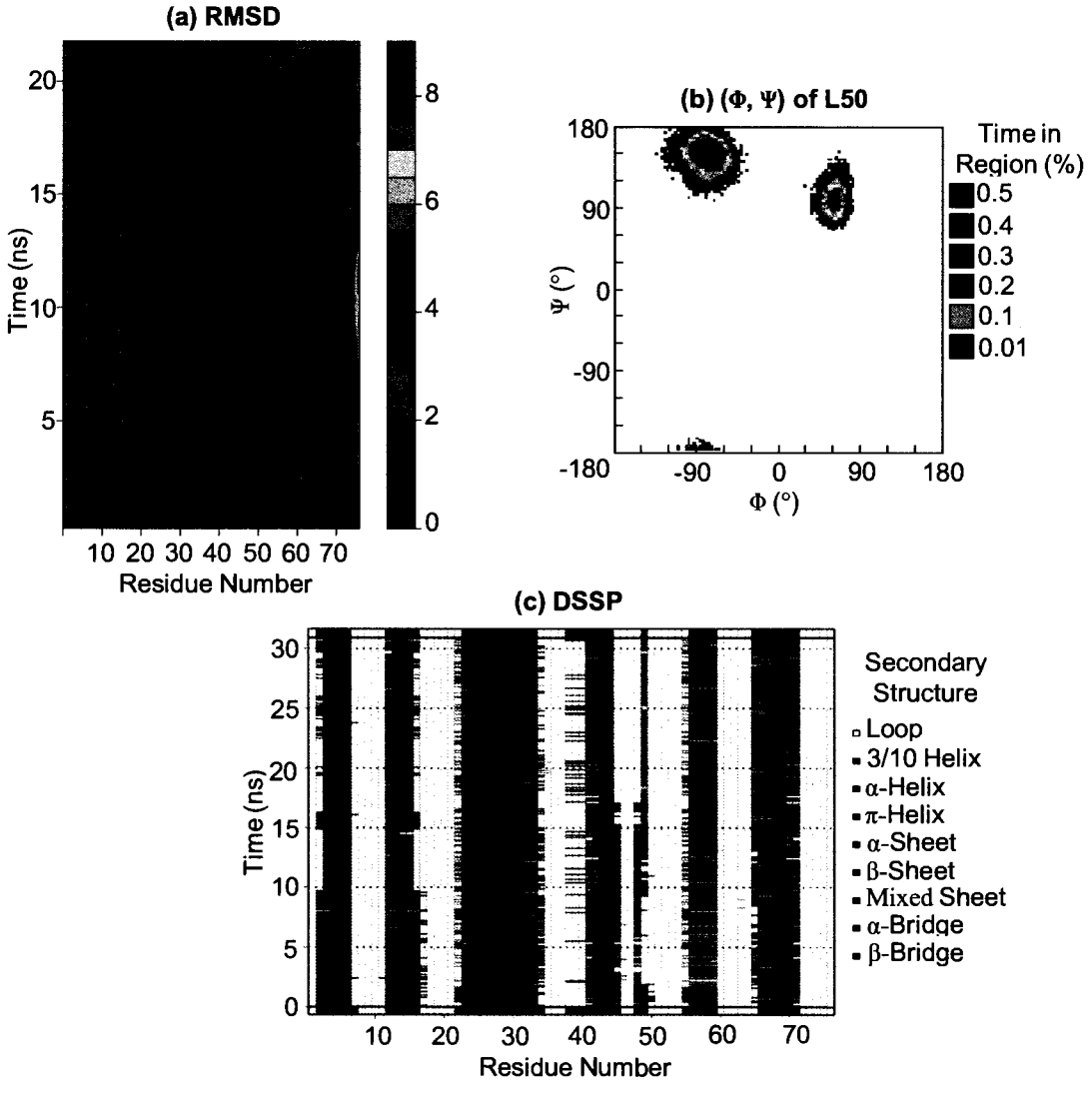


Figure 1.2: Example analyses of the protein ubiquitin (*1ubq*). **(a)** Root-mean-square deviation (RMSD) over time. **(b)** (Φ, Ψ) angle distributions of residue L50. **(c)** Dictionary of Secondary Structure of Proteins (DSSP) over time.

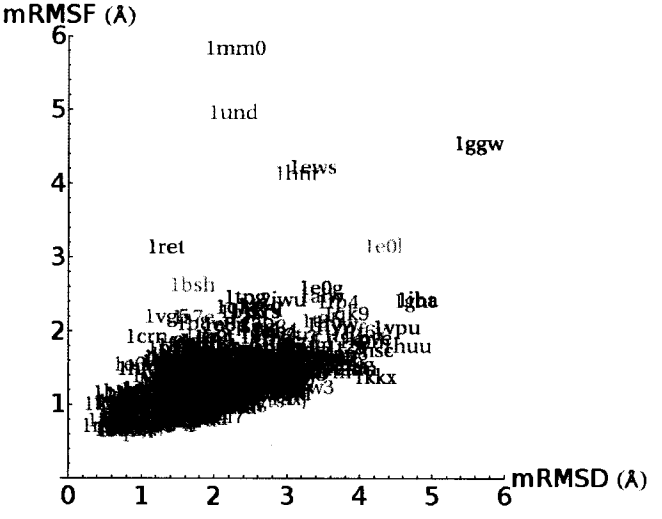


Figure 1.3: Plot of PDB codes of all targets in Dynameomics by mRMSD and mRMSF. Those proteins whose distance from the origin is ≥ 4 Å were categorized as high-risk and examined for potential instability.

Chapter 2

DYNAMEOMICS: LARGE-SCALE ASSESSMENT OF NATIVE PROTEIN FLEXIBILITY

2.1 Summary

Structure is only the first step in understanding the interactions and functions of proteins. In this paper, we explore the flexibility of proteins across a broad database of over 250 solvated protein molecular dynamics simulations in water for an aggregate simulation time of 6 μs. These simulations are from our Dynameomics project, and these proteins represent approximately 75% of all known protein structures. We employ principal component analysis of the atomic coordinates over time to determine the primary axis and magnitude of the flexibility of each atom in a simulation. This technique gives us both a database of flexibility for many protein fold families and a compact visual representation of a particular protein's native-state conformational space, neither of which are available using experimental methods alone. These tools allow us to better understand the nature of protein motion and to describe its relationship to other structural and dynamical characteristics. In addition to reporting general properties of protein flexibility and detailing many dynamic motifs, we characterize the relationship between protein native-state flexibility and early events in thermal unfolding and show that flexibility predicts how a protein will begin to unfold. We provide evidence that fold families have conserved flexibility patterns, and family members who deviate from the conserved patterns have very low sequence identity. Finally, we examine novel aspects of highly inflexible loops that are as important to structural integrity as conven-

tional secondary structure. These loops, which are difficult if not impossible to locate without dynamic data, may constitute new structural motifs.

2.2 Introduction

Much scientific effort has been spent attempting to catalog, describe, observe, and understand protein structure and function. Even when the structure of a protein is known, this knowledge is often not sufficient to elucidate details of the protein's function or its mode of action, both of which are pieces of information that are frequently of much greater importance to biologists than structure. As biologists increasingly seek to understand and modify aspects of cellular behavior and as protein databases gather more high-resolution three-dimensional structures, the ability to understand key features of a protein's dynamic behavior becomes more important.

Flexibility is critical in determining protein behavior and function. Because proteins are not static entities (as they are represented in structural databases) and because crystal structures do not necessarily represent a protein in its active conformation, any attempt to determine potential biochemical interactions of a protein from these data suffers from a lack of information about its motion. A quantitative description of a protein's flexibility provides a summary of its dominant dynamical modes and significant information about potential conformations available to it. Flexibility may also provide insight into unfolding and folding pathways because a protein is most likely to start unfolding, and to finish folding, at a site that is highly mobile. Thus, flexibility may affect not only function but also unfolding and stability.

Molecular dynamics (MD) is a common method for determining protein motion over time. MD provides the researcher with snapshots of a protein's conformation at regular time intervals. These data, when saved at frequent enough intervals, behave as a stop-motion photography film and can be analyzed by

mathematical and statistical techniques to further explore protein motion.

The Dynameomics project (Beck et al., 2008) is a large-scale effort to simulate a protein from every protein fold family (Day and Daggett, 2003). The Dynameomics database (Kehl et al., 2008; Simms et al., 2008) currently contains 450 proteins, each of which has been simulated for at least 21 ns at a temperature of 298 K. Additionally, it contains at least two unfolding simulations of each protein at 498 K for 31 ns and at least three short (2 ns) simulations at 498 K. These simulated target proteins form a data set that spans a considerable portion of the protein universe, representing $> 75\%$ of all known protein folds.

Here we focus on the analysis of general features of protein flexibility of the native-state proteins in the Dynameomics project, resulting in a database of protein flexibilities. For three of these highly populated folds, we compared 36 family members to determine if flexibility is conserved across a fold family. Then we compare native-state flexibility with unfolding behavior to explore the relationship between flexibility and the mechanism of early unfolding. Finally, we searched our database of flexibilities for unstructured regions whose flexibility was uncharacteristically low, and we use these findings to demonstrate how flexibility may be useful for determining intrinsic properties of structure that are difficult to elucidate with other techniques.

2.3 Methods

Simulations were performed with explicit water using our in-house developed simulation package in lucem molecular mechanics (Beck et al., 2008; Beck and Daggett, 2004) and our previously described protein and water force fields (Levitt et al., 1995, 1997). Simulation details can be found elsewhere (Beck et al., 2008). For each simulation, atomic coordinates from all but the first nanosecond of our trajectories were downloaded from our in-house developed

data warehouse (Simms et al., 2008) into Mathematica Version 5.2 (Wolfram Research, 2005) for analysis. The first nanosecond was omitted to allow for equilibration. For each picosecond of the simulation, the protein structure was aligned to the initial structure using a rigid least squares fitting of Cα atoms with the structure's center of mass held at the origin (Kearsley, 1989). The coordinates of each non-hydrogen atom were centered by subtracting the atom's mean position. Principal component analysis (PCA) was performed on these centered coordinates via singular value decomposition of their correlation matrix. This procedure yields, for each atom, three principal component vectors, u_1, u_2, and u_3, each of which encapsulates a variance s_1, s_2, and s_3, respectively, the sum of which is the total variance of the atom's trajectory. These values were placed back into our database for further analysis. The first principal component, u_1, which encapsulates the largest portion of the variance in the trajectory, was taken as the primary axis of flexibility while the standard deviation of the trajectory along that axis, $b = \sqrt{s_1}$, was taken as the primary measure of the flexibility in angstroms (Å). The flexibility vector for a given atom was thus taken to be bu_1, the vector in the direction of the first principal component whose length is the standard deviation of the movement along that axis. The total number of proteins/simulations analyzed was 253 (5.56 μs total) and the total number of atoms analyzed was 505,702 in 32,306 residues. These 253 targets include the 188 targets described in Table S1 of Beck et al. (2008), as well as the 65 targets listed in Table 2.1 here.

Once this flexibility information was collected and placed in the data warehouse, various statistical analyses and visual inspections of the trajectories were performed. Flexibility was visualized in two ways. The first involved plotting the flexibility vector (bu_1) for each atom onto the mean structure of the simulation; the vectors were also plotted in reverse because the principle component represents a trend along an axis with the atom at the origin. The

second method involved coloring the reference structure based on its calculated flexibility (b) along the sequence.

2.3.1 Analysis of Secondary Structure

Each secondary structure element was separated and categorized for analysis. Atoms were considered part of a secondary structure element if they existed in that element for at least 75% of the simulation according to the DSSP algorithm (Kabsch and Sander, 1983). Turns were determined according to the criteria outlined by Kuntz (1972) and labeled as such if the residue was not previously part of another secondary structure element and was in a turn conformation for at least 75% of the simulation. In the case of α-helices and β-sheets, the directions of the flexibility vectors were compared to the principal components of the Cα atoms of their respective secondary structure units (i. e., the consecutive Cα atoms belonging to a β-strand or α-helix).

2.3.2 Comparison of Fold Family Flexibility

Three fold families were examined to compare the flexibilities of family members: engrailed homeodomain three-helix bundles (3HB), Src homology 3 (SH3) domains, and ubiquitin-like folds (UBX). Twelve proteins from each family were analyzed, details of which can be found in Table 2.2. Correlations of flexibility were calculated for each pair of proteins in a single family using equivalent residue ranges based on the DaliLite server's alignment of the mean structures (Holm and Park, 2000).

2.3.3 Comparison of Native-State Flexibility to Early Unfolding Events

Unfolding trajectories were simulated at 498 K for at least 31 ns Day and Daggett (2005). Three proteins were chosen (*1enh*, *1shg*, and *1ubq*), one from

each fold family, each of whose native-state flexibility vectors were compared to their unfolding pathways.

2.4 Results

Here, we focus on the use of principal components of atomic trajectories to analyze the main chain Cα flexibility of proteins in MD simulations using a technique formally described by Teodoro et al. (2003). This technique provides the magnitude and primary axis of an atom's movements. We performed the analysis on all targets in our Dynameomics project for which we had completed at least 21 ns of simulation at 298 K and for which all of our standard analyses had been run, a total of 253 proteins when this project began.

2.4.1 General Properties of Flexibility

The data collected in the analysis of the 253 solvated protein MD simulations yielded several broad statistics concerning flexibility (Tab. 2.3). The distribution of flexibilities of all simulations can be seen in Figure 2.1a. Approximately 85% of the first principal components covered more than half of the variance of a given atom's trajectory. The distribution of the portion of variance covered by the first principal component is shown in Figure 2.1b. If atoms with very low flexibilities (< 0.5 Å) are excluded, 91% of the first principal components cover more than half of the variance. If only those atoms with higher than average flexibility are examined, this percentage climbs to 98%, and among the upper outliers (flexibility > 1.7 Å) the flexibility covers a mean of $76 \pm 10\%$ of the variance. The distribution of the flexibility of all atoms can be seen in Figure 2.1c. Less than a third of the variance in all atoms is covered by the final two principal components together. The ellipsoid formed by the standard deviations along each principal component for an atom represents that atom's probable

occupancy. The mean anisotropy of these ellipsoids, defined as the ratio of the shortest to the longest semi-axes of the ellipsoid, for Cα atoms in our simulation was 0.48, excluding upper outliers. This indicates that the atoms in our data set have distributions about their mean positions that are marginally less spherical than the experimentally derived data set from 68 proteins examined by Kondrashov et al. (2007), whose mean anisotropy was 0.51.

The average correlation of a protein's flexibility to the mean Cα root-mean-square fluctuation (RMSF) about the average was 0.74. The correlation between the average C root-mean-square deviation (RMSD) and the average flexibility was 0.75. The average correlation between Cα, Cγ, and Cζ flexibility and mean solvent-accessible surface area (SASA) by residue was 0.25, 0.33, and 0.47, respectively. The Spearman correlation between flexibility and hydrophobicity (Black and Mould, 1991) by amino acid type was 0.58; if Pro is excluded, this correlation rises to 0.65.

2.4.2 Properties of Secondary Structure Flexibility

In general, both β-strands and α-helices have flexibility vectors that are more parallel to their principal axes (i. e., stretching/compressing the structure) at their termini than in the middle. Histograms of the absolute values of the dot products of the flexibility vectors with the principal components of the secondary structure units, representing the degree of alignment of the vectors to the principal axes (1 indicating parallel vectors and 0 indicating perpendicular vectors), are shown in Figure 2.2. In the case of α-helices with at least two turns, the principal axis (the first principal component of the Cα atoms) of the helix is approximately parallel to the axis of the helix while the secondary and tertiary axes point outward toward the loops. In the case of β-strands, the principal axis of the strand lies along the backbone of the strand. A summary of the flexibilities of Cα atoms by residue and secondary structure can be found

in Table 2.4.

2.4.3 Fold Family Flexibility

We examined 12 proteins from each of three fold families: one all α-helical–the three-helix bundle fold (3HB), one all β-sheet – the SH3 fold family, and one with both an α-helix and β-sheet – the ubiquitin fold family (UBX) (Table 2.2).

The threeα-helix bundle fold (3HB) contains members that are among the fastest folding and unfolding proteins. Each protein contains relatively rigid α-helices and flexible loop regions. The mean C flexibility for the α-helices and loops is 0.76 ± 0.31 Å and 1.43 ± 0.83 Å, respectively. Residues of the α-helices flex perpendicular to the axis of the helix (Fig. 2.3a) in all cases except two helices of *1e17*. The residues flexing highly parallel in *1e17* are E13, L14, I15, Q17, A18, and I19 in the first helix (Fig. 2.3b) and L29, A30, Q31, I32, Y33, E34, and R38 in the second (Fig. 2.3c). Other helices in the 3HB family tend to contain Glu, Lys, Val, and Phe residues but fewer Leu and Ile residues. Table 2.6 shows the comparison of a collection of 3HB proteins with an average correlation of the magnitude of the flexibility of 0.76 with values ranging from 0.70 to 0.92. The final member is a significant outlier (*1kkx* vs. *1enh*) with $R = 0.38$.

The SH3 fold family consists of highly inflexible β-strands in barrel-like orientations. The mean flexibility of these β-strands is 0.47 ± 0.21 Å and of the loop regions is 1.29 Å ± 0.68 Å. No obvious global patterns exist in the directions of the flexibilities. Table 2.6 shows a collection of the SH3 family members compared to each other with an average correlation of flexibility magnitudes of 0.81 with values ranging from 0.73 to 0.96. There is a significant outlier with $R = 0.45$ (*1gcp* vs. *1ihv*).

The UBX fold contains both β-sheets and an α-helix; all of the UBX members studied have the helix docked against the β-sheet except for *1kot*, which additionally contains several external helices. The mean flexibility of the β-

strands is 0.53 ± 0.27 Å, and 0.53 ± 0.24 Å for α-helices. In each of the UBX proteins, the residues of the helix and strands exposed to each other flex more readily along the axis between them. Table 2.6 shows a collection of UBX family members compared to each other with an average correlation of flexibility magnitudes of 0.72 with values ranging from 0.61 to 0.77. A significant outlier with $R = 0.44$ (*1kot* vs. *1h8c*) is also included.

Overall, the correlation in flexibility between family members is highest when the sequence identity is high. The correlation between sequence identity and per-residue flexibility correlation is 0.76. The 3HB family members studied here have the lowest average sequence identity (Tab. 2.6). In addition to the comparisons provided, correlations were calculated between every possible pair of simulated proteins in a given family. The correlations tended to be high, with a small number of outlying low values. Of the 198 intrafamily protein pairs, only 11 had correlations below 0.1; excluding these, the average correlation between the flexibility magnitudes of two proteins in the same family was 0.62. Of the 11 pairs with low correlation, all belonged to either the 3HB or UBQ family, and their average sequence identity was $9\% \pm 6\%$.

2.4.4 Native-State Flexibility and Early Unfolding Events

The native-state flexibility of the engrailed homeodomain (*1enh*), of the 3HB family, was compared to early events in its thermal unfolding pathway (Fig. 2.4a). The protein can be broken down into segments in order of decreasing flexibility: its N terminus (1.48 Å); (H3C) the C-terminal end of H3; (L1) the flexible residue Y25 between H1 and H2; H3; (L2) the joint between H2 and H3; H2; and H1 (0.31 Å). The first significant unfolding event (within the first 0.1 ns) is the undocking of H3 in conjunction with the lifting of the flexible N-terminal tail (regions N and L2). This is followed by unwinding of the flexible C terminus (H3C). These events begin very early, around 0.3 ns, with

a stretching of the helix toward the C terminus, and they are complete by 3.5 ns, before the other two helices have begun to unwind significantly. Another early unfolding event is the movement of H1 from a position parallel to H2 to a skew position at approximately a right angle. The helices pivot around Y25 between 0.7 ns and 2.4 ns. The N-terminal end of H3 begins to unwind around 1.6 ns (H3 and L2) and is complete by 3.2 ns; the N-terminal end of H1 does not begin to unwind until 3.8 ns.

The native-state flexibility of the SH3 domain of alpha spectrin (*1shg*) was similarly compared to early events in its thermal unfolding (Fig. 2.4b). This protein is very inflexible overall (0.5 Å) and, in order of decreasing main-chain flexibility, consists of the C terminus (0.9 Å); (H) a single α-helical turn near the C terminus; (NT) the NT Src Loop; (DL) the Distal Loop; (RT) the RT Loop; and (S) four β-strands. The flexibility vectors of the end of the C terminus and of the helical turn have very strong components in the direction away from and toward the protein. The first event in the unfolding pathway of *1shg* (in the first 0.3 ns) is an extension of the C terminus (region C) away from the protein in the direction that the flexibility vectors point, accompanied by the undocking of RT (H and RT). From 0.1 to 0.2 ns, S2 and S3 (separated by NT) shift alignment. Around 0.3 ns, S4 pivots on DL and separates from S3. This is accompanied, around 0.5 ns, by the twisting of RT and the pivoting of S1 around RT. It is not until 0.8 ns that any of the β-strands bend significantly (S).

The protein ubiquitin (*1ubq*) is an inflexible protein (0.5 Å) consisting of four β-strands (S) and an α-helix (between S2 and S3) connected by four loops. Its most flexible regions are the four C-terminal residues (2.29 Å), (L1) Loop1, (L3) Loop3, and (L4) Loop4. The flexibility vectors of the C terminus point away from the body of the protein (Fig. 2.4c). By 0.6 ns of the simulation, the entire protein expands via the separation of S2 from S1 and the undocking of the helix from L4 via movement of L1 and L3. Although the C terminus is highly flexible

and moves considerably, it does not play a significant role in unfolding. Around 0.3 ns, L4 extends, eventually leading to the separation of S4 from S3 and S1 (between 0.5 and 0.7 ns).

2.4.5 Inflexible Loops

There are 21 loops or unstructured regions consisting of ≥ 6 residues in the ensemble of targets with an average Cα flexibility of ≥ 0.5 Å and an additional 353 moderately flexible unstructured regions with an average of < 1.0 Å. Seven of the highly inflexible loops are buried or partially buried in a protein, but 14 of them are exposed to solvent. Table 2.6 details these 21 regions; we highlight three of these regions below.

The ribosomal protein L14 (*1whi*) has a highly inflexible loop (mean flexibility is 0.47 Å) with sequence A11, D12, N13, S14, G15, A16, and R17 (Fig. 2.5a). A domain from bovine mitochondrial F1-ATPase (*1e1q*, residues 24-93), of the α/β-subunits F1 ATPase/thrombin family, has a highly inflexible region (0.37 Å) exposed to solvent with sequence L44, R45, N46, V47, Q48, A49, and E50 (Fig. 2.5b). A loop near the ice-binding surface of type III antifreeze protein from ocean pout (*1ops*, residues 2-65), of the β-clips II family, has a highly inflexible region with a mean flexibility of 0.45 Å and with sequence V26, T27, N28, P29, I30, G31, and I32 (Fig. 2.5b).

2.5 Discussion

2.5.1 General Properties of Flexibility

Large-scale MD flexibility analysis has never been applied to data mining on the scale of hundreds of proteins. By employing the basic technique of Teodoro et al. (2003) with our database of MD simulations, we have collected considerable information regarding the general flexibility of proteins, as well as uncov-

ered both anomalies and patterns concerning protein dynamics.

The distribution of the variance captured by the first principal components of the Cα trajectories and the correlation with the more conventional Cα RMSF supports the validity of using the first principal component of the trajectory as a measurement of an atom's flexibility. The most flexible Cα atoms have first principal components that cover the greatest portion of their total variance, and very inflexible Cα atoms have principal components that cover less of their variance. This observation suggests that the atoms for which flexibility analysis is most like RMSF are those that are least flexible. This observation additionally suggests that highly rigid atoms, such as those found in β-strands, undergo small fluctuations with less directed distributions about a mean position, while very flexible atoms, such as those in loops, oscillate along predictable trajectories. The primary difference between flexibility and RMSF is encapsulated in these observations; while RMSF measures all fluctuations from a mean structure, flexibility analysis isolates the key features of the motion of an atom. In addition to giving a direction to the atom's motion, flexibility filters out an atom's less significant and noisy motions and gives a measure of the fluctuation of an atom along its most significant mode. The distribution of flexibility shows that very few atoms are highly rigid compared to the number that are slightly flexible ($\geq$ 1 Å) and that a small number of atoms are very flexible ($\geq$ 5 Å), which occurs primarily in tails and loops.

The correlation between average Cα RMSD and flexibility shows that highly flexible proteins are very poorly captured by a small number of static structures and supports the notion that flexibility should be taken into account in docking and other structure analyses. The correlation between average Cα RMSF and flexibility is expected because of the underlying similarity in what they measure. This correlation supports the fact that they are related without suggesting that they are the same. One might expect a high correlation to SASA,

because surface residues would seem to be more mobile than buried residues. However, the correlations between SASA and flexibility are low because SASA is a very noisy measurement, although there is a higher correlation for side chain atoms.

The agreement of the anisotropy of atomic flexibilities to the anisotropy derived from crystallographic anisotropic displacement parameters, as examined by Kondrashov et al. (2007), strongly supports the validity of this flexibility metric. The slight decrease in the anisotropy of our simulations (0.48 vs. 0.51) may be due either to differences in the sampling of the data sets (68 vs. 253 proteins) or to the dynamical differences between atoms in solution and in crystals.

2.5.2 Properties of Secondary Structure Flexibility

The flexibility of individual amino acids by secondary structure tends to be highly variable due to the large data set and effects of averaging. A few exceptions to this emerge, however, notably the rigidity of His and Trp in β-strands or Ile in α-helices. These data suggest that the insertion of, for example, His into a β-strand or Ile into an α-helix would cause it to be more rigid. Additionally, it is apparent that the flexibilities of hydrophilic and polar residues are slightly higher on average than those of hydrophobic and nonpolar residues (Table 2), with a few exceptions. This trend can be easily explained by the tendency of nonpolar residues to cluster tightly with other nonpolar residues as opposed to polar and hydrophilic residues, which often interact with solvent. The correlation ($R_s = 0.65$, excluding Pro) between hydrophobicity and flexibility additionally supports this explanation. The appearance of cystine (Cys) as the least flexible amino acid is not surprising because we separated reduced cysteine (Cyh) and oxidized cystine.

The dot products of the Cα atoms with the principal axes of their secondary structure measure the angle that the motion of the atom makes with the sec-

ondary structure element. Both the principal axis of a secondary structure unit and the flexibility vector for any given Cα atom are unit vectors; thus the dot product will always range from 0 (perpendicular vectors) to 1 (parallel vectors). The distributions of these dot products are noisy due to the relative rigidity of secondary structure combined with the previous observation that rigid atoms have less ordered distributions about a mean than highly flexible atoms, which tend to flex more strongly along a single axis. Nonetheless, slight trends are apparent. In both α-helices and β-strands, there is a slight tendency for the flexibility vectors of a secondary structure unit's Cα atoms to be perpendicular to its primary axis and for its second principal component to be parallel to the flexibility vectors. In the case of α-helices, this trend indicates that the flexibility vectors point most strongly outward/inward, away from and toward the center of the helix. In the case of β-strands, this trend indicates that flexibility vectors point least strongly along the backbone of the strand and more strongly in the direction of the bends of the backbone (the direction of the second principal axis) than from side to side. The trend is more pronounced in α-helices than β-strands, which can be predicted by the higher flexibility of α-helices as well as the tendency of β-strands to curve and bend (thereby preventing the principal axis from being as consistent). Additionally, in the case of α-helices, the trend is slightly more pronounced at the ends of helices than for the middle residues, showing that the ends of helices flex more readily outward from the central axis.

2.5.3 Trends in Fold Family Flexibility

The examination of the flexibilities of fold families begs the question of whether there are fundamental rules that tie sequence and local structure to flexibility. The average flexibilities of secondary structure within a fold family differ from the overall averages, suggesting that some trends between the flexibilities of

members of various fold families exist. The secondary structure of the 3HB and SH3 fold families consist only of α-helices and only of β-sheets, respectively. In both 3HB and SH3, the flexibilities of their secondary structures are lower than expected from the overall averages. The lack of trends in the flexibility vectors in the SH3 fold family, which is rich in β-strands, agrees with the previous observation that β-strands are highly inflexible and therefore tend to have less directed fluctuations. These data, along with the observations concerning the trends in the directions of the flexibility vectors in the 3HB and UBX families and the high correlations between fold family members, suggest that there are motifs in the specific flexibilities of fold families, though these trends may be subtle. Additionally, the correlation between a protein's sequence identity and the closeness of its flexibility to other family members suggests that sequence modulates the flexibility. For example, *1e17*, whose helical flexibility vectors varied from the other members of the 3HB family, has helical sequences that are quite different from the other family members; features such as the lack of Lys and presence of Ile, a highly inflexible residue in α-helices, explain some of these differences.

Additionally, the high correlations of the magnitude of the flexibility between equivalent structural regions of family members suggest that families have characteristic flexibility patterns. Notably, comparison of arbitrary α-helices to each other and arbitrary β-strands to each other produces very low correlations (mean correlation < 0.2), so the relationship observed here is not dependent only on the makeup of secondary structure in each family. Not every member of a fold family adheres strictly to these flexibility patterns, however, as shown by the small number of pairs of proteins with low flexibility correlations. This is not surprising considering the structure and sequence diversity of the fold families examined and demonstrates that the local chemical environment of a residue, and not just its local backbone configuration and chain

topology, determine its flexibility. Nonetheless, the high correlation between most pairs of fold family members indicates that the similarity between two proteins' flexibilities correlates with the similarity of their structures and sequences. Future work will extend this observation to examine in detail how the local chemical environment of a residue influences its flexibility.

2.5.4 Native=State Flexibility and Early Unfolding Events

The comparison to unfolding simulations shows a relationship between the flexibility of a residue at 298 K and the early steps in the thermal unfolding pathway in the proteins examined here. There is a nearly step-by-step correlation between high flexibility and the order of unfolding. These data suggest that native-state dynamics are closely related to unfolding and folding dynamics, in agreement with our findings in the first simulations of protein unfolding (Daggett and Levitt, 1992). Later, Hespenheide et al. (2002) explored the relationship between flexibility and unfolding pathways in simulations of 10 monomeric proteins and compared the results to hydrogen-deuterium exchange experiments (Li and Woodward, 1999). They found that the folding cores of proteins with the greatest structural stability against denaturation could be determined by flexibility. Here we extend this work to the level of hundreds of proteins, further tying flexibility to instability by showing that flexible Cα sites are the most likely candidates for early unfolding.

2.5.5 Inflexible Loops

The sheer number of inflexible loops (21 with flexibility 0.5 Å and 353 with flexibility < 1.0 Å) is surprising and suggests that there may be a number of structured loops that are not recognized as secondary structure, although they are as rigid as conventional secondary structure. Because the inflexible secondary

structure units that form a protein's backbone and core generally determine its structure, it is useful to consider the possibility of additional rigid structural units that may be important in the determination of structure. This hypothesis was examined before by Leszczynski and Rose (1986) in their study of Ω-loops. While many of the loops examined here have occasional Ω character, partially due to the broad definition of Ω-loops, many of them do not share the motif of being tightly packed internally. The three cases examined here are each interesting for different reasons. The second loop in *1whi* (ribosomal protein L14 family) contains a pair of hydrogen bonds between the side chains of Asp 12 and Ser 14. Notably, this loop is highly conserved among species and is responsible for mediating interactions between the neighboring loops in the β-barrel of this protein (Davies et al., 1996). The loop in *1e1q* (α/β subunits of F1 ATPase family) sits at the interface between and subunits of ATPase (Abrahams et al., 1994) and contains a pair of hydrophobic residues (Leu 44 and Val 47) in close proximity, forming a small hydrophobic cluster. Such interactions indicate that sequence is an important predictor of flexibility. The loop in *1ops* (antifreeze protein III-like family) does not contain hydrogen bonds or sites for potential hydrophobic interactions, though it is highly hydrophobic. The C-terminal end of the protein runs through it, however, which may lock it down. This loop provides rigid support for the ice-binding surface of this antifreeze protein (Yang et al., 1998).

Each of the 21 least flexible regions fit into one of five categories: (1) those that are sterically hindered, (2) those with internal hydrophobic contacts, (3) those with internal polar contacts, (4) those with partial secondary structure character, and (5) those with external contacts. Internal hydrophobic contacts between side chains appear to play a stabilizing role in many of these, and such contacts represent the most commonly appearing motif. These loops specifically coincide with the Ω-loop motif. The regions with hydrophobic internal

contacts often had backbones with characteristic and very high curvature such that close contacts could be made between side chains at the center. The distribution of amino acids in all of the inflexible loop regions, as well as in only the 21 least flexible regions, contained no significant deviations from the distribution of amino acids in all proteins; however, the distribution of amino acids in those regions with internal hydrophobic contacts was heavily skewed toward Leu and Pro and moderately skewed toward Val and Ile. Gly and Thr are the only other amino acids with a high frequency in this set. Pro appears near the point of highest curvature in several of these regions and may be important for forming this motif by introducing a kink in the segment. Future analysis will explore the extent to which these observations are examples of structural motifs that imply a predictable quality to flexibility based on sequence and structure and whether these potential motifs can further be tied to structural stability.

2.6 Conclusions

Protein flexibility is a useful means of extracting information from individual protein trajectories as well as related sets of trajectories. Protein flexibility bears a strong relationship to unfolding and can be used to predict early steps in unfolding. The ability of flexibility to elucidate regions of interesting structure has been demonstrated by the identification of inflexible loops that constitute new structural motifs. Finally, the correlation of flexibility with structure and the inherent flexibility differences between fold families are potentially very useful for understanding how different arrangements of structure can lead to different dynamics and function.

Table 2.1: Sixty-five protein targets analyzed in addition to the 188 targets in Table S1 of Beck et al. (2008)

PDB Code	Description
1a1x	HMTCP-1 Chain A
1a3a	IIA Mannitol From E. coli
1bgw	Topoisomerase Residues 410-1202
1bm0	Human Serum Albumin
1cd5	Glucosamine-6-Phosphate Deaminase From E. coli, T Conformer
1cfe	NMR Structure Of P14A
1ciy	Insecticidal Toxin
1crz	E. coli TOLB Protein
1d0b	Internalin B Leucine Rich Repeat Domain
1dd5	Thermotoga Maritima Ribosome Recycling Factor, RRF
1dhn	7,8-Dihydroneopterin Aldolase from Staphylococcus aureus
1dx7	Light-Harvesting Complex 1 Beta Subunit From Rhodobacter sphaeroides
1dxk	Metallo-Beta-Lactamase from Bacillus cereus 569/H/9 C168S Mutant
1dzo	Truncated PAK Pilin from Pseudomonas aeruginosa
1e17	DNA Binding Domain of the Human Forkhead Transcription Factor AFX (FOXO4)
1ef1	Moesin Ferm Domain/Tail Domain Complex
1epu	Neuronal SEC1 from Squid
1ey1	E. coli NUSB
1f43	MATA1 Homeodomain
1f7t	Holo-(Acyl Carrier Protein) Synthase
1fhq	FHA2 Domain of RAD53
1fna	Tenth Type III Cell Adhesion Module of Human Fibronectin

Table 2.1: (continued)

PDB Code	Description
1fuo	Fumarase C with Bound Citrate
1fva	Bovine Methionine Sulfoxide Reductase
1fx2	Adenylate Cyclases from Trypanosoma brucei
1fyv	TIR Domain of Human TLR1
1g03	N-Terminal Domain of HTLV-I CA1-134
1g61	M. jannaschii EIF6
1gc7	Radixin Ferm Domain
1gcp	VAV SH3 Domain
1gef	Archaeal Holliday Junction Resolvase HJC
1gl5	SH3 Domain from the TEC Protein Tyrosine Kinase
1gso	Glycinamide Ribonucleotide Synthetase (GAR-SYN) from E. coli
1h8c	UBX Domain from Human FAF1
1h8h	Bovine Mitochondrial F1-ATPase
1hf8	N-Terminal Domain of Clathrin Assembly Lymphoid Myeloid Leukaemia Protein
1hic	Hirudin (1-51)
1hpl	Horse Pancreatic Lipase
1i42	UBX Domain from P47
1igp	Recombinant Inorganic Pyrophosphatase from E. coli
1ihc	Gephyrin N-terminal Domain
1ihv	DNA Binding Domain of HIV-1 Integrase
1ijy	Cysteine-Rich Domain of Mouse Frizzled 8 (MFZ8)
1ile	Isoleucyl-TRNA Synthetase

Table 2.1: (continued)

PDB Code	Description
1jaw	Aminopeptidase P from E. coli Low pH Form
1kkx	DNA-binding domain of ADR6
1kot	Human GABA Receptor Associated Protein GABARAP
1kra	Klebsiella Aerogenes Urease
1mmo	Monooxygenase Oxidoreductase
1qau	Oxidoreductase
1qcv	Rubredoxin Variant (PFRD-XC4) Folds Without Iron
1qk9	Domain from MECP2 That Binds to Methylated DNA
1qly	SH3 Domain from Brutons Tyrosine Kinase
1ryu	SWI1 ARID DNA-Binding Protein
1sgk	Nucleotide-Free Diphtheria Toxin
1swb	Apo-Core-Streptavidin
1tnr	Soluble Human 55 kD TNF Receptor-Human TNF-Beta Complex
1tx4	RHO/RHOGAP/GDP(DOT)ALF4 Complex
1whi	Ribosomal Protein L14
2a36	N-terminal SH3 domain of DRK
2dik	R337A Mutant of Pyruvate Phosphate Dikinase
2lis	Red Abalone Lysin Monomer
3fib	Recombinant Human Gamma-Fibrinogen Carboxyl Terminal Fragment (Residues 143-411)
3gb1	B1 Domain of Streptococcal Protein G
7hsc	Heat Shock Cognate-70 kD Substrate Binding Domain

Table 2.2: Proteins that were analyzed and compared by fold family

Fold Family	PDB Code	Description
3HB	*1e17*	DNA Binding Domain of the Forkhead Transcription Factor AXF
3HB	*1f43*	MATA1 Homeodomain
3HB	*1kkx*	DNA Binding Domain of ADR6
3HB	*1ryu*	SWI1 ARID
3HB	*1enh*	Engrailed Homeodomain
3HB	*1bw6*	Human Centromere Protein B (Cenp-b) DNA Binding Domain RP1
3HB	*1ba5*	DNA Binding Domain of Human Telomeric Protein, HTRF1
3HB	*1du6*	Truncated PBX Homeodomain
3HB	*1apl*	Mat α2 Homeodomain
3HB	*1ret*	DNA Binding Domain of $\gamma\delta$ Resolvase
3HB	*1bw5*	Homeodomain of Rat Insulin Gene Enhancer Protein ISL-1
3HB	*1ug2*	Mouse 2610100b20rik Hypothetical Gene Product
SH3	*1gcp*	Sulfite Reductase Hemoprotein
SH3	*1gl5*	SH3 Domain from TEC Protein Tyrosine Kinase
SH3	*1shf*	SH3 Domain in Human FYN
SH3	*1shg*	SH3 Domain of α Spectrin
SH3	*1ihv*	DNA Binding Domain of HIV-1 Integrase
SH3	*1qly*	SH3 Domain from Brutons Tyrosine Kinase

Table 2.2: (continued)

Fold Family	PDB Code	Description
SH3	*2a36*	N-terminal SH3 Domain of DRK
SH3	*1ujy*	SH3 Domain in RAC/CDC42 Guanine Nucleotide Exchange Factor 6
SH3	*1ugv*	SH3 Domain of Human Olygophrein-1-like Protein
SH3	*1cka*	N-Terminal SH3 Domain of C-crk
SH3	*2hsp*	SH3 Domain of Phospholipase Cγ
SH3	*1spk*	RSGI RUH-010
UBX	*1h8c*	Ubiquitin-like domain from FAF1
UBX	*1i42*	Ubiquitin-like domain from P47
UBX	*1kot*	Human GABA Receptor Associated Protein GABARAP
UBX	*1ubq*	Ubiquitin
UBX	*1a5r*	Sumo-1
UBX	*1ef5*	Ras-Binding Domain of RGL
UBX	*1rlf*	Ras-Binding Domain of RLF
UBX	*1iyf*	Ubiquitin-like Domain of Human Parkin
UBX	*1j8c*	Ubiquitin-like Domain of HPLIC-2
UBX	*1v5t*	Ubiquitin-like Domain of Mouse Hypothetical 8430435i17rik Protein
UBX	*1gb4*	Hypothetical Variant of the B1 Domain from Streptococcal Protein G
UBX	*1ssn*	Sakstar Variant of Staphylocinase

Table 2.3: Flexibilities for various atom groups over all simulations analyzed.

Atom Group	Mean (Å)
All atoms[1]	1.257 ± 0.94
Cα	1.009 ± 0.76
Cγ	1.250 ± 0.84
Backbone Atoms[1]	1.013 ± 0.75
Side-chain Atoms[1]	1.332 ± 0.97

[1] Hydrogen atoms were not included.

Table 2.4: Flexibility of Cα atoms by secondary structure and residue.

Residue	β-strand, parallel	β-strand, anti-parallel	α-helix	loop/none	turn	overall mean
GLY	0.56 ± 0.24	0.65 ± 0.32	0.84 ± 0.40	1.35 ± 0.90	1.20 ± 0.62	1.22 ± 0.77
ALA	0.52 ± 0.19	0.61 ± 0.29	0.86 ± 0.53	1.30 ± 0.99	1.31 ± 0.82	1.04 ± 0.71
VAL	0.51 ± 0.21	0.59 ± 0.33	0.79 ± 0.43	1.11 ± 0.86	1.15 ± 0.64	0.86 ± 0.58
LEU	0.48 ± 0.22	0.60 ± 0.33	0.81 ± 0.59	1.01 ± 0.65	1.00 ± 0.49	0.86 ± 0.56
ILE	0.51 ± 0.24	0.61 ± 0.28	0.72 ± 0.34	1.06 ± 0.80	1.07 ± 0.63	0.82 ± 0.50
SER	0.59 ± 0.34	0.67 ± 0.37	0.86 ± 0.48	1.37 ± 1.02	1.32 ± 0.78	1.18 ± 0.82
THR	0.49 ± 0.20	0.68 ± 0.35	0.86 ± 0.52	1.18 ± 0.85	1.18 ± 0.57	1.02 ± 0.67
CYS	0.53 ± 0.18	0.59 ± 0.25	0.50 ± 0.22	0.88 ± 0.42	0.93 ± 0.30	0.77 ± 0.35
MET	0.50 ± 0.14	0.62 ± 0.28	0.90 ± 0.80	1.34 ± 1.25	1.28 ± 0.75	1.08 ± 0.94
PRO	0.61 ± 0.29	0.64 ± 0.23	0.91 ± 0.40	1.24 ± 0.92	1.10 ± 0.53	1.16 ± 0.80
ASP	0.59 ± 0.27	0.65 ± 0.30	0.90 ± 0.47	1.19 ± 0.83	1.40 ± 0.92	1.11 ± 0.73
ASN	0.53 ± 0.19	0.63 ± 0.23	0.86 ± 0.49	1.20 ± 0.79	1.23 ± 0.62	1.09 ± 0.67
GLU	0.53 ± 0.21	0.64 ± 0.39	0.86 ± 0.52	1.27 ± 1.04	1.18 ± 0.70	1.04 ± 0.75
GLN	0.62 ± 0.28	0.68 ± 0.35	0.92 ± 0.52	1.22 ± 0.92	1.25 ± 0.71	1.05 ± 0.69
HIS	0.46 ± 0.11	0.72 ± 0.47	0.91 ± 0.71	1.27 ± 1.05	1.28 ± 0.77	1.10 ± 0.87
LYS	0.52 ± 0.20	0.67 ± 0.38	0.84 ± 0.45	1.19 ± 0.87	1.17 ± 0.78	1.01 ± 0.66
ARG	0.58 ± 0.23	0.60 ± 0.36	0.81 ± 0.46	1.18 ± 0.84	1.13 ± 0.58	0.96 ± 0.61
PHE	0.57 ± 0.32	0.60 ± 0.27	0.79 ± 0.46	1.11 ± 0.86	1.05 ± 0.66	0.88 ± 0.59
TRP	0.47 ± 0.14	0.61 ± 0.28	0.78 ± 0.42	1.13 ± 0.75	1.23 ± 0.88	0.92 ± 0.56
TYR	0.62 ± 0.25	0.60 ± 0.33	0.83 ± 0.50	1.11 ± 0.85	0.97 ± 0.54	0.91 ± 0.62
CYH	0.68 ± 0.24	0.75 ± 0.52	0.79 ± 0.39	1.04 ± 0.57	0.86 ± 0.50	0.88 ± 0.48
Overall Mean	0.53 ± 0.23	0.63 ± 0.33	0.84 ± 0.50	1.21 ± 0.88	1.19 ± 0.67	1.01 ± 0.67

Table 2.5: Flexibility correlation between various fold family members.

Fold	Protein 1	Protein 2	Equivalent Residue Ranges[1]	Sequence Identity	Correlation
3HB	*1enh*	*1e17*	3-23 ⇔ 93-113; 30-39 ⇔ 119-128; 41-53 ⇔ 146-158	7%	0.70
3HB	*1enh*	*1f43*	3-6 ⇔ 1-4; 7-55 ⇔ 7-55	4%	0.70
3HB	*1enh*	*1kkx*	7-9 ⇔ 34-36; 12-25 ⇔ 37-50; 26-40 ⇔ 55-69; 41-51 ⇔ 71-81; 52-55 ⇔ 83-86	9%	0.38
3HB	*1enh*	*1ryu*	3-7 ⇔ 1-5; 12-21 ⇔ 51-60; 22-25 ⇔ 63-66; 26-39 ⇔ 69-82; 41-56 ⇔ 93-108	18%	0.92
3HB	*1e17*	*1f43*	96-100 ⇔ 12-16; 101-110 ⇔ 18-27; 117-130 ⇔ 28-41; 134-137 ⇔ 42-45; 144-155 ⇔ 47-58	4%	0.77
3HB	*1e17*	*1kkx*	117-129 ⇔ 11-23; 139-146 ⇔ 74-81; 147-154 ⇔ 83-90; 156-159 ⇔ 91-94; 177-181 ⇔ 99-103	5%	0.72
3HB	*1e17*	*1ryu*	101-113 ⇔ 50-62; 116-119 ⇔ 63-66; 121-138 ⇔ 71-88; 144-157 ⇔ 91-104; 158-161 ⇔ 106-109; 177-181 ⇔ 112-116	9%	0.74

[1] Structural alignments based on the mean backbone structure with DaliLite were used to find equivalent residue ranges.

Table 2.5: (continued)

Fold	Protein 1	Protein 2	Equivalent Residue Ranges[1]	Sequence Identity	Correlation
SH3	*1shg*	*1gcp*	12-20 ⇔ 598-606; 20-34 ⇔ 612-626; 39-45 ⇔ 634-640; 54-62 ⇔ 649-657	19%	0.83
SH3	*1shg*	*1gl5*	6-47 ⇔ 180-221; 48-62 ⇔ 223-237	25%	0.78
SH3	*1shg*	*1shf*	6-46 ⇔ 84-124; 47-62 ⇔ 127-142	33%	0.96
SH3	*1shg*	*1ihv*	9-10 ⇔ 225-227; 13-14 ⇔ 230-231; 30-37 ⇔ 238-245; 42-57 ⇔ 250-265	18%	0.84
SH3	*1shg*	*1qly*	6-46 ⇔ 1-41; 47-61 ⇔ 43-58	32%	0.95
SH3	*1shg*	*2a36*	7-37 ⇔ 1-31; 38-61 ⇔ 33-57	24%	0.93
SH3	*1gcp*	*1gl5*	596-605 ⇔ 183-195; 614-631 ⇔ 195-212; 634-642 ⇔ 214-222; 645-655 ⇔ 224-234	22%	0.70
SH3	*1gcp*	*1shf*	599-600 ⇔ 88-89; 621-627 ⇔ 105-111; 633-655 ⇔ 117-139	28%	0.82
SH3	*1gcp*	*1ihv*	598-599 ⇔ 225-226; 603-604 ⇔ 230-231; 624-631 ⇔ 238-245; 636-640 ⇔ 246-250; 643-655 ⇔ 256-268	5%	0.45
SH3	*1gcp*	*1qly*	602-603 ⇔ 6-7; 608-609 ⇔ 12-13; 611-629 ⇔ 13-31; 636-640 ⇔ 35-39; 646-649 ⇔ 43-46; 659-660 ⇔ 56-57	24%	0.73
SH3	*1gcp*	*2a36*	596-606 ⇔ 2-12; 614-642 ⇔ 14-42; 645-660 ⇔ 43-58	16%	0.96

[1] Structural alignments based on the mean backbone structure with DaliLite were used to find equivalent residue ranges.

Table 2.5: (continued)

Fold	Protein 1	Protein 2	Equivalent Residue Ranges[1]	Sequence Identity	Correlation
UBX	*1ubq*	*1h8c*	1-37 ⇔ 6-42; 39-45 ⇔ 43-49; 46-49 ⇔ 52-55; 52-75 ⇔ 59-82	14%	0.77
UBX	*1ubq*	*1i42*	1-6 ⇔ 295-300; 9-15 ⇔ 305-311; 17-25 ⇔ 312-320; 27-35 ⇔ 321-329; 40-46 ⇔ 336-342; 47-53 ⇔ 347-353; 55-70 ⇔ 354-369	13%	0.68
UBX	*1ubq*	*1kot*	1-7 ⇔ 31-37; 11-38 ⇔ 48-75; 40-59 ⇔ 77-96; 60-63 ⇔ 100-103; 64-72 ⇔ 106-114	6%	0.61
UBX	*1kot*	*1i42*	1-4 ⇔ 282-285; 27-37 ⇔ 291-301; 43-55 ⇔ 303-315; 57-72 ⇔ 316-331; 76-80 ⇔ 333-337; 81-84 ⇔ 340-343; 88-91 ⇔ 350-353; 105-110 ⇔ 361-366	10%	0.72
UBX	*1kot*	*1h8c*	29-38 ⇔ 5-14; 45-73 ⇔ 14-42; 76-83 ⇔ 43-50; 84-88 ⇔ 53-57; 89-97 ⇔ 61-69; 106-117 ⇔ 71-8	25%	0.44

[1] Structural alignments based on the mean backbone structure with DaliLite were used to find equivalent residue ranges.

Table 2.6: Inflexible loop regions of proteins.

PDB	Residue Range	<Cα Flexibility>	Proposed Explanation
1fkb	64-70	0.33	The loop is sterically hindered by a more superficial loop.
1gpr	14-20	0.37	Region is internal to the protein and not exposed to solvent.
1e1q	44-50	0.37	Hydrophobic contact between V47 and L44
1ops	34-43	0.41	Region fluctuates between 3/10 and -helix.
1gpr	151-157	0.42	R152 and E153 side-chains form polar and H-bond network with nearby β-strands.
1whi	11-17	0.42	D12 and S14 form an internal H-bond.
1vmo	26-32	0.42	Y30 forms external polar contact K134; region is hindered by tight curvature.
1g61	12-19	0.43	Region contains internal tightly-packed hydrophobic contacts.
1ops	17-22	0.44	Region contains internal tightly-packed hydrophobic contacts.
1ops	50-59	0.45	Region has occasional β-strand character is contains H-bonds to nearby loop and β-strand.
1ops	26-32	0.45	Loop contains tightly-packed hydrophobic side-chains and is hindered by N-terminus, which runs through it.

Table 2.6: (continued)

PDB	Residue Range	<Cα Flexibility>	Proposed Explanation
1ris	81-86	0.45	R82 forms internal salt-bridge with N84 and external salt-bridge with E22.
3fib	215-229	0.46	Loop is structurally fixed by several internal hydrophobic contacts and held in place by contacts with nearby β-strands.
2hnp	257-263	0.47	Hydrophobic center of loop sits in a hydrophobic pocket where it is stabilized by nearby α-helices.
1chu	271-277	0.47	D273 forms polar contact with T274.
1fqn	29-38	0.47	Region has occasional β-strand character and makes H-bonds with nearby β-strands.
1dyw	131-136	0.48	Loop contacts and is sterically hindered by superficial loop.
1whi	88-103	0.49	Region containes several internal hydrophobic contacts and external polar contacts.
1e1q	64-70	0.49	Region contains internal hydrophobic contacts and polar contact between E67 and N65.
1ubq	16-22	0.49	Loop's hydrophobic interior is surrounded by polar side-chains.
1ge8	54-64	0.5	K60-V64 form occasional β-sheet; P54-S59 have tightly-packed hydrophobic core.

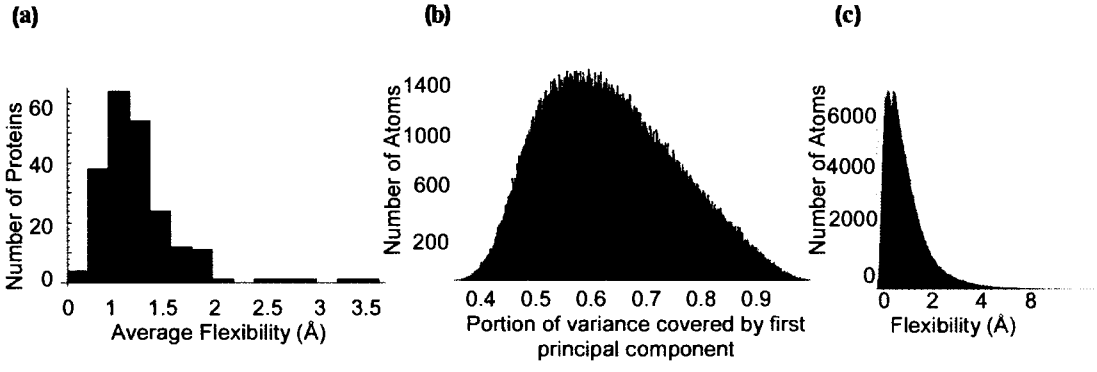


Figure 2.1: General properties of protein flexibility. **(a)** Histogram of proteins by average flexibility (square root of the variance represented by the first principal component of an atom's trajectory). **(b)** Histogram of the portion of the variance covered by the first principal component of each atom's trajectory. High coverage means that most of the movement of that atom is encapsulated by its flexibility. **(c)** A histogram of the flexibilities of all atoms analyzed.

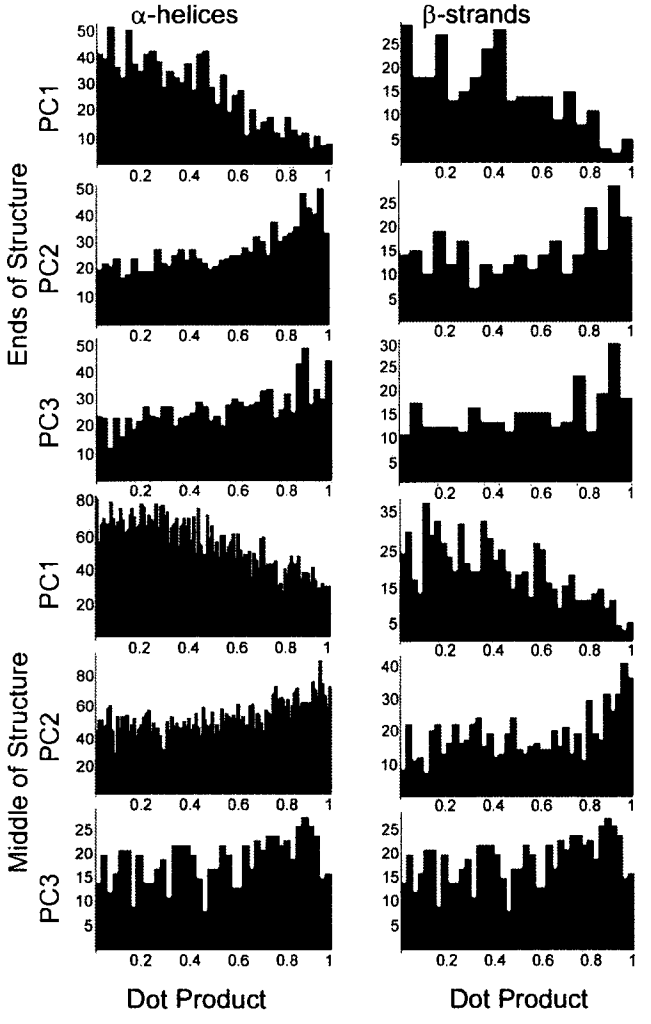


Figure 2.2: Histograms of the absolute values of the dot products of the principal axes of secondary structure elements with the end or middle residues of each. A dot product of 1 indicates parallel vectors while a dot product of 0 indicates perpendicular vectors. The y-axis of each graph is the number of proteins, while the x-axis is the dot product.

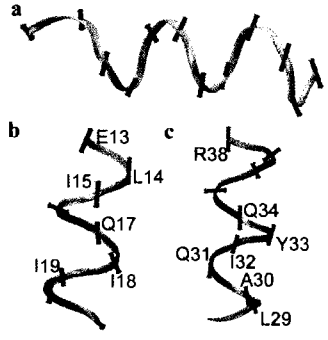
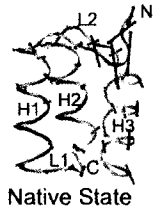


Figure 2.3: Protein backbones with flexibility vectors shown as vectors with lengths equal to the Cα flexibility in angstroms. **(a)** An α-helix of *1f43* with flexibility vectors perpendicular to the principal axis of the helix. **(b)** First α-helix of *1e17* with flexibilities parallel to the principal axis of the helix. **(c)** Second α-helix of *1e17* with flexibilities parallel to the principal axis of the helix. Backbones are colored black to white by flexibility with darker regions being the least flexible.

Figure 2.4: Flexibility representation, unfolding snapshots, and significant unfolding events of three proteins. All proteins are colored blue-green-red by flexibility magnitudes. Flexibility vectors are shown in red. Vectors are displayed at twice their length for clarity. **(a)** Engrailed homeodomain (*1enh*); the transition state is at 362 ps. **(b)** α-Spectrin (*1shg*); the transition state is at 110 ps. **(c)** Ubiquitin (*1ubq*); the transition state is at 611 ps.

(a) Engrailed Homeodomain

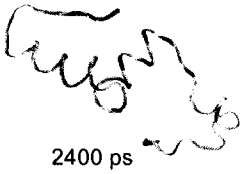


Native State

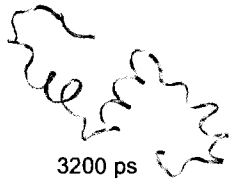
362 ps

600 ps

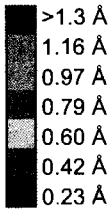
1600 ps



2400 ps

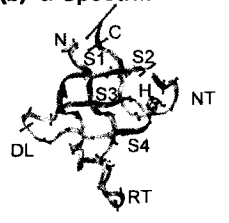


3200 ps



Time (ns)	Event	Segment	Residues	<C_α Flexibility> (Å)
0.1	Movement of N-terminus,	N	R3 – S9	1.48
	Undocking of H3	L2	G39 – N41	0.5
0.3-3.5	Unwinding of H3C	H3C	Q50 – I56	0.65
0.7-2.4	Pivoting of H1 and H2	Y25	Y25	0.63
1.6-3.2	Unwinding of H3 N-terminal end	H3	E42 – I56	0.59
		L2	G39 – N41	0.5
3.8-4.5	Unwinding of H1 N-terminal end	H1N	S10 – Q12	0.46
4.1-5.1	Unwinding of H3 middle	H3	E42 – I56	0.59

(b) α-Spectrin

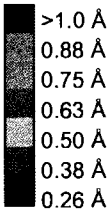


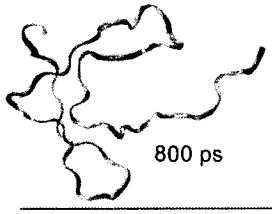
Native State

110 ps

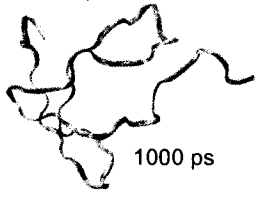
400 ps

600 ps





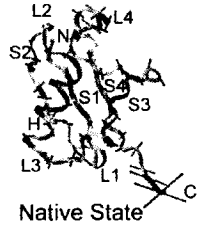
800 ps



1000 ps

Time (ns)	Event	Segment	Residues	<C_α Flexibility> (Å)
0-0.3	Extension of C	C	V53 – D57	0.90
	Undocking of RT	RT	Y8 – M20	0.52
		H	P49 – Y52	0.60
0.1-0.2	Alignment change of S2 and S3	NT	N30 – W36	0.60
0.3-1.0	Pivoting of S4 Separation of S3 and S4	DL	V41 – R44	0.57
0.5-0.8	Twisting of RT Pivoting of S1	RT	Y8 – M20	0.52
0.8-1.0	Distortion of S1	S1	K1 – L7	0.35

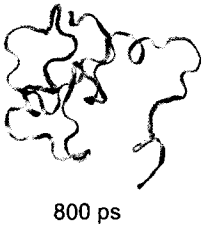
(c) Ubiquitin



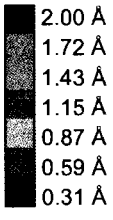
Native State

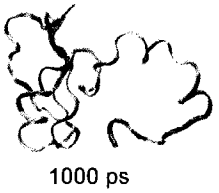
400 ps

611 ps

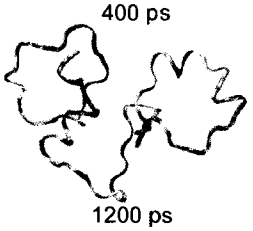


800 ps





1000 ps



1200 ps

Time (ns)	Event	Segment	Residues	<C_α Flexibility> (Å)
0-0.6	Separation of S2 from S1,	L1	L8 – T9	1.27
	Separation of H from L4	L3	I35 – D39	0.95
0.3-0.6	Expansion of L4	L4	A46 – S65	0.78
0.5-0.7	Separation of S3 and S4	L4	A46 – S65	0.78
0.6-0.7	Separation of S4 and S1	L4	A46 – S65	0.78
0.8-1.2	Formation of α-helix at S4 C-Terminal End	S4	T66 – L71	0.47
		C	R72 – G76	2.29

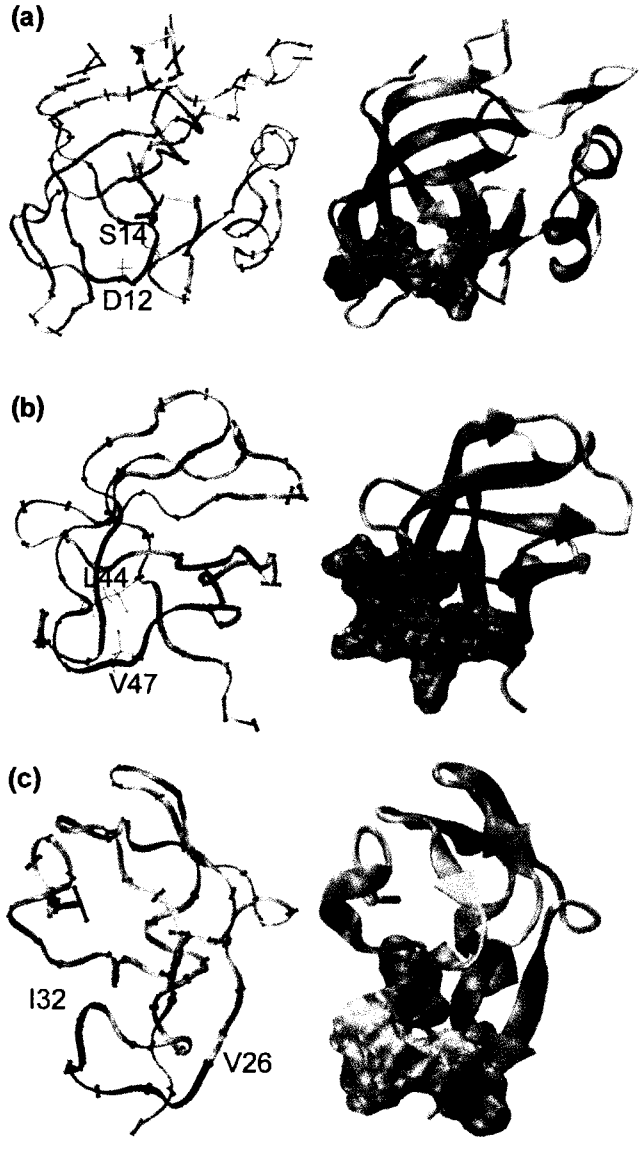


Figure 2.5: Three proteins with inflexible regions with each inflexible region colored in black. In each row, the right column contains the protein with the surface of the inflexible region shown colored black to white by Cα flexibility with darker regions being the least flexible. **(a)** The ribosomal protein L14. The inflexible loop begins with residue 11 on the left and loops around to residue 17 on the right. Side chains for S12 and D14 are shown and colored by atom. Hydrogen bonds are shown by dotted lines. **(b)** Bovine mitochondrial F1-ATPase. Side chains for residues L44 and V47 are displayed. **(c)** Ocean pout antifreeze III protein.

Chapter 3

WAVELET ANALYSIS OF PROTEIN MOTION

3.1 Summary

As high-throughput molecular dynamics simulations of proteins become more common and the databases housing the results become larger and more prevalent, more sophisticated methods to quickly and accurately mine large numbers of trajectories for relevant information will have to be developed. One such method, which is only recently gaining popularity in molecular biology, is the continuous wavelet transform, which is especially well-suited for time course data such as molecular dynamics simulations. We describe techniques for the calculation and analysis of wavelet transforms of molecular dynamics trajectories in detail and present examples of how these techniques can be useful in data mining. We demonstrate that wavelets are sensitive to structural rearrangements in proteins and that they can be used to quickly detect physically relevant events. Finally, as an example of the use of this approach, we show how wavelet data mining has led to novel discoveries that appear to be related to the mechanism of the protein $\gamma\delta$ resolvase.

3.2 Introduction

Molecular dynamics (MD) has become a common method for studying the motion of proteins over time, and it is the only available technique for examining continuous fine granularity motion at atomic resolution. By numerically integrating Newton's equations of motion, one can produce a series of snapshots of a protein's trajectory through time. These snapshots, when saved at suffi-

ciently high resolution, serve as stop-motion photography and provide a great deal of information about how proteins behave.

The Dynameomics project (Beck et al., 2008; van der Kamp et al., 2010) is a large-scale MD effort to simulate a representative from every protein fold family (Day and Daggett, 2003). The Dynameomics database (Kehl et al., 2008; Simms et al., 2008) currently contains over 2200 proteins, each of which has been simulated for at least 31 ns at a temperature of 298 K. Additionally, it contains at least two unfolding simulations of each protein at 498 K for 31 ns and at least three short (2 ns) simulations at 498 K for a total of ~11,000 simulations. These simulated target proteins are selected from our updated consensus domain dictionary (Schaeffer et al., 2010) based on procedures developed by Day et al. (2003) These targets constitute a data set that spans a considerable portion of the protein universe, representing more than 80% of all known protein domains. The majority of the remaining 20% of the domains are not in fact autonomous self-contained folds. Consequently, the simulation portion of the Dynameomics project is complete; thus we now turn to mining and using this database.

Because of the incredible amount of information stored in the Dynameomics database, which contains 104 times as many structures as the Protein Data Bank (PDB) Berman et al. (2000), analysis is often challenging. Although a vast array of analysis techniques exist for the examination of individual trajectories, these techniques are designed to shed light on the cause and effect of events specific to one protein. Determining the often subtle similarities and differences between hundreds of simulations has never before been possible, and new analysis techniques that focus on hypothesis generation rather than mere description are necessary.

Wavelet analysis is a signal processing technique that has been around since the early 1900s (Haar, 1910), but it has only recently begun to gain popu-

larity in molecular biology (reviewed by Liò (2003)). Wavelets have been specifically suggested as powerful tools in MD Askar et al. (1996), but until now have remained unexplored in this field. Like the Fourier transform, wavelets give information about the frequency domain of a signal, but, unlike the Fourier transform, which gives only average information about each frequency, wavelets give instantaneous information about how a particular frequency is localized in time. Consequently, one can obtain considerable information about the modes of a particular signal without losing information about when these modes occur or how variable they are (Fig. 3.1). The continuous wavelet transform (CWT) is a wavelet technique that offers high resolution information about a signal at any scale. For our purposes, a signal is the trajectory of an atom over time. The CWT is defined as

$$W^{(\psi,s)}(t) = \frac{1}{\sqrt{s}} \int q(\tau)\psi^* \left(\frac{\tau - t}{s}\right) d\tau, \tag{3.1}$$

where s is the scale of the wavelet, t is time, $q(\tau)$ is the signal over time, $\psi(t)$ is the wavelet function, τ is the variable of integration, and $*$ denotes the complex conjugate. Conceptually, this is equivalent to sliding a given wavelet function along the signal and calculating the match of the signal to the wavelet at each time. The wavelet is scaled (or horizontally stretched) by some amount determined by the scale s in order to examine various wavelengths in the signal. In order for wavelets to produce finite values localized in time, they are required to be localized in time and frequency space, meaning they and their Fourier transforms must approach zero as time or frequency approaches negative or positive infinity. We additionally require that they have unit power ($\int_{-\infty}^{\infty} |\hat{\psi}(\omega)|^2 d\omega = 1$ where $\hat{\psi(\omega)}$ is the Fourier transform of $\psi(t)$) in order to make them comparable across scales. Wavelets are also required to have a mean of zero. Examples of wavelet functions are shown in Figure 3.2.

For a discrete signal q of length n, the wavelet coefficients $\mathbf{W}^{(\psi,s)}$ for a scale s and a wavelet function ψ are calculated using Equation 3.2, a discrete version of Equation 3.1:

$$W^{(\psi,s)}(t) = \frac{1}{\sqrt{s}} \sum_{j=0}^{n-1} q_j \psi^* \left(\frac{j-k}{s} \right). \tag{3.2}$$

The resulting coefficients can then be examined in terms of time and scale (or wavelength) as shown in Figure 3.1c. The coefficients can be calculated very efficiently using the discreet Fourier transform and convolution theorem (Arfken, 1985). Further details including complete Mathematica codes for calculating wavelets are included in Appendix C. Because each wavelet function has a unique shape, the scale of a wavelet does not always correspond perfectly to the wavelength at which it best matches the signal. For example, the Paul wavelet (Fig. 3.2b), when scaled by s, matches a sine or cosine wave with a wavelength of approximately $1.389s$. The Morlet wavelet (Fig. 3.2b), on the other hand, would match a wavelength of $1.01s$. These parameters can be calculated using the method outlined by Meyers et al. (1993). Parameters as well as equations for each of the wavelets used in this paper are given in Table 3.1.

Once wavelet coefficients have been calculated, one may determine which scales and times are significant and which are not. To demonstrate how this can be done, suppose that we believe our signal follows white noise, meaning that at every wavelength, the signal will tend to have the same amplitude of motion. We would thus expect that at any given time t the square of the absolute value of the wavelet coefficient for a wavelength λ would be approximated by the variance of the original signal; note that the absolute value is used because the wavelet coefficient may be a complex number. Generally speaking, we can expect that a wavelet coefficient will be normally distributed around the expected value, thus the square of its absolute value, assuming the coefficients are complex numbers, will be distributed by $\chi_2^2\sigma^2/2$. By extension, if

we believe that the mean amplitude of our signal is distributed by the function $\nu(\lambda)$ and that the wavelet coefficients will be normally distributed around their mean amplitudes, then we expect the square of the absolute values of our wavelet coefficients to be distributed by $\chi_2^2\sigma^2\nu(\lambda)/2$. Using this distribution, we can choose any significance level and examine only those regions of time whose power is in the upper portion of the expected distribution, just like in a standard t-test. For a more complete theoretical description of the continuous wavelet transform, please refer to Daubechies (1992). A practical guide to wavelets is discussed by Torrence and Compo (1998). Implementation details, including an exact algorithm, are given in Appendix C.

We begin by showing what wavelet analysis provides for a simple 3-helix bundle fold (the engrailed homeodomain, EnHD). Then we analyzed simulations from all 807 of the targets in our Dynameomics database; we demonstrate the utility of wavelet analysis by focusing on two proteins: endonuclease A (*1cem*) and profilin (*1ypr*). We compare these wavelet spectra to other analysis methods as well as to the trajectories themselves. With these two proteins, we show that wavelet analysis can be used to discover several kinds of interesting events in a simulation. We then show how wavelet signatures can serve as an excellent high-throughput metric for identifying subtle features and interactions in a trajectory that are not always obvious using traditional techniques. As an example, we show how wavelet spectra locate an event in the simulation of $\gamma\delta$ resolvase that explains how the protein achieves the flexibility required to bind DNA.

3.3 Methods

3.3.1 Molecular Dynamics Simulations

Simulations were performed with explicit water using our in-house developed simulation package in lucem molecular mechanics (Beck et al., 2008; Beck and Daggett, 2004) and our previously described protein and water force fields (Levitt et al., 1995, 1997). Simulation details can be found elsewhere (Beck et al., 2008). Here we are focusing on the 298 K trajectories. For each simulation, atomic coordinates from all but the first 1 ns of our trajectories were analyzed from our in-house developed database (Simms et al., 2008). For each ps of the simulation, the protein structure was aligned to the initial structure using a rigid least squares fitting of Cα atoms with the structure's center of mass held at the origin (Kearsley, 1989). Haar, Morlet, and Paul wavelet analyses were performed on each Cα atom's trajectory over time; these wavelet data were then loaded into Mathematica (Wolfram Research, 2008) for further analysis. The total number of proteins/simulations analyzed was 807 ($\sim$17 μs total), which represents all 'simulatable' (self-contained folds) in our new 2009 consensus domain dictionary (Schaeffer et al., 2010), which is an updated version of our 2003 domain dictionary (Day and Daggett, 2003).

3.3.2 Wavelet Analysis

We chose to use the continuous wavelet transform because of its ability to retain very finely detailed information at a wide range of wavelengths. Scales were chosen to fit Equation 3.3,

$$s_k = 105 \cdot 2^{k/8}, k = 0, 1, ...59, \tag{3.3}$$

giving a range of 60 scales from 105 ps to 17.5 ns. The granularity for our simulations is 1 ps, so this range of scales captures both the fast (100 ps) and the relatively slow (10-20 ns) motions that occur in our simulations. Additionally, the large number of wavelet scales gives a very fine resolution.

Three wavelet functions were chosen in order to capture the variety of motion that can occur in a simulation. The Morlet wavelet (Goupillaud et al., 1984) consists of a plane wave tempered by a Gaussian. The Morlet has both a real and imaginary component, such that it can capture both the amplitude of the motion and the phase. It best matches motions that are sinusoidal in nature. The Haar wavelet (Haar, 1910) is a very simple wavelet that is zero everywhere except for immediately before and after 0 where it is 1 and -1, respectively. The Haar wavelet best matches sudden changes in a signal and square waves. The Paul wavelet (Addison et al., 2002) is essentially a complex version of the famous Mexican hat wavelet, which is based on the derivative of the Gaussian function. It is similar to the Morlet wavelet but decays more quickly, giving it better resolution in time and lower resolution in frequency. Notably, the imaginary portion of the Paul wavelet can match sigmoidal signals quite well. All wavelets were initially scaled so as to have a single period of approximately 21 ns. Plots of the three wavelets are shown in Figure 3.2. Example wavelet spectra for the Cα atom of Arg29 of EnHD are show in Figure 3.3. These spectra demonstrate that the Morlet, Paul, and Haar wavelets have different sensitivities in time and frequency but maintain the same general trends.

Because the amount of data generated by a single wavelet analysis is so immense (60 times as much data as the simple x, y, z coordinates), an efficient method of compression had to be employed. (For perspective, the 807 proteins hold $\sim$161.5 billion Cartesian coordinates for analysis, not including solvent.) Based on the observation that wavelet spectra tend to be smooth, we chose to

save each spectrum by approximating it with cubic splines. For each scale, s, four splines were fitted using general least-squares for every period of the wavelet function. For example, a scale of 10.5 ns would be estimated using 8 splines, uniformly distributed. To assure that this technique did not lose excessive amounts of data, we calculated the total square deviation of each scale for every atom. In all cases, the mean square deviation was less than 1/100th of the variance of the original values.

In order to determine which pieces of a wavelet spectrum are of interest, we used the basic significance testing method outlined by Torrence and Compo (1998). Because the square of the absolute value of a wavelet coordinate is distributed by $\chi_2^2 \mu\lambda\sigma^2/2$, where the variance of the signal is σ^2 and the mean expected Fourier power (amplitude) of a particular wavelength λ is μ_λ, we only need to know the mean Fourier power for a wavelength to determine statistical significance of the oscillations occurring at any given time for that wavelength. We calculated the Fourier spectrum, f_λ, for each of our wavelengths over every atom's trajectory, q, according to Equation 3.4 and found that the mean Fourier power, $|f_\lambda|^2$, was approximately described by the equation $\mu_\lambda = \lambda^{1.43}/155 + 20$, where λ is the wavelength measured in picoseconds. Equation 3.4 is similar to the calculation of a single Fourier coefficient but at an arbitrary wavelength. The calculation is made over only part of the signal in order to prevent incomplete sinusoidal waves from biasing the magnitude of the calculation.

$$f_\lambda = \frac{1}{\lambda\lfloor N/\lambda \rfloor} \sum_{k=n-\lambda} N - 1 e^{\frac{-2\pi i k}{\lambda N}} q_k, \tag{3.4}$$

For each wavelet spectrum, we extracted regions whose values were statistically in the upper 20% of the expected power distribution as strong oscillations of a particular wavelength. For each scale, s, regions within $s/2$ ps of the beginning or end of the trajectory were ignored in order to avoid the edge effects

inherent with a finite signal. Additionally, the first nanosecond was ignored to allow for equilibration. For each picosecond, the wavelength at which a given Cα atom was oscillating according to this analysis was recorded. Whenever multiple frequencies occurred at the same time, the one with the stronger oscillation (the greater outlier in its distribution) was used. These data thus formed an "oscillation map" of the wavelengths that were most prevalent at every picosecond for each Cα atom in a given protein.

In order to demonstrate the utility of these oscillation maps, we examined their general properties for all 807 proteins. We hypothesized that an atom experiencing no significant wavelet oscillations over a time regime would be characterized by very little motion or by rapid vibrations, likely due to heat. Similarly, we hypothesized that those residues with low frequency wavelets would be characterized by structural rearrangements and large motions during the time of those wavelets. To test this, we randomly chose 100 residues and time regions from our 807 proteins requiring only that the wavelets for the residue be of a uniform frequency over that time. Time regions were allowed to be low frequency ($\lambda > 1$ ns), high frequency ($\lambda < 1$ ns), or no frequency (no significant wavelets) for the entire region in question. These residues were then scored as either arbitrary vibrations or large movements/rearrangements with the actual values of the wavelets during each time region concealed. The results were then tallied and compared. To demonstrate our specific findings, we present wavelets for two proteins: profilin (ProF; *1ypr*) and endonuclease A (CelA; *1cem*). Finally, to show how wavelets can be used to mine simulations, we compared the low frequency distributions of all Cα atoms and examined the simulations of those with the highest low-frequency wavelet content. The trajectory of one such pair of atoms, G101 and M103 of $\gamma\delta$ resolvase, revealed a novel mechanism in which helix αE changes conformation during DNA binding.

3.4 Results and Discussion

In general, the Morlet and Paul wavelets were a better fit for MD trajectories than the Haar wavelet. At a given wavelength, the Paul wavelet tended to give the best resolution in time, the Morlet wavelet tended to give the best resolution in frequency, and the Haar tended to lag behind both. This comparison is demonstrated in Figure 3.3 for the simple 3-helix bundle fold of EnHD. There were no residues in all of our simulations that could be differentiated from white noise more than 20% of the time using the Haar wavelet; thus, we do not consider it further (note that nothing in Fig. 3.3c is statistically distinct from white noise).

In the 807 protein data set, high frequency oscillations ($\lambda < 1$ ns) were common, occurring 22% of the time, but they were frequently correlated with thermal vibrations. Midrange and low frequencies occurred 30% of the time and were almost always correlated with motions ranging from slight rearrangements to loss or gain of secondary structure to broad shifts in backbone conformation. When scored by hand, regions of time with no significant wavelets correlate with arbitrary vibrations 78% of the time while low frequency wavelets correlate with structural movements and rearrangements 73% of the time. High frequency wavelets correlated with movements and rearrangements 50% of the time and with arbitrary vibrations 50% of the time.

Proteins with very stable trajectories have considerably fewer significant oscillations than those that were unstable. EnHD, for example, exhibits only a small amount of motion at the N-terminal tail. Only 20% of the time is there a significant oscillation with $\lambda > 1$ ns not occurring in the N-terminal tail (Fig. 3.4a). Conversely, proteins that undergo considerable rearrangement from their crystal structures have more low frequency oscillations. The DNA-binding domain of ADR6 (*1kkx*) is a protein with a similar topology to the en-

grailed homeodomain, but which was deemed unstable by our simulation. It undergoes a large set of helical rearrangements in the beginning of its trajectory after which it moves less but has an exposed hydrophobic core. Low frequency oscillations occur in 35% of this simulation, most of which correlate with the protein's overall shifts (Fig. 3.4b).

Given that low frequency wavelets correlated strongly with overall rearrangements in a protein simulation, we searched all 807 simulations for wavelet coordinates that whose wavelength was at least 1 ns and whose significance was in the top 5% of the expected power distribution. Two proteins stood out: endoglucanase A (CelA) and profilin (ProF). We examine these proteins in more detail here.

The catalytic core of CelA is an all-helical protein in the α/α toroids family (Fig. 3.5a). The simulation of CelA contains moderate rearrangement of several mobile loops early on and several subtle changes that occur throughout the simulation. The Paul oscillation map and the root mean square fluctuation (RMSF) plot for CelA are shown in Figure 3.6a. RMSF is a commonly used metric for the amount of fluctuation occurring in a residue over time relative to its average position. Three main regions are of interest in this wavelet map, the first of which is an empty region around 5-10 ns near residue 125 followed by the low frequencies around 14 ns. The corresponding structures for these regions are shown in Figure 3.5b. Another interesting region is the low frequency block near residue 250 throughout the middle of the simulation. The structures for this region are compared with the region absent of low frequencies at the end of the simulation in Figure 3.5c. Finally, Figure 3.5d shows the subtle helical shift that occurs near residue 350 early in the simulation. These fluctuations are not visible on the RMSF spectrum due to their subtle nature and their relatively small movements. RMSF and other traditional analyses often fail to detect small movements, even when they are significant, due to

their focus on the amount of change rather than the quality of change. Wavelet analysis finds these motions despite their subtlety because they are ordered rearrangements.

The protein profilin is a member of the profilin-like family (Fig. 3.7a) that binds actin and regulates the growth of actin filaments. The simulation of ProF, in contrast to CelA, undergoes a few fast rearrangements in the first few ns of the simulation after which little significant motion is observed. The simulation is very stable with even the most flexible residue having a mean RMSF of only $\sim$0.76 Å. When examining the Morlet oscillation map of ProF (Fig. 3.6b), one is immediately drawn to the low frequency block throughout the middle of the simulation between residues 55 and 60. This midrange oscillation occurs for a long period of time and is focused around a band of of residues from A53-N58 (Fig. 3.7). These residues are in a helix near the binding interface with actin, and S57 participates directly in actin binding. Above this band (further along the sequence) are several other shorter-lived bands of low-frequency motion containing 6 other actin-binding residues (M68, L70, R71, H81, D82, and G85). In the crystal structure, S57 points outward into solvent and away from the other binding residues, but during time frame highlighted by the low frequency wavelets from $\sim$4.5 ns until $\sim$14 ns, the helix containing S57 unravels from the C-terminal end, keeping the loop containing S57 and N58 in tact and pushing them toward the other active site residues slightly (Fig. 3.7).

Figure 5 shows the RMSF for CelA and ProF over time. For these proteins, their RMSF profiles are essentially uncorrelated with their wavelet maps. Notably, there is a slight increase in the RMSF of the region S122-A153 for CelA during the longer wavelengths near 15 ns. However, regions E245-Y275 and S335-T360 show virtually no distinctive patterns in the RMSF spectra. Similarly, the regions around S57 and N58 of ProF show little correlation with the wavelets and, in fact, do not tend to change much over time. Thus,

wavelet analysis was able to effectively screen for and detect interesting motion within two unrelated proteins where conventional analysis failed. Searching a database of multiple simulations of 807 proteins and > 100 μs of simulation time for interesting events is a daunting task. In order to expedite this process, we hypothesized that individual residues dominated by low frequency movements were most likely to be involved in significant conformational events. Accordingly, we examined the trajectories of Cα atoms in our simulations that had the highest portion of significant low frequency (> 1 ns) motion according to the Paul and Morlet wavelets. Two such atoms, both in the upper 5% of the distribution, belong to G101 and M103 of $\gamma\delta$ resolvase (*1gdt*). $\gamma\delta$ resolvase is a 183-residue protein belonging to the resolvase and DNA invertase family that forms a homodimer in solution (Yang and Steitz, 1995). It is known that G101 is a critically flexible residue situated between β-strand 5 and α-helix E (Fig. 3.8a) that allows αE to pivot away from αD during DNA binding (Li et al., 2005), but how this event occurs is unclear.

In our simulation of the monomer of $\gamma\delta$ resolvase, we observed a slight unraveling of helix αE and β-strand 5 around 3.5 ns as well as periodically throughout the simulation (Fig. 3.8b). These movements were the cause of the low frequency motion highlighted by wavelet analysis. Closer examination revealed that this separation is accompanied by the formation of an Ω-loop between β5 and αE with G101 at its tip. This loop is stabilized by the movement of the side-chain of M103 from a solvent-accessible state into a hydrophobic pocket consisting of I90, F92, and I97 where it displaces the Cγ of T99 (Fig. 3.8c and 3.8d). During this motion, T99 rotates out of the pocket, maintaining its hydrogen bond with the amide of I90 and allowing it to easily rotate back into the pocket when M103 leaves. The result of this event is a slight turning of αE and a loosening of loop 5E, making further rearrangement of αE, such as that required for strand exchange, possible. Interestingly, methionine can be

reversibly oxidized, increasing its polarity and hydrophilicity, a process proposed to be involved in protein regulation (Stadtman et al., 2003; Santarelli et al., 2006). Theoretically, an oxidized M103 or a mutation such as M103D could stabilize the solvent-accessible state (αE closed) while a reduced M103 or a mutation such as M103L could stabilize the Ω-loop (αE open). Thus, an automated screen for Cα atoms in the upper 5% of the distribution with respect to low frequency motion led to the discovery of interesting cyclic conformational behavior that may be linked to function.

The wavelet analyses explored here are a very effective method of examining both very large and very subtle types of motions occurring in a protein over time. We have demonstrated that wavelets are capable of picking out multiple types of distinct movements that occur within a protein that may not be easy to find via visual inspection of the trajectory or by using traditional analysis methods (for example, CelA, ProF). Additionally, wavelets are capable of pinpointing when a change is occurring in time, allowing them to be used as a high-throughput screening technique for simulations (as with $\gamma\delta$ resolvase). It is not surprising that the Haar wavelet fit our data poorly. The Haar is, by nature, designed for square waves and discrete jumps, neither of which we observe in our simulations. The Paul wavelet, which approximates the Haar wavelet in a smooth form, was much more useful for our purposes. Both the Paul and the Morlet wavelet provided good results, though the Paul is theoretically better suited for analysis across time due to its high temporal resolution. Although it is initially surprising that wavelets would be able to detect non-oscillatory movements, such as a helical rearrangement, it should be noted that a sigmoidal trajectory by an atom can easily match the imaginary part of an appropriately scaled Paul wavelet (Fig. 3.2b). Thus, the Paul wavelet should not be thought of purely as an indicator of oscillation, but rather as an indicator of non-random motions. The fact that wavelet significance testing is not depen-

dent on the amplitude of the oscillation additionally confers an advantage, in that large motions do not necessarily drown out smaller motions as is often the case in analyses such as RMSF. For example, a large hinge motion between two regions of a protein would not prevent a smaller change in secondary structure within one region from being detected.

Wavelets show clear sensitivity and specificity to all ranges of structural rearrangement in a simulation, including many that are not visible using traditional analyses such as RMSF. This is potentially of great use for studying the effects of mutation, pH, and/or temperature on a structure, as these changes can be difficult to detect. The motions highlighted from CelA (Fig. 3.5) demonstrate the range of wavelet sensitivity, as these motions include a large loop rearrangement (Fig. 3.5b), a small change in contacts and secondary structure position (Fig. 3.5c), and a subtle change in the arrangement of two helices (Fig. 3.5d).

Wavelets also show promise for detecting biochemically relevant motions that can be otherwise very subtle and difficult to find. Notably, the Cα RMSFs for the oscillating region in ProF are relatively low and show no particular distinction over the time range during which the helical unwindings were occurring (Fig. 3.7b). In fact, compared with the oscillation maps, the RMSF profile shows very little differentiation over time.

Notably, the Paul and Morlet wavelets excel at detecting different kinds of events. While the Paul wavelet showed excellent sensitivity to changes and rearrangements in protein structure, the Morlet showed sensitivity to periodic oscillations. This sensitivity suggests that the Morlet wavelet may be useful in detecting interactions and communication in long simulations while the Paul wavelet may additionally be useful in examining changes in simulations and simulations in which rearrangements are expected to occur, such as in high-temperature unfolding simulations.

Perhaps most critically, all of these advantages of wavelets can be used in a high-throughput fashion to screen and isolate events in large simulations or sets of simulations, as illustrated with $\gamma\delta$ resolvase. Finding an event of interest by hand in even 0.1 μs of simulation data of a single protein is a daunting task and is virtually impossible for our now complete database containing $\sim$11,000 simulations of all protein folds. As high-throughput computation becomes more common, methods for mining the resulting data, such as wavelets, are becoming more important.

3.5 Conclusions

Wavelet analysis is a powerful tool that can be used to quickly and automatically isolate distinct motions of interest in a protein simulation. Due to their ability to locate subtle changes without being drowned out by larger more obvious motions, wavelets represent an ideal method for screening simulations to quickly pinpoint changes or structural rearrangements and for comparing biochemical differences in simulations, due to mutation, pH, or temperature changes, for example. Additionally, wavelets can be used to scan large databases of simulations for biochemically relevant events, such as the motion of a catalytic site and other portions of the structure that may interface with it.

Table 3.1: Formulas and wavelengths for wavelets used in this chapter.

Wavelet	Formula	Wavelength of $\mathbf{W}^{(\psi,s)}$ (ps)
Morlet (frequency = 2π)	$\psi(t) = \pi^{-1/4} e^{-t^2/2} e^{2\pi i t}$	$1.01s$
Paul (order = 4)	$\psi(t) = 8\sqrt{\frac{2}{35\pi}}(1 - it)^{-5}$	$1.389s$
Haar	$\psi(t) = \begin{cases} 1 & -1/2 \leq t < 0, \\ -1 & 0 \leq t < 1/2, \\ 0 & \textbf{otherwise.} \end{cases}$	$0.87s$

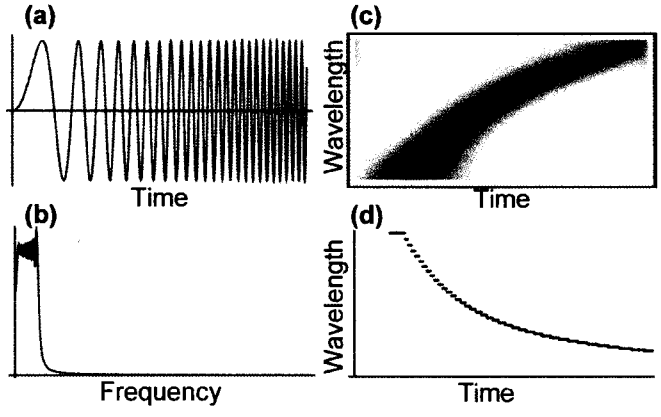


Figure 3.1: Comparison of Fourier transform and the continuous wavelet transform. **(a)** A signal whose frequency increases over time. **(b)** The absolute value of the Fourier transform of the signal in a. **(c)** The continuous wavelet transform of the signal in a. Notably, the wavelet transform shows clearly that the signal is increasing in frequency over time while the Fourier transform shows only that low frequencies are dominant. **(d)** Plot of the significant wavelength over time of the signal in **(a)**, calculated by taking the most significant wavelet wavelength from **(c)** at each time with a minimum significance of 0.2.

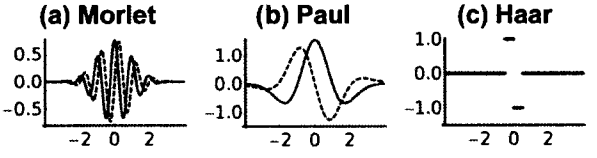


Figure 3.2: Plots of the three wavelets used in this study, as described in Table 3.1, each plotted from -4 to 4 with scale $s = 1$. Solid lines represent the real parts while dashed lines represent the imaginary parts. **(a)** The Morlet wavelet. **(b)** The Paul wavelet. **(c)** The Haar wavelet.

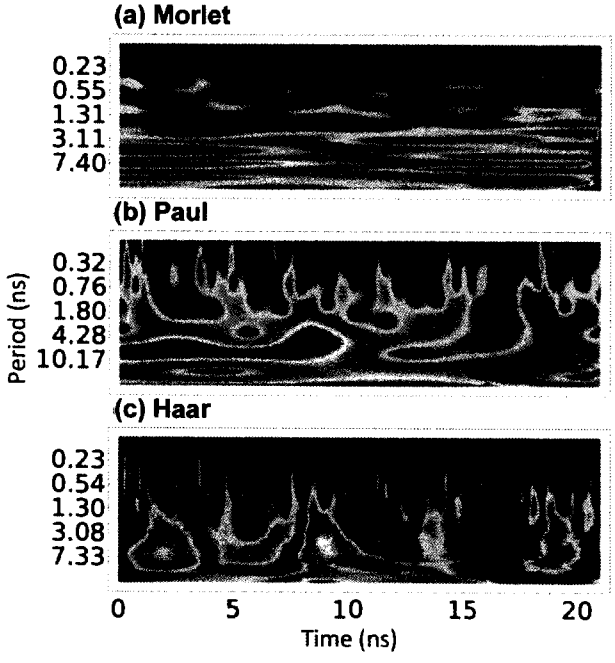


Figure 3.3: Plots of the wavelet analyses of the Cα atom of R29 of the engrailed homeodomain (*1enh*). The absolute value of each wavelet coordinate is shown with low values illustrated in blue. No scale is given because wavelet values are in arbitrary units. **(a)** The Morlet wavelet. **(b)** The Paul wavelet. **(c)** The Haar wavelet. The scales of each are not identical as they are not directly comparable.

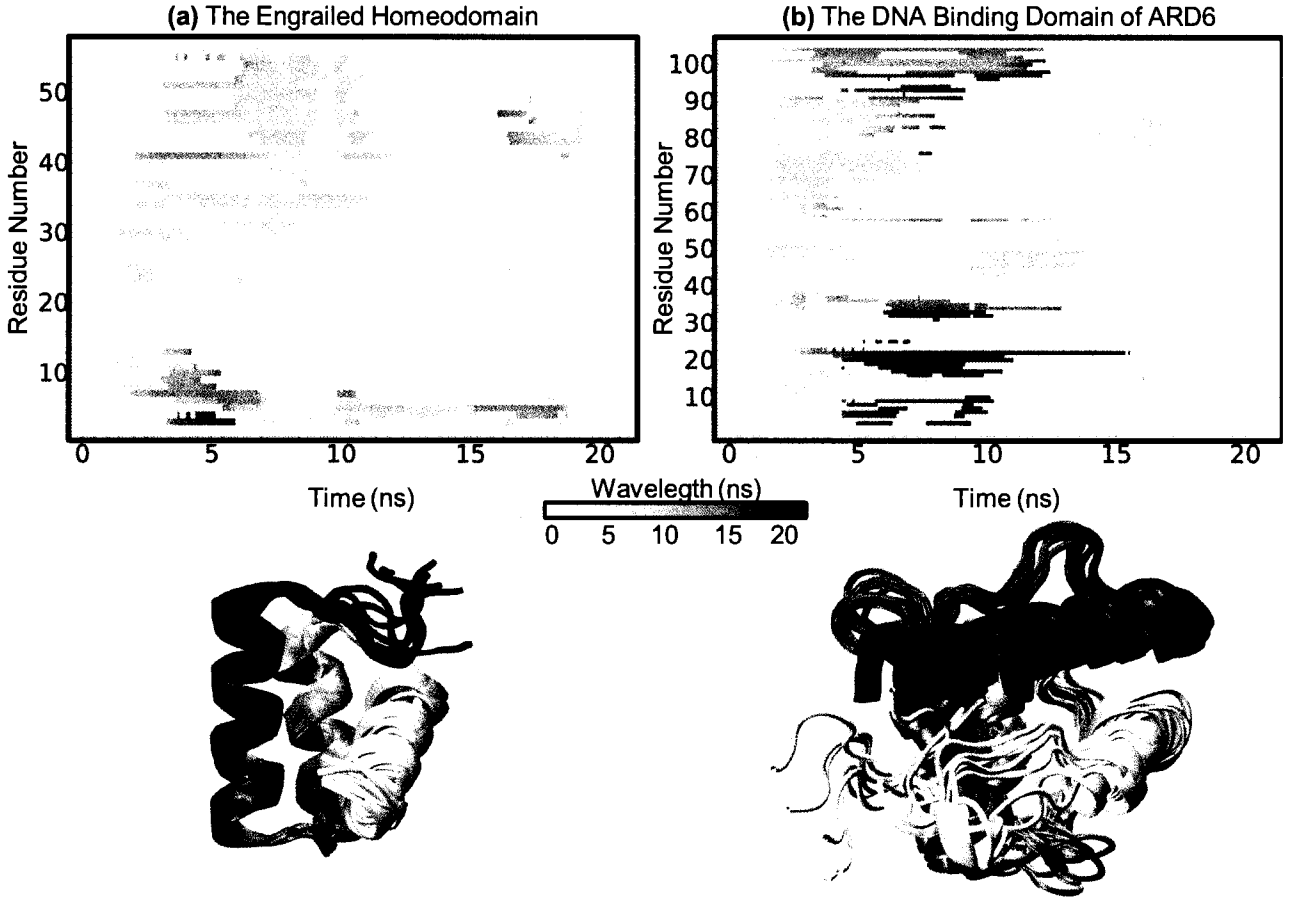


Figure 3.4: Wavelet maps and structures from the simulations of **(a)** the engrailed homeodomain (*1enh*) and **(b)** the DNA-binding domain of ADR6 (*1kkx*). The engrailed homeodomain has few significant motions outside of its tails. This is shown clearly in the wavelet map, which has few low frequency oscillations. The simulation of ADR6 was unstable, and its rearrangements are highlighted by the many low frequency wavelet coordinates.

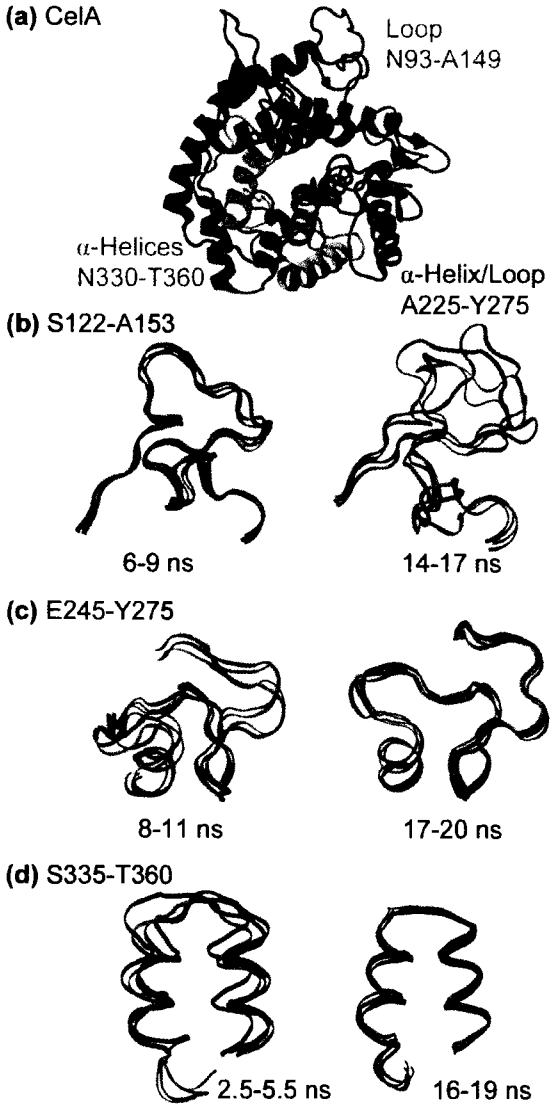
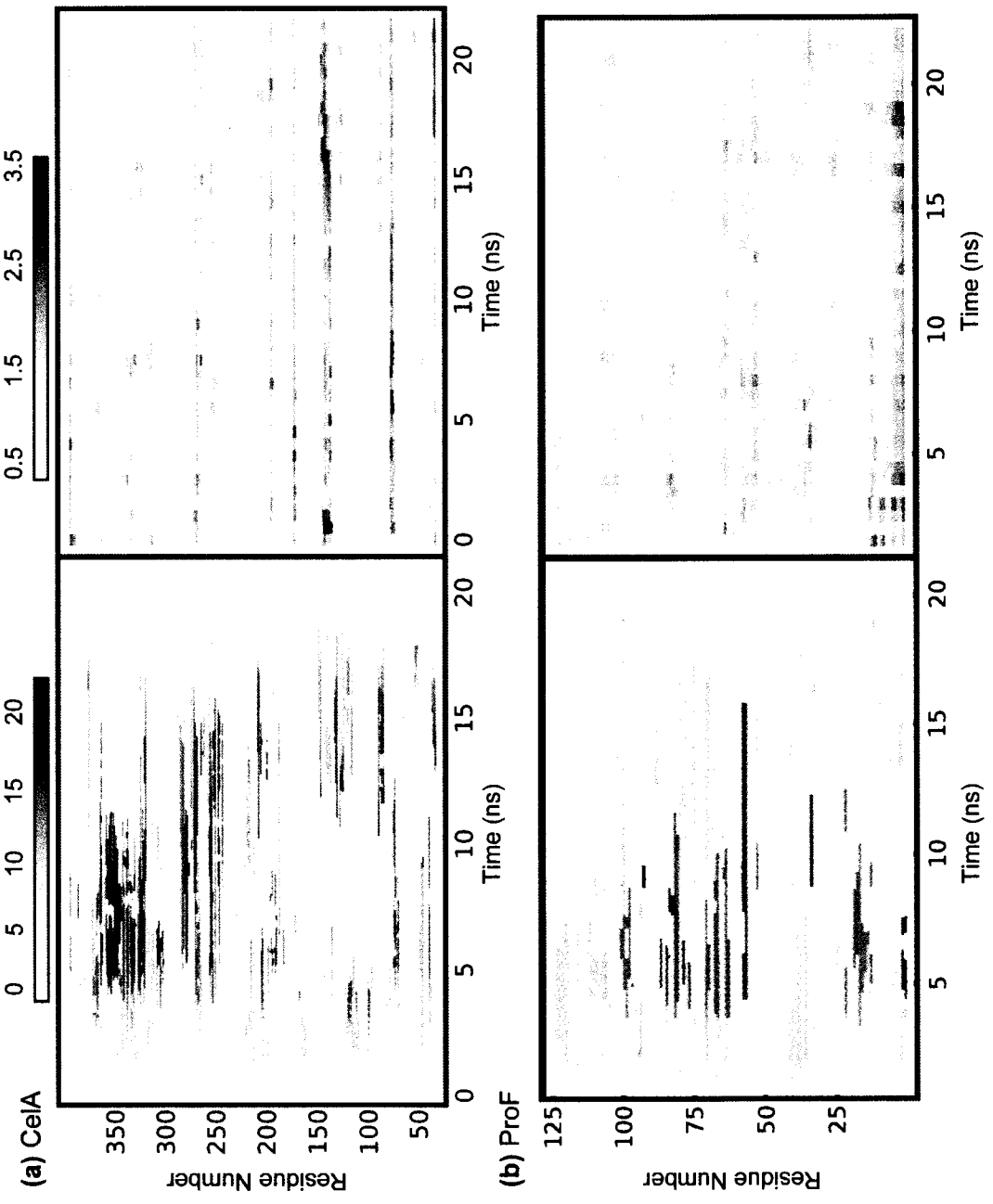


Figure 3.5: **(a)** Protein structures and notable structural features of the protein Endonuclease A (*1cem*; CelA) taken at 10 ns in its simulation. **(b)** Region S122-A153 of CelA colored red, green, blue, magenta in temporal order. **(c)** Region E245-Y275. **(d)** Region S335-T360. In each instance, the time period whose wavelet coordinates were significant in the low frequency range are mobile while the time period whose wavelet coordinates were not significant in the low frequency range is stationary.

Figure 3.6: Wavelet maps and RMSF plots of **(a)** Endoglucanase A (*1cem*; CelA) and **(b)** profilin (*1ypr*; ProF). The wavelet maps show the most statistically significant wavelength of each Cα atom occurring at each time. Notably, RMSF maps and wavelet maps are not correlated in time.


(a) CelA
Wavelenth (ns)
0 5 10 15 20
RMSF (Å)
0.5 1.5 2.5 3.5
Residue Number
350 300 250 200 150 100 50
0 5 10 15 20
Time (ns)
0 5 10 15 20
Time (ns)
(b) ProF
Residue Number
125 100 75 50 25
5 10 15 20
Time (ns)
0 5 10 15 20
Time (ns)

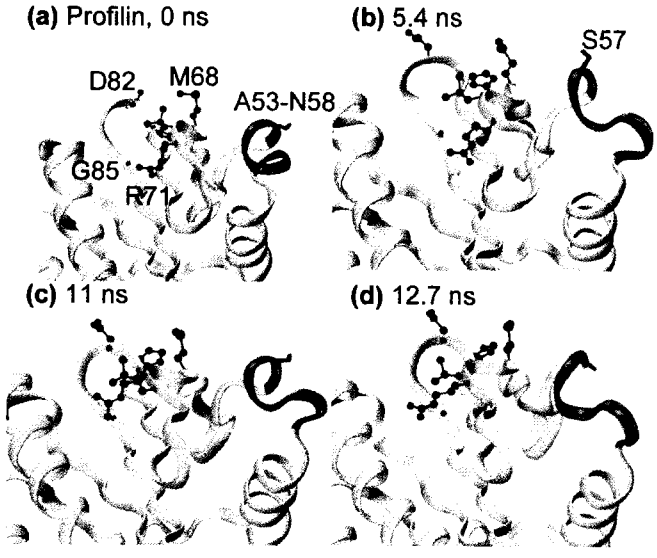


Figure 3.7: Changes in profiling (ProF) binding residue S57. Helix $\alpha 3$, containing S57, is shown in red. Side-chains of actin binding residues highlighted by wavelet analysis are shown in black, and the side-chain of S57 is shown in red. **(a)** Minimized crystal structure. **(b)** 5.4 ns, **(c)** 11 ns, and **(d)** 12.7 ns. During this time period, the $\alpha 3$ twists significantly and unravels from the N-terminal end, changing the orientation of S57 to the binding site.

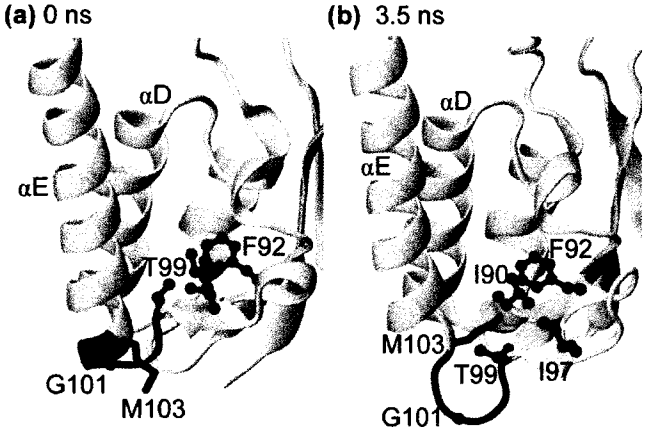


Figure 3.8: The protein $\gamma\delta$ resolvase (*1gdt*). The side-chains of residues forming a hydrophobic pocket (I90, F92, I97, T99, and M103) are shown in black while the backbones of residues 99-103 are shown in red. **(a)** Residues near loop 5E in the minimized crystal structure. **(b)** Residues near loop 5E at 3.5 ns. Near 3.5 ns the end of helix αE and part of loop 5E unwind to form an Ω-loop. This motion flips the side-chain of residue M103 into the hydrophobic pocket shown in black while pushing residue G101 into solvent, stabilizing the alternate conformation. Both M103 and G101 are known to be important for the binding and flexibility of αE and were identified as highly significant during this time range by wavelet analysis.

Chapter 4

UNDERSTANDING THE MOLECULAR BASIS OF DISEASE IN SINGLE NUCLEOTIDE POLYMORPHISM VARIANTS USING WAVELET ANALYSIS

4.1 Summary

Single nucleotide polymorphisms (SNPs) are frequently associated with disease yet their effects and modes are often very difficult to characterize. Although molecular dynamics (MD) techniques are capable of determining precise structural and dynamic differences between SNP variants, they are often difficult and time-consuming to interpret and analyze. Wavelet analysis has shown promise in quickly screening MD simulations for events of interest, and could potentially make the comparison and analysis of simulations much easier, faster, and more quantitative. Here, we use wavelet analysis to analyze MD simulations of two SNP variants for each of four methyltransferase proteins. The variant proteins each displays subtle structural changes and took over a year to fully characterize using traditional analyses. We compare the results of wavelet analysis to those of traditional analyses and demonstrate that wavelet analysis can quickly and quantitatively determine differences between two highly similar simulations without requiring the large time investment that is traditionally needed.

4.2 Introduction

Single nucleotide polymorphisms (SNPs) are ubiquitous mutations that frequently lead to or are associated with disease, altered protein function, or

changes to a protein's structure and dynamics. In many cases, the phenotypic effects of a SNP are known while the underlying mechanism of the mutation is not. In these cases it is instructive to use molecular dynamics (MD) simulations to study the affects of the SNP on protein structure and dynamics in order to inform further experimental research. Critical to this process is the need for MD simulations to be quickly analyzed and easily interpreted.

Methyltransferases comprise a large family of proteins involved in many aspects of molecular biology including protein/DNA repair (Bugni et al., 2007; Pahlich et al., 2006; Weinshilboum, 2006), transcriptional regulation (Saito et al., 2001), hormone signaling, neurotransmission, and drug metabolism. Multiple SNPs have been identified in several methyltransferases (Saito et al., 2001) including protein *L*-isoaspartate *O*-methyltransferase (PIMT), histamine *N*-methyltransferase (HNMT), thiopurine *S*-methyltransferase (TPMT), and catechol *O*-methyltransferase (COMT). The variant alleles of TPMT (A80P), HNMT (T105I), and COMT (V108M) are linked with increased risk for hematopoetic toxicity (Coulthard and Hogarth, 2005; Deeken et al., 2007); age-related disease (Ligneau et al., 1998; Morisset et al., 2000; Panula et al., 1997) and alcoholism (Oroszi et al., 2005; Reuter et al., 2007); and breast cancer (Dawling et al., 2001; Goodman et al., 2002; Huang et al., 1999; Lavigne et al., 1997; Matsui et al., 2000; Mitrunen et al., 2002; Sazci et al., 2004; Tan et al., 2003; Thompson et al., 1998; Wedren et al., 2003; Yim et al., 2001) and neuropsychiatric disorders (Bilder et al., 2004), respectively, while the PIMT V119I heterozygosity is linked to successful aging (DeVry and Clarke, 1999). These proteins and the effects of their mutations are summarized in Table 4.1. All of these polymorphisms decrease enzymatic activity and protein stability. Interestingly, all of the variant residues are located on the protein surface ~16-20 Å from the site at which the enzyme's co-substrate S-adenosylmethionine (SAM) binds (Fig. 4.1). Crystal structures of both of the polymorphic variants

of COMT (Rutherford and Daggett, 2008) and HNMT (Horton et al., 2005, 2001) bound with SAM and a substrate analogue have been solved. The active sites and overall structures of both variants of each protein are virtually identical, indicating that the overall changes in the proteins due to the mutation are extremely subtle.

In order to determine how these mutations affect methyltransferase structure and dynamics, we performed multiple MD simulations of wild-type and variant of PIMT (Rutherford and Daggett, 2009b), TPMT (Rutherford and Daggett, 2008), HNMT (Rutherford et al., 2008b), and COMT (Rutherford et al., 2006; Rutherford and Daggett, 2009a) apoproteins. Extensive analyses, including structural visualizations, solvent-accessible surface area (SASA), root-mean-square fluctuations (RMSF) and deviations (RMSD), and changes in contacts initially revealed that the introduction of the larger polymorphic residues (PIMT V119I, HNMT T105I, TPMT A80P, COMT V108M) alters local side-chain packing, distorting the orientation of active-site residues and increasing the solvent exposure of the active site $\sim$20 Å away from the site of mutation. Notably, the effects of the V119I mutation on PIMT structure and dynamics were very subtle, making it difficult to characterize using conventional analysis techniques at all, while the the A80P mutation in TPMT caused the loss of an entire α-helix. Exact descriptions of the subtle structural effects of these mutations required several years of analysis. In order for MD to inform experimental research, novel analyses capable of quickly and confidently identifying subtle changes such as those that occur in these methyltransferase and other SNP-associated proteins are necessary.

Wavelet analysis is a technique that is gaining popularity in molecular biology (reviewed by Liò et al. (Liò, 2003)). This method allows one to search a particular signal (for us, the trajectory of an atom over time) for a set of motions at various scales. It is similar to the Fourier transform, which allows one

to examine the frequencies (different scales of sinusoidal motion) of a signal. However, unlike the Fourier transform, wavelet analysis retains information about when a motion is occurring (Fig. 4.2). This makes wavelet techniques ideal for studying signals in which certain frequencies of motion are dominant only at a specific time but not necessarily throughout. Wavelet analysis is performed by sliding a wavelet function $f(t)$ along a signal $q(t)$. The wavelet function is scaled to a variety of sizes, and the resulting wavelet coefficients, $W_{s,t}$, represent the likeness of the signal near time t to the motion of interest when the motion has been scaled by s. Detailed methods for the wavelet analysis of protein trajectories are given by Benson et al. (Benson and Daggett, 2010a).

Here, we examine the MD simulations of PIMT, HNMT, TPMT, and COMT apoproteins using wavelet analysis and compare the results for each of the SNP-associated variants with those obtained through conventional analyses of the simulations. Triplicate simulations are used to determine if the changes identified by wavelet analysis are associated with the SNP with statistical confidence. We show that wavelet analysis is a powerful technique for quickly and quantitatively identifying small differences between MD systems, such as SNPs.

4.3 Methods

4.3.1 Protein Preparation

The protein preparation and MD simulations of PIMT, HNMT, TPMT, and COMT have been described in depth elsewhere (Rutherford et al., 2006; Rutherford and Daggett, 2008; Rutherford et al., 2008b; Rutherford and Daggett, 2009b). Crystal structures of PIMT (*1i1n* (Skinner et al., 2000), residues 2-225), TPMT (*2bzg* (Wu et al., 2007), residues 17-245), HNMT (*2aot* (Horton et al., 2005), residues 5-292), and a homology model of human COMT

based on the 2 Å crystal structure of rat COMT (*1vid* (Vidgren et al., 1994), residues 4-216), which has 81% sequence identity with human COMT, were used as starting structures. Notably, the crystal structures of human wild-type and V108M COMTs have since been solved (*3bwm* and *3bwy* (Rutherford et al., 2008a)) and are completely consistent with the homology model; in fact, the Cα RMSD between *1vid* and *3bwm* is only 0.4 Å. Although simulations have been performed with the human structures, they have not been analyzed to the extent required for comparison in this paper. The proteins simulated include wild-type (119V) and V119I PIMT; wild-type (105T) and T105I HNMT; wild-type (80A) and A80P TPMT, and wild-type (108V) and V108M COMT.

4.3.2 Molecular Dynamics Simulations

MD simulations of all of the apoproteins were performed using the *in lucem* molecular mechanics (*il*mm) simulation package (Beck et al., 2008) using protocols and a potential energy function that have been described elsewhere (Beck and Daggett, 2004; Levitt et al., 1995). The simulations include all hydrogen atoms and explicit flexible waters (Levitt et al., 1997). Proteins were solvated in a periodic rectangular box with walls extending at least 10 Å from all protein atoms. The solvent density was set to 0.993 g/ml for water at 37 °C (Kell, 1967). Once the solvent density was set, the box volume was held fixed and the NVE microcanonical ensemble (constant number of particles, volume, and energy) was employed. A 10 Å force-shifted non-bonded cutoff was used and updated every 2 steps (Beck et al., 2005), and a time step of 2 fs was used in all calculations. All simulations were a minimum of 31 ns with structures saved every 1 ps for analysis. Three independent simulations were performed for each system, for a total simulation time of 744 ns.

4.3.3 Wavelet Analysis

In order to examine the effects of a mutation on protein motion, we performed wavelet analysis on all 24 simulations, leaving out the first nanosecond to allow for thermal equilibration. We used the Paul wavelet with order 4 (Addison et al., 2002) (Fig. 4.2b), which is particularly effective in detecting subtle rearrangements in simulations (Benson and Daggett, 2010a), with 41 scales from 1 ns to 16 ns according to the formula $s_k = 2^{k/10}$ with $k \in \{0, 1, ...40\}$. Note that a maximum scale of 16 ns does not imply that only half of a 31 ns simulation was analyzed but rather that we did not examine motions that would have required more than a continuous 16 ns window to occur. The equation for the Paul wavelet is given in Equation 4.1.

$$f(t) = \frac{8\sqrt{\frac{2}{35\pi}}}{(1 - it)^5} \tag{4.1}$$

Wavelet coefficients were calculated for each Cα atom after translation and rotation had been removed from the system using a rigid least-squares method (Kearsley, 1989) as described by Benson and Daggett (Benson and Daggett, 2010a). For each atom at each picosecond, the wavelet coefficients for each scale were examined and the scale with the strongest confidence as described by Benson and Daggett (Benson and Daggett, 2010a) was identified as the dominant wavelet for that atom at that time. If no wavelet coefficient at any scale was significant at the 80% level, the motion at that time was not considered to match the Paul wavelet.

Dominant wavelets were compared across simulations by counting the number of picoseconds at which a given wavelet scale was dominant for a particular variant of PIMT, HNMT, TPMT, or COMT over all three of its simulations. For a single protein this allows us to calculate 41 p-values per atom, one for each wavelet scale, using a standard two-tailed student T-test comparing the

three simulations of wild-type to V108M COMT. These p-values can then be used to calculate the overall p-value for each atom giving the confidence that a difference in wavelet-significant motion exists for that atom between the two variants of COMT. Scales at which no motion was present in either variant were excluded from the overall p-value calculation while scales at which only one variant had motion were given a p-value of 0.1. These combined p-values were compared to results found via traditional analyses of each simulation by Rutherford et al. (Rutherford et al., 2006; Rutherford and Daggett, 2008; Rutherford et al., 2008b; Rutherford and Daggett, 2009a,b, 2010).

4.4 Results

The simulations of PIMT show many changes between the 119V and 119I proteins (Rutherford and Daggett, 2009b). However, all of these changes are extremely subtle compared to those seen in the simulations of HNMT, TPMT, and COMT. Traditional analysis techniques identified rearrangements in helices $\alpha 1$, $\alpha 3$, $\alpha 4$, and αC and in strands $\beta 3$ and $\beta 9$ (Fig. 4.3a) caused by the V119I mutation. Wavelet analysis of PIMT highlights a great number of residues, including residues in helices αA (H14, K18, N19), αB (K25, V30), $\alpha 1$ (M63, L70), and $\alpha 3$ (T128) and in strands $\beta 1$ (L82) and $\beta 5$ (I179, V182), as well as in the $\alpha 2$-$\beta 2$ (G100, C101), $\beta 5$-$\beta 6$ (P184), and $\beta 7$-αC (L215) loops.

HNMT maintains its overall structure throughout all simulations, but contains several notable differences between the wild-type and T105I proteins (Rutherford et al., 2008b). These primarily occur around the mutation in helices αA, $\alpha 1$, $\alpha 2$, $\alpha 3$, and $\alpha 4$ and in strand $\beta 3$ (Fig. 4.3b). Wavelet analysis of HNMT finds 18 residues with significant differences between variants, most of them also clustered around the mutation site. These encompass helices αA (H12, E17, N24), $\alpha 1$ (E28, D37, I44), $\alpha 2$ (E65, L68), $\alpha 3$ (Y98, K99), $\alpha 4$ (R126), and αB (S177, D180, L182) and loops $\alpha 2$-$\beta 2$ (G80) and $\alpha 6$-$\beta 6$ (L213).

The A80P mutation of TPMT disrupts the last turn of helix α2, causing distortions of the helices near the mutation site (Rutherford and Daggett, 2008). Helices α2 and α3 move apart from each other and helix α2 unravels slightly near the C-terminus where the mutation lies. Nearby helices α1 and αA are also affected (Fig. 4.3c). Wavelet analysis finds differences between the variants throughout the protein, particularly in helices αA (E27, D31, K32), α1 (H52), α2 (M76, A80P, G83), α3 (G95), α4 (T141), and αB (K228). The loops α1-α2 (A39, E43), α1-β1 (G59), βA-βB (T113, E114), and β5-β6 (H192, P195, F197).

The V108M mutation of COMT caused a number of subtle changes throughout the protein (Rutherford et al., 2006; Rutherford and Daggett, 2009a). Primarily, secondary structure elements near the polymorphic site (αB, α2, and α3) interact differently with the 108V residue than the Met. Additionally, helices α4 and α6 are more prone to distortion in the V108M variant (Fig. 4.3d). Wavelet analysis shows differences in every helix but α5 (L14; L26, D30; K46; G70; N100, R101; K128; F179, H182, R184), along with strand β1 (L63) and loops αA-αB (E18, P19), αB-α1 (A39), β1-α2 (A67), α2-β2 (P82), and at the C-terminus (P215).

4.5 Discussion

Overall, there was good agreement between the wavelet analysis described here and the traditional analysis described by Rutherford et al. However, there were a small number of disagreements. Generally speaking, wavelet analysis accurately identified residues in protein regions that were highlighted as different according to previous studies. However, wavelet analyses proved to be more sensitive than the traditional analyses, identifying some additional residues that demonstrated dynamic differences between proteins. Although this is preferable for the purpose of screening simulations for events of inter-

est, it can be adjusted by requiring a lower p-value or by requiring a higher significance level in selecting the dominant wavelets.

Among the proteins examined here, wavelets were by far the least accurate at determining the changes that occurred in the simulations of PIMT (Fig. 4.1a). Only two secondary structures were accurately identified in PIMT: $\alpha 1$ and $\alpha 3$, though, notably, the α-helix between them was correctly not identified. This lack of effectiveness was likely due to the nature of the changes in PIMT, which were by far the most subtle of all four of the proteins studies here; according to Rutherford et al. (Rutherford and Daggett, 2010), "PIMT displayed the subtlest structural effects upon substitution in the hotspot region." This suggests that there is a lower limit to the effectiveness of wavelet screening.

Wavelet analysis was extremely accurate at identifying changes in the HNMT protein (Fig. 4.1b). The only change that was missed was in $\beta 3$, which has increased flexibility in the mutant (105I) simulations (Rutherford et al., 2008b). Notably, this β-strand is solvent-exposed and already somewhat flexible; in fact, the difference in flexibility of the Cα atoms between variants is < 0.25 Å. According to Rutherford et al., $\beta 3$ undergoes a slight reorientation, but this reorientation happens quickly and is barely visible across simulations. Additionally, wavelets only identified a single change that Rutherford et al. did not identify, and this was helix αB. Helix αB is a slightly bent helix that is highly stable in all simulations. When we examined the wavelet coordinates for this helix, we noticed a long stretch of low-frequency wavelets from ~ 10 ns to ~ 20 ns located near the C-terminal end of helix αB in the first run of the 105I variant only. Upon examination of this trajectory, we discovered that the C-terminal end of this helix pivots relative to the rest of the helix and eventually unwinds (Fig. 4.4). Notably, this change is quite subtle and occurs in only one of the three 105I simulations. It is fitting, then, that wavelets

highlighted it with a lower p-value than other changes that occur in these simulations.

The protein TPMT undergoes several changes similar to HNMT that are well characterized by wavelet analysis (Fig. 4.1c). Wavelet analysis finds differences in all of the areas with observed changes according to Rutherford et al. (Rutherford and Daggett, 2008) and identifies only two neighboring loops and two helices, α4 and αB, as significantly different. Although α4 and αB are not mentioned in by Rutherford et al., it is important to note that there are slight reorientations associated with them. In the 80A variant, α4 has a less ordered orientation overall and is shifted from the 80P variant (Fig. 4.5a). Similarly, αB is highly disordered in both of the variants but is especially mobile in the 80P variant (Fig. 4.5b). These changes are both significant compared with other structural changes that occur in the protein. Thus it is not surprising that wavelet analysis identified them.

The COMT variants undergo changes similar to those seen in HNMT, probably due to the similar position and nature of the mutation. Wavelet analysis correctly identifies αB, α2, α3, α5, and α6 as areas in which significant changes occur. Interestingly, the strand β1 was the only β-sheet identified at the $p < 0.05$ level by wavelet analysis in all four proteins. Upon closer examination of the wavelet coordinates for this β-strand, we noticed that run 2 of the COMT 108M variant had a patch of significant low frequency ($\lambda \approx 12$ ns) motion occurring from 6.5 to 18.5 ns during the simulation where β1 and β4 bend apart from each other during this time range in this run only (Fig. 4.6). Although this motion was not addressed by Rutherford et al. (Rutherford et al., 2006; Rutherford and Daggett, 2009a), this motion explains why wavelet analysis highlighted this particular β-strand.

It is worth highlighting the fact that Rutherford et al. did not discuss three of the events discovered by wavelet analysis that we discuss here. This is likely

due to a combination of several things: 1) that the changes were not deemed structurally relevant to the mutation and its effects, 2) that the changes occurred in single simulations rather than consistently across identical runs, and 3) that the changes were difficult to locate without the aid of wavelet analysis. While all of these may be legitimate reasons not to report the events, one cannot argue that these events did not occur. Clearly wavelet analysis is capable of highlighting changes between variant proteins and structural events in a simulation that would take considerable effort to uncover without them.

4.6 Conclusions

As simulations become a more common way of exploring biological questions and generating hypotheses for further experimentation, it is critical that the initial analyses of these simulations be quick and efficient. Wavelet analysis has proved to be a very effective technique for accurately identifying all but the most subtle differences between protein simulations. Not only does wavelet analysis quantify these differences, but it is capable of identifying changes that may be missed or extremely elusive by traditional analysis. Critically, wavelet analysis can be automated and performed quickly, saving considerable time and effort.

Table 4.1: List of methyltransferase variants, their effects, and their associated diseases.

Protein	Variant	Primary Effect	Associated Disease
PIMT	119V	30% higher affinity for endogenous substrates[a]	heterozygosity: none
	119I	20% higher specific activity[a]	homozygosity: age-related disorders[a]
HNMT	105T	none	alcoholism[b], age-related disease[c]
	105I	2-fold lower activity[d]	none
TPMT	80A	none	none
	80P	no activity, low protein levels [e]	hematopoetic toxicity[f]
COMT	108V	none	none
	108M	30% lower activity, reduced stability[g]	breast cancer[h] and neuropsychiatric disease[i]

[a]DeVry and Clarke (1999); David et al. (1997); [b]Oroszi et al. (2005); Reuter et al. (2007); [c]Ligneau et al. (1998); Morisset et al. (2000); Panula et al. (1997); [d]Horton et al. (2001); Preuss et al. (1998); Scott et al. (1988); Girard et al. (1994); Price et al. (1993); [e]Krynetski et al. (1995); Tai et al. (1997, 1999); [f]Coulthard and Hogarth (2005); Deeken et al. (2007); [g]Rutherford et al. (2008); Shield et al. (2004); [h]Dawling et al. (2001); Goodman et al. (2002); Huang et al. (1999); Lavigne et al. (1997); Matsui et al. (2000); Mitrunen et al. (2002); Sazci et al. (2004); Tan et al. (2003); Thompson et al. (1998); Wedren et al. (2003); Yim et al. (2001); [i]Bilder et al. (2004)

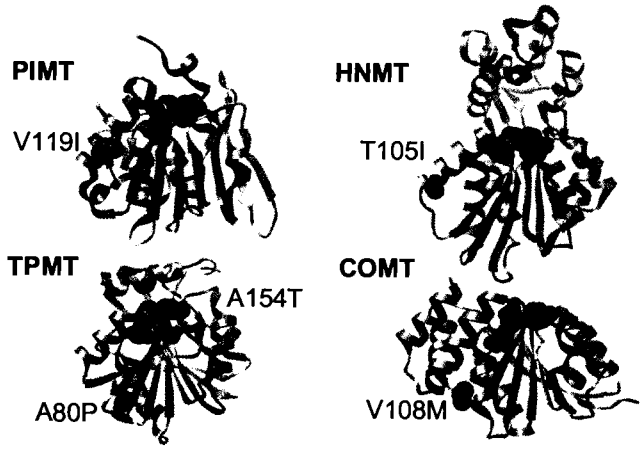


Figure 4.1: Common polymorphisms in four methyltransferase proteins. Ribbon structures of PIMT (*1i1n*), HNMT (*2aot*), TPMT (*2bzg*), and COMT (*3bwy*) are shown aligned by the core 7-stranded β-sheet. Polymorphic residues and S-adenosylmethionine (SAM) are shown in space-filling representation and colored in red and blue, respectively.

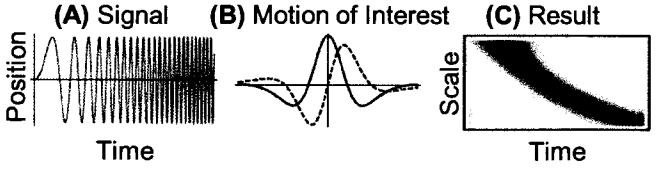


Figure 4.2: Explanation of wavelet analysis. **(a)** An example signal (e. g. the motion of an atom over time) whose frequency increases over time. **(b)** The Paul wavelet: a sinusoidal wave modified by a Gaussian. The real part is shown solid while the imaginary part is dashed. **(c)** The result of the wavelet analysis, showing that as time increases, the scale of motion decreases.

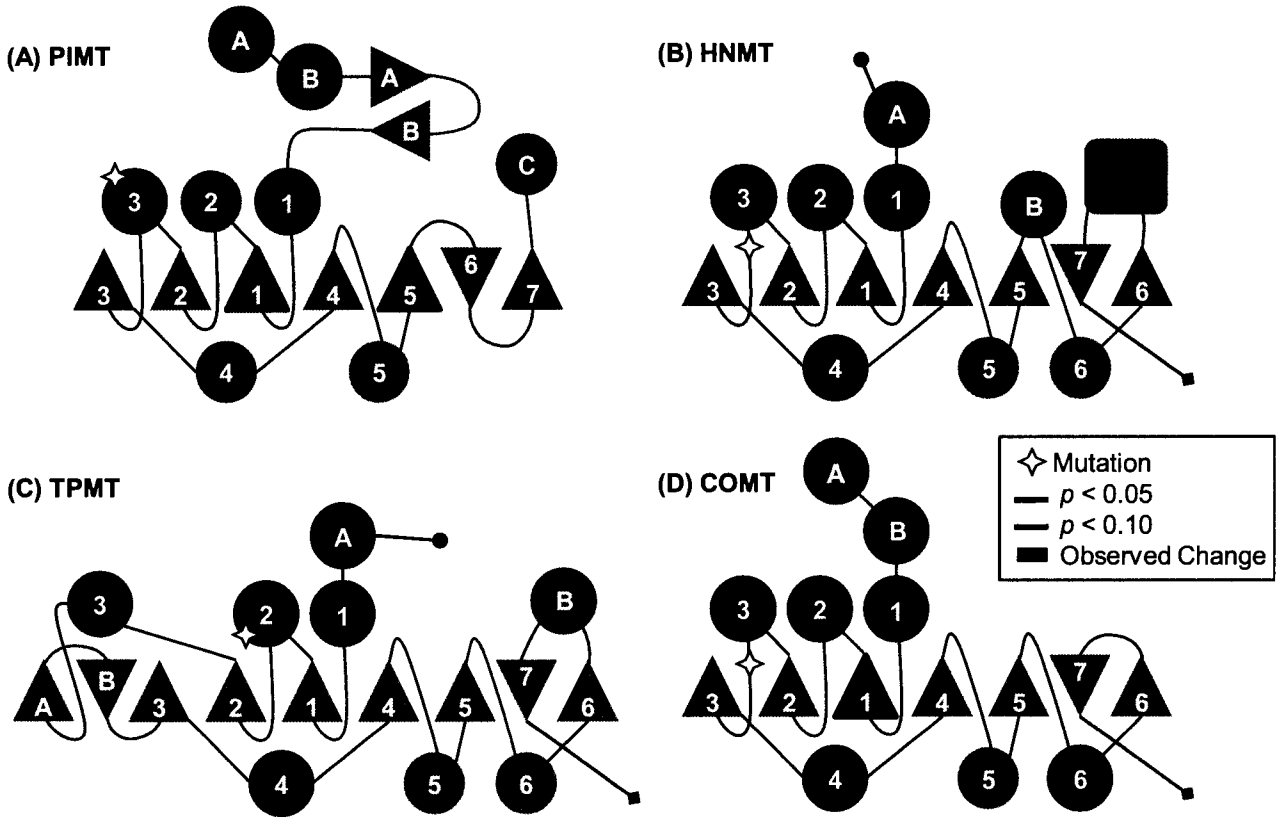


Figure 4.3: Results from wavelet analysis of the four methyltransferase proteins **(a)** PIMT, **(b)** HNMT, **(c)** TPMT, and **(d)** COMT. The proteins are shown schematically with circles representing α-helices and triangles representing β-sheets. The underlying fold similarities between proteins are numbered congruently. Significant changes between variants that were observed by Rutherford et al. are shown in green while red and purple borders indicate differences according to wavelet analysis at the $p < 0.05$ and $p < 0.1$ levels respectively.

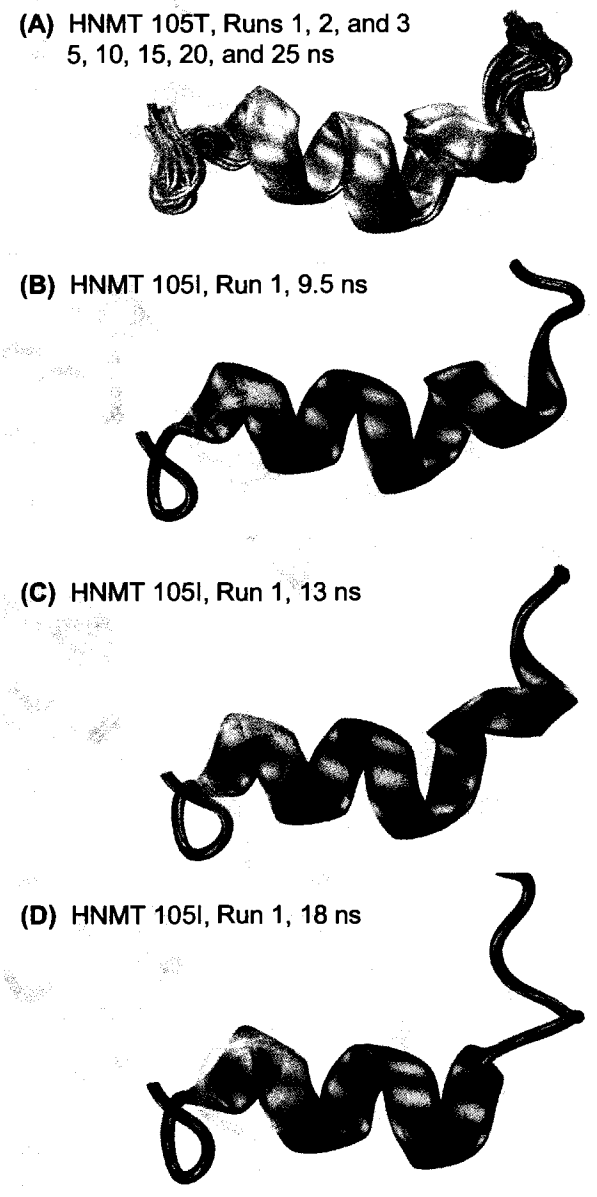


Figure 4.4: Helix αB of HNMT. **(a)** αB in all three runs of the 105T variant at 5, 10, 15, 20, and 25 ns each. **(b)** αB of run 1 of variant 105I at 9.5 ns, before the helix begins to unwind. **(c)** αB of run 1 of variant 105I at 13 ns. **(d)** αB of run 1 of variant 105I at 18 ns after the helix has begun to unwind.

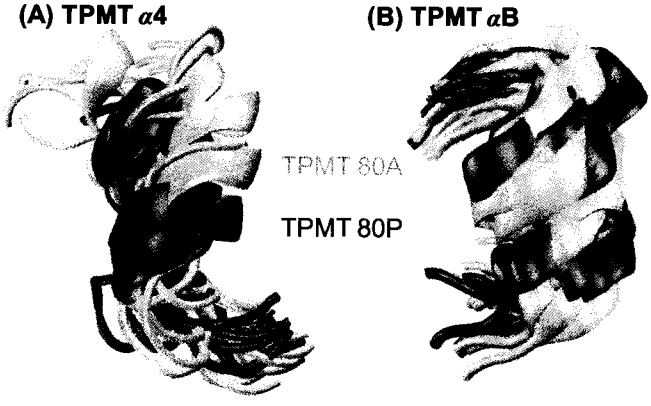


Figure 4.5: Helices **(a)** α4 and **(b)** αB of TPMT shown for triplicate simulations of both the wild-type (cyan) and A80P (red) variants. Snapshots were taken at 5, 10, 15, and 20 ns for each run for a total of 24 structures shown. Although αB is generally disordered, there is a noticeable difference in the orientation of α4 between the two variants.

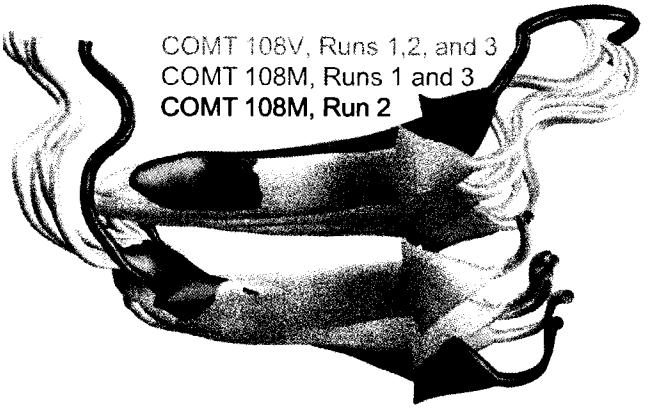


Figure 4.6: Strands $\beta 1$ and $\beta 4$ of COMT. Snapshots are taken from the range 10-16 ns for each simulation. The unique simulation in which the β-strands spread apart is colored red.

Chapter 5

A GRAPH THEORETIC APPROACH TO INDEXING PROTEIN DYNAMICS

5.1 Summary

Graphs are a natural and efficient tool for representing protein structure, yet graph theoretic approaches remain under-developed in the fields of molecular biology and molecular dynamics (MD). Many graph theoretic approaches have been used in the past with varying levels of success, but all have taken very simplistic approaches to representing the chemistry of a protein. We propose a novel graph representation of proteins that accurately encapsulates chemical and steric properties. We use this representation to capture the dynamic chemical environments of atoms and groups of atoms in MD simulations by analyzing the probability that a chemical group are in contact at specific times. We use these probabilities to index the dynamic chemical environment of each chemical group and demonstrate that these indices can be used to locate structural regions that are both chemically and dynamically similar but that cannot be located by simpler means such as searches based on residue packing. Finally, we compare these indexed chemical environments to chemical shifts from NMR and find good agreement between the two.

5.2 Introduction

Molecular dynamics (MD) simulations are a powerful and increasingly popular method for studying the motions of proteins at extremely fine resolution in time and space. In MD, Newton's equations of motion are integrated nu-

merically using a potential function to define interactions between individual atoms. At regular intervals, coordinates for the atoms in the system are written out, creating as a time-course movie for further analysis.

The Dynameomics project (Beck et al., 2008; van der Kamp et al., 2010) is a large-scale project whose goal is to simulate a member from every protein fold family (Day et al., 2003; Schaeffer et al., 2010). For 807 different fold families, at least one simulation has been performed for at least 31 ns at 298 K and at least two long simulation (31 ns) and three short simulations (2 ns) have been performed at 498 K. The Dynameomics database (Simms et al., 2008) is the largest database of its kind, containing $> 10^8$ structures in over 53 TB; it is more than four orders of magnitude larger than the PDB.

5.2.1 Graphs

Graph theory is a field that has found an incredibly diverse range of applications yet whose application to MD and protein chemistry in general has remained surprisingly rare. A graph G is defined as an ordered pair, (V, E), where V is a set of vertices or nodes and E is a set of edges connecting nodes ($E \subseteq V \times V$). In a graph, nodes represent objects or pieces of data while edges represent relationships between them. Graphs can be directed or undirected; in a directed graph, the edge (u, v) is not the same as the edge (v, u). For example, in a graph describing evolutionary relationships, an edge (A, B) might indicate that A is an ancestor of B. Additionally, edges can have weights that describe their properties; the weight of an edge e is usually written as a function, $w(e)$. Graphs can also be simple graphs or multi-graphs. In multi-graphs edges of the form (u, u) are allowed. Finally, graphs may be connected or disconnected; in a connected graphs, every node is connected to every other node by a path of edges. In this paper, we will be dealing exclusively with connected simple undirected weighted graphs, an example of which can be found in Figure 5.1.

5.2.2 *Protein Structure Graphs*

Graph representations of protein structure have taken many forms in protein science, most of which represent individual residues as nodes. Important variants of this motif include those that represent Cα distances as edges (Webber et al., 2001; Vendruscolo et al., 2002), those that represent sequence similarity patterns as edges (Giuliani et al., 2002), and those that represent correlated motions as edges (Amadei et al., 1993). These methods have been reviewed by Krishnan et al. (2008). Additionally, edges in a protein structure graph can represent atom-atom contacts within a certain cutoff distance or as determined by Delaunay tessellation (Delaunay, 1934; Huan et al., 2004).

Each representation of protein structure has its merits and has been used to elucidate a large range of structural properties. For example, Cα distance graphs have been used in the identification of amino acids that play a key role in folding (Vendruscolo et al., 2002), and contact graphs have been used to show that functional residues are those with high closeness values (Amitai et al., 2004). A more complete review of applications of various graph theory approaches in protein structure can be found in Böde et al. (2007).

Although several representations of protein graphs exist, to our knowledge no attempt has been made to distinguish between the varied properties of the amino acids in a protein aside from defining each amino acid as a unique type of node. While this approach is intuitive, it has the drawback of treating very similar nodes exactly the way it treats very different nodes. For example, Leu and Ile are as different as Leu and Arg. Additionally, it can bias calculations by treating large residues such as Trp identically to small residues such as Gly. Using such a representation, Trp is bound to be more central to the graph simply because it is so much larger thus can form contacts with so many more residues than Gly. Similarly, if a Val contacts the Cη of a Trp while a Leu

contacts its O, the Val and Leu do not have the same relation as if they were contacting the Cα and O of a Gly, respectively, yet residue-based graphs will often treat these cases identically. Accordingly, it does not stand to reason that, when a Trp that has been identified as important in a graph, the entire Trp is important. We propose a method of representing graphs that clusters atoms into nodes based on their chemical properties and covalent bonds. This representation conserves the similarity between amino acids and captures their overall properties without compromising the power of graphs to simplify and summarize a protein structure for additional analysis.

5.2.3 Protein Structure Comparison

The ability to compare protein structures and substructures is a very important area of molecular biology. The entire field of docking and computational drug design is based around the ability to locate structural motifs. Additionally, the ability to describe the chemical environment of a residue and to detect when it is similar to that of another is a critical piece of NMR. As important as these topics are, however, very few strides have been made in them, and attempts to characterize the often changing chemical environment of a residue that consider the dynamics of a protein are virtually nonexistent. Here we propose a graph construction and indexing strategy for proteins that considers not only the chemical similarity between residues and pieces of residues but that additionally is capable of characterizing the dynamic nature of proteins.

5.3 Methods

5.3.1 Molecular Dynamics Simulations

Simulations were run using our in-house simulation package, *in lucem* Molecular Mechanics (*il*mm) (Beck et al., 2008), and force fields for protein and

water that are described elsewhere (Beck and Daggett, 2004; Levitt et al., 1995, 1997). This paper examines 31 ns each of the 807 proteins from our Dynameomics project (Beck et al., 2008; van der Kamp et al., 2010) at 298 K for a total of ~ 17 μs.

5.3.2 Graph Construction

Proteins were divided into graph nodes by residue. Atoms in each residue were clustered according to their bonds and chemical properties, and each cluster became either a nonpolar, dipolar, positive, negative, partially positive, or partially negative node. Attempts were made to keep nodes approximately the same size: roughly 2-4 heavy atoms; though a few single atom nodes were allowed, and rings of 5 or 6 atoms were allowed to be a single node (e. g. Phe, Pro). Nodes with a total charge of > 0.5 or < -0.5 were considered positive or negative respectively; nodes with a total charge between 0.15 and 0.5 or between -0.15 and -0.5 were considered partially positive or partially negative, respectively. Nodes whose net charges were between -0.15 and 0.15 were considered nonpolar or dipolar depending on the distribution of charge within the node. Hydrogen atoms were considered part of the heavy atom to which they were attached and counted as part of that atom's charge, but were considered in deciding whether a node was dipolar or nonpolar. The protein backbone was divided into alternating dipolar (C, O, N) and nonpolar (Cα, Cβ) nodes. Any atom whose charge was > 0.5 or < -0.5 was restricted from being a member of a nonpolar, dipolar, or partially-charged node (e. g. Cγ in Asp, Cδ in Glu). Figure 5.2 shows two peptides containing each of the 20 amino acids with the individual nodes circled. The result of this process is a graph in which each node contains some small number of heavy atoms with similar chemical properties and edges represent covalent bonds between nodes. Notably, all heavy atoms are a member of exactly one node.

5.3.3 Graph Analysis

For each simulation, the covalent graph was established according the the above rules and maintained throughout the simulation using special covalent edges between nodes. For every picosecond of simulation, additional edges between nodes were recorded as contact edges based on an atom-atom cutoff distance of 4.6 Å; two nodes were considered to be in contact if at least one atom from the first node was within 4.6 Å of at least one atom in the second node. Additionally, Delaunay tessellation (Delaunay, 1934) was applied such that two atoms were not considered to be in contact if another atom occluded them.

This analysis created an ordered set of graphs, one for each picosecond of simulation, with identical nodes, identical covalent edges, and a changing set of contact edges. Because contact edges frequently are inconsistent and rapidly changing over a period of time in which two nodes would generally be considered to be "in contact," we applied a Gaussian smoothing to the edges. For each potential edge between every pair of non-covalent nodes, u and v, a discrete time-course signal, $C_{(u,v)}$, was constructed with each picosecond, $C_{(u,v)}(t)$, equal to 1 if u and v were in contact and 0 otherwise. A normal Gaussian curve with a standard deviation of 250 ps and a mean of 0 ($N(0, 250)$) was convolved with $C_{(u,v)}$ giving the signal $\boldsymbol{q}_{(u,v)}$, where $q_{(u,v)}(t)$ is equal to the probability that nodes u and v will be in contact if a time-point is randomly chosen according to the normal distribution $N(t, 250)$. Thus, $q_{(u,v)}(t)$ will be 0 if and only if u and v are never in contact and will be 1 if and only if u and v are always in contact. If u and v are always in contact from time $\tau - 250$ to $\tau + 249$, then $q_{(u,v)}(\tau) \geq 0.68$. Put simply, the value $q_{(u,v)}(t)$ is an index of connectedness or probability of contact of nodes u and v around time t. Covalently-bonded nodes are always considered to have an index of connectedness of 1.

5.3.4 *Node Communication*

In order to measure the degree to which two nodes, u and v, communicated with each other at a time t, we define the communication index between two nodes, $\kappa_{u,v}(t)$ to be the reciprocal of the length of the shortest path between u and v in the graph whose edge weights for any pair of nodes, a and b, are equal to $1/q_{a,b}(t)$ (or ∞ if $q_{a,b}(t) = 0$). This is described in Equation 5.1, where $P_{a,b}(t)$ is the set of all paths leading from node a to node b at time t and where each path is a set of edges.

$$\kappa_{u,v}(t) = \min \left\{ \frac{1}{\sum_{(a,b)\in S} 1/q_{(a,b)}(t)} \text{ where } S \in P_{(u,v)}(t) \right\} \tag{5.1}$$

Note that because all of our proteins are connected graphs, $0 < \kappa_{u,v}(t) \leq 1$. For example, two nodes that are always in contact would have a communication index of 1 at all times. If nodes a and b have a probability of contact of $1/2$ at time t, nodes b and c have a probability of contact of $3/4$ at time t, and a and c have no shorter route between them than a-b-c, then $\kappa_{a,c}(t) = 1/(2+4/3) = 3/10$.

To examine the chemical environment of a single node, we created a hash which counted, for each node, the number of nonpolar, dipolar, positive, and negative nodes whose indices of communication with the given node were $= 1$, > 0.75, > 0.5, > 0.25, and > 0 separately. Partially positive and partially negative nodes were considered positive and negative respectively due to their rarity. This gave each node's hash a type (nonpolar, dipolar, positive, or negative) and a 20-dimensional vector describing the kinds of nodes with which it communicated at a given time. Hash vectors were calculated for each of the nodes at 250 ps intervals. These hash vectors were sorted so that searches could be performed efficiently.

Several nodes and times were chosen at random from the set of all

nodes/times in Dynameomics and compared to all other nodes. Nodes with identical environment profiles were examined for similarity in their surroundings, Although it is difficult to quantify a dynamic chemical environment, several examples are presented here for the reader to evaluate. Additionally, we hypothesized that two similar nodes in similar chemical environments would have similar root mean square fluctuation (RMSF) values during the time window near their matching times. Although the entire set of nodes/times was far too exhaustive ($> 4 \cdot 10^9$ node/time pairs) for an all-versus-all comparison, we compared windowed RMSF values for a random subset of 100 nodes and their matching node/time pairs.

5.3.5 NMR Chemical Shift Comparison

In order to evaluate whether the chemical environment indices were accurate in characterizing the chemical environments of particular nodes, we examined the chemical shifts of the H atoms, as used in Nuclear Overhauser Effect Spectroscopy (NOESY). We obtained chemical shifts for 67 proteins with identical or nearly identical sequences to our proteins, as determined by BLAST, from the BioMagResBank (Ulrich et al., 2007). Residues from these sequences were aligned to our proteins using the Needleman-Wunch algorithm, BLOSUM62 matrices, and a gap of -4.

To locate protein nodes with similar chemical environments in terms of chemical shifts, we calculated an overall chemical environment score based on expected contacts over the entire simulation. For each node, we counted the expected number of contacts to each type of node (nonpolar, dipolar, positive, and negative/partially negative) with a probability of contact greater than 0.75, greater than 0.5, or greater than 0.25 over the entire simulation. This gave us counts of the nodes very frequently in contact (probability > 0.75), the nodes of each type often in contact (probability > 0.5), and the nodes occasionally in

contact (probability > 0.25) with the node in question, giving us a 12 dimensional vector of expected contacts with each node type at each probability level over the course of the simulation. The Euclidean distance between the vectors for each pair of nodes was then taken and the chemical shifts of the 100 pairs with the most similar chemical environments (all within 0.03 expected contacts of each other) were compared.

5.3.6 Graph Hubs

Previous research has examined network hubs in protein structure graphs. Brinda and Vishveshwara (2005) defined a hub as a residue that contacted at least 4 other residues with an interaction strength (defined by the number of interacting atoms normalized by the residue sizes) over a certain threshold and found that large planar residues such as Trp, His, Arg, and Phe were the most likely to form hubs. Here we extend this research by applying it to a much larger dataset that includes proteins simulated at 298 K in addition to minimized crystal and NMR structures using our own higher-resolution graph definition.

5.4 Results

Graph sizes ranged from very large (1,291 nodes) to very small (101 nodes) with an average of ~ 3.1 nodes per residue. Nonpolar nodes were by far the most common, comprising 45.4% of the dataset while dipolar nodes were a second at 41.3%. Positive and negatively charged nodes made up only 8.3% and 5.0% respectively.

Chemical environments for nodes with identical environment profiles were very similar in virtually all cases. In general, the number of node/time pairs with identical profiles was correlated with the packing density around the

nodes in question. Loosely packed nodes tended to create more hits than densely packed nodes. For any given node and time, by far the most common profile match was the same node at alternate times throughout the simulation. In addition to being somewhat trivial, these matches were so common that we will ignore them henceforth. The standard deviations of the RMSF of node/time pairs with identical communication profiles were uniformly low. The values ranged from below our level of precision to 0.672 Å with a mean of 0.105 Å. The number of node/times with identical profiles and the standard deviations of their RMSFs were not significantly correlated ($r \approx 0.374$).

Three node/time pair matches with identical communication profiles are presented here for the reader's examination. They were chosen based on the disparate properties of the queried node and on the relative rarity of the node/time pair. Figure 5.3 shows the first of these: a comparison between the negatively charged node made of the Oϵ atoms of residues D222 of Arginyl-TRNA Synthetase (*1bs2*, Fig. 5.3a) and D316 of Lactate Dehydrogenase (*1ceq*, Fig. 5.3b). In each case the node communicated with a positively charged residue (Arg or Lys), a His, a Tyr, a Phe, and a hydrophobic group (Ile or Val). Notably, in both cases, the Phe and the Asp which communicate strongly are adjacent in sequence while the hydrophobic Val/Ile and the positively charged Arg/Lys are within four residues in sequence.

The residue F22 of Bleomycin Resistance Protein at 20.5 ns (*1byl*, Fig. 5.4a) and the phenyl group of W125 in Benzoylformate Decarboxylase at 6.5 ns (*1bfd*, Fig. 5.4b) also had identical communication profiles. Both groups are located in hydrophobic pockets of their respective proteins, surrounded closely by hydrophobic residues such as Val, Ile, Met, Leu, and the phenyl rings of Trp and Tyr. Interestingly, both additionally have moderate communication with a His and Arg residue.

Figure 5.5 shows two positively charged nodes. The first is K203 from Lyti-

cus Protease 1 at 26.75 ns (*1arb*, Fig. 5.5a), and the other is R21 from Human Lysozyme at 1 ns (*1b5u*, Fig. 5.5b). Interestingly, each of these nodes has a nontrivial amount of communication with other positively-charged nodes. Both communicates with at least two Arg or Lys residues. Additionally, both are immediately in contact with a Tyr and communicate weakly with a Pro and an Asp.

Overall evaluation of graph nodes of the 100 pairs of nodes whose chemical environments were most similar over the course of their entire simulation for found an average difference in chemical shifts of 0.37 ppm ($\sigma = 0.28$ ppm). If the same value is calculated based on residue type instead of the chemical environments, the average difference is 0.74 ppm ($\sigma = 0.60$). Notably, the original calculation does not implicitly include residue type and frequently matches backbone nodes without identical residue types; in fact, only 5 of the 100 pairs of nodes had identical residue types.

The likelihood of a given node, residue, or residue/node (e. g. the positively charged Cγ node of Asp or the dipolar side-chain node of Gln) being a hub was similar for most residues and residue/nodes. Of the individual nodes, dipolar nodes were the most likely to be hubs with nonpolar nodes close behind. Negatively and positively charged nodes were each much less likely to be hubs than either dipolar or nonpolar nodes. Virtually all residues were similarly likely to be hubs, however, with Ile, Gln, Val, Asn, and Phe at the top. These were followed by Leu, Trp, Ala, Gly, Thr, Tyr, and Arg.

5.5 Discussion

Our protein structure graphs differ in several critical ways from traditional graph representations of proteins. Primarily, our graphs break the protein down into smaller subunits with similar or interdependent chemical properties, allowing simultaneously for a greater level of specificity in the meaning

of a contact and for a much simpler set of nodes, making certain graph computations more tractable. The one drawback of our schema is that the number of nodes increases by a factor of ~ 3. We did not find, however, that the time required for any calculations became unwieldly as a function of graph size.

The choice of a dynamic measure of communication also confers strong advantages over previous graph theoretic techniques, which were designed to study static protein structures. Proteins are not static molecules, and the set of atoms with which a given atom is in contact changes constantly. One cannot capture the set of residues that influence a given residue with a single static structure. Thus, while previous graph theoretic techniques have proven useful for studying protein structure, they must be reevaluated as biologists look more at the dynamic properties of a protein and not merely the static ones. By using a multi-layered communication profile that examined not only those nodes that a given node is constantly or frequently in contact with at a given time, we implicitly examine short- and long-range contacts. Those nodes that communicate with a given node with a low probability at time t are likely to be in direct communication with it at a more distant time. We suspect that it is this ability to capture all of a node's influences in a single vector that gives the technique its high accuracy.

The extremely low standard deviation of RMSF for the average set of node/time pairs with identical communication profiles is a strong indication that the profiles accurately capture the critical structural elements of a protein that contribute to dynamics. This coincides with the fidelity of the matches shown in Figures 5.3, 5.4, and 5.5. Each of these figures shows similar nodes with similar contacts at similar positions in their environments. It is additionally clear that a residue-based node system would not have been capable of finding these matches due to the similar properties but different amino acid labels of nodes such as Arg and Lys, which can occasionally serve similar roles

in a protein due to their positive charges.

The comparison of nodes by chemical shifts and graph-based chemical environments supports the accuracy of the graph representation as a means of capturing the chemical environment of an individual node. The average difference in chemical shift of 0.37 ppm may be insufficient for use in NOESY experiments and structure determination, but it is equivalent to chemical shift prediction systems such as SPARTA (Shen and Bax, 2007). It is worth pointing out that this accuracy was obtained despite the fact that chemical shifts are based on the H atom only while the graph-based chemical environments that were compared are for the closest equivalent, which is the O-C-N backbone node. Additionally, NMR experiments are performed over hundreds of ms, while our simulations are performed over only 31 ns.

The finding that dipolar and nonpolar nodes were slightly more likely to be hubs was expected as one would expect positively and negatively charged nodes to exist on the surface of the protein away from dense contacts. The residue-based findings were less predictable, however. Brinda and Vishveshwara (2005) found that planar residues (Trp, His, Phe, Tyr, and Arg) were the most likely to to form hubs with Ile and Leu likely to form weak hubs when the threshold for being considered a hubs was weaker (required fewer contacts). Our results support these claims but additionally find that Gln and Asn are likely to form hubs. Although this may be initially surprising, Asn and Gln both have side chains with hydrophilic and polar components which are similar to those of the protein backbone. It is worth noting that our calculations are not directly comparable to those of Brinda and Saraswathi because while theirs were calculated over crystal structures and were designed for static structures only, our calculations were performed over trajectories of dynamic proteins and were designed to consider non-static conformations. For example, an Asn residue that moved rapidly between states communicating

with several other residues might not be considered a hub at any given instant by Brinda and Saraswathi but could be considered one by our calculations. Thus, our findings should be considered a dynamic extension to their work and show that Asn and Gln may play a more critical role in protein structure once a protein is no longer in its low-temperature crystalline state.

One potential critique of our graph construction is that certain residues appear identical despite being different. The side-chains of an Arg and a Lys are not, for example, the same, but their graph representations are the same. It is important to keep in mind, however, that although the nodes are represented the same conceptually, the atoms that make up the node are still different, thus the positive node on the Arg is likely to have a different contact profile than the positive node on the Lys, making it difficult for them to occupy the same place in a node's communication profiles spuriously. In fact, it would be very unusual for an Arg and a Lys to occupy the same position in a profile for a node whose packing was very dense due to the difference in the size of the nodes. Figure 5.5 shows an Arg and Lys with similar profiles but which are not densely packed, for example.

It is worth noting that several concepts in this paper can be generalized readily. Among the most important of these are as follows:

Nodes The nodes that we define to construct our graphs could be specified or generalized. An obvious generalization of our nodes is to use residues as nodes, which has been well studied. Alternately, one could use common chemical groups such as phenyl groups and carboxylic acid groups.

Smoothing The choice of a standard deviation of 250 ps for the Gaussian smoothing reflected our desire to examine primarily a small window of time without completely losing the information about contacts as far as 1 ns in time. Very long simulations are being performed now which may

benefit from much larger standard deviations.

Probability The communication profiles in this paper were made up of counts of those nodes with probabilities > 0, > 0.25, > 0.5, > 0.75, and $= 1$. Obviously the nodes whose communication probabilities are equal to 1 are almost exclusively those nodes that are covalently bonded to a given node, thus occupy a critical role in the profile, but alternate profiles can easily be constructed if one is more interested in events occurring very close in time, for example. Additionally, removing the nodes with communication between 0 and 0.25 from a node's profile would tend to make the profile more focused on its immediate neighbors.

Although it is beyond the scope of this paper to examine the full range of possibilities using the graph theoretic techniques we have shown here, there is reason to believe that permutations of these techniques would be highly useful in a great deal of protein structure research.

5.6 Conclusions

We have described and improvement upon previous research in protein structure graphs and demonstrated their utility in examining protein structure. We have demonstrated that our metric of node communication profiles is an effective way to quantify the chemical environment of an individual node or residue in a protein. We have verified this with comparison to NMR chemical shifts. Finally we describe ways in which dynamic protein graphs can be generalized in order to encourage future research with protein graphs.

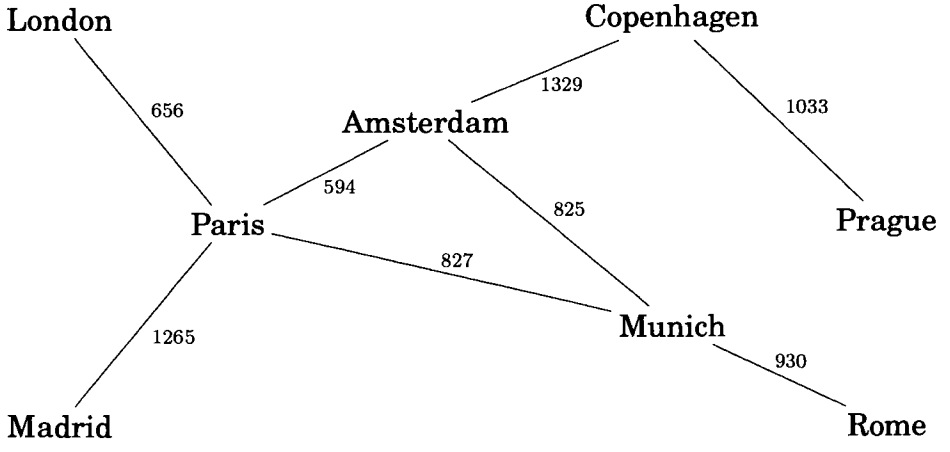


Figure 5.1: An example graph showing the distances between cities in Europe in km. In this graph, the set of nodes is {London, Copenhagen, Amsterdam, Paris, Prague, Vienna, Munich, Madrid, and Rome}. The set of edges is {(London, Amsterdam), (Copenhagen, Paris), (Copenhagen, Prague), (Amsterdam, Paris), (Paris, Munich), (Paris, Madrid), (Vienna, Rome), (Prague, Vienna)}. The graph is undirected, meaning the edge (Prague, Vienna) is the same as the edge (Vienna, Prague), and the edges have weights (the distances between cities). The graph is also simple and connected.

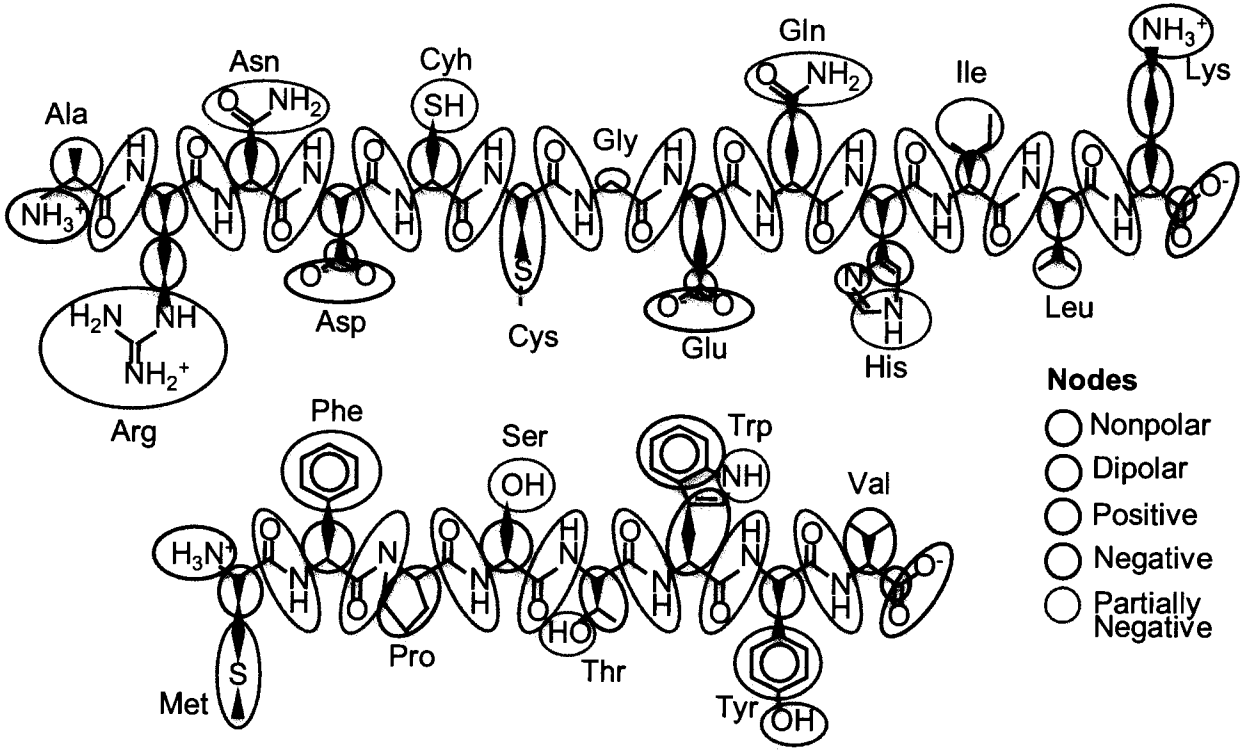


Figure 5.2: The graph representation of two short peptides with sequences ARNDCCGEQHILK and MFPSTWYV. Notably, only His has a partially negative node, and there are no partially positive nodes in our Dynameomics protein structures, but the node is supported for future compatibility. When two nodes are adjacent (contain atoms connected by covalent bonds), they are linked by a covalent edge in the graph.

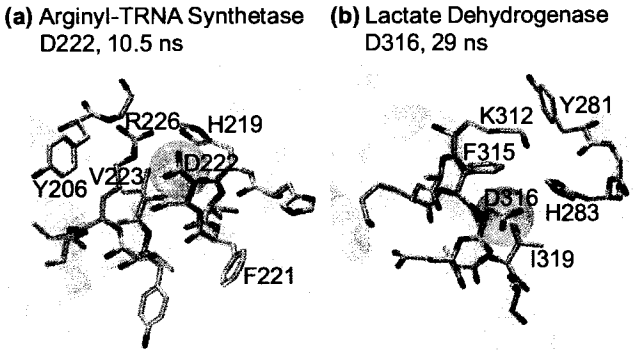


Figure 5.3: Two negatively charged nodes with identical node communication profiles and the residues with at least one atom within 10 Å. **(a)** The Oϵs of Arginyl-TRNA Synthetase (*1bs2*) residue D22 at 10.5 ns. **(b)** The Oϵs of Lactate Dehydrogenase (*1ceq*) residue D316 at 29 ns. Both nodes have close contacts with either an Arg or a Lys, a His, and either an Ile or a Val residue as well as more distant contacts with a Tyr and a Phe.

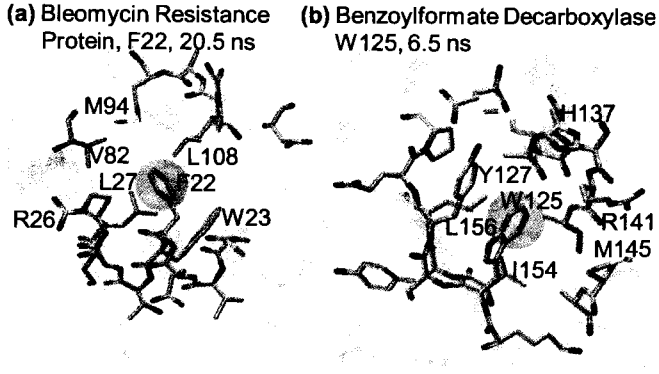


Figure 5.4: The residues with at least one heavy atom within 10 Å of two phenyl groups of residues **(a)** F22 of Bleomycin Resistance Protein (*1byl*) at 20.5 ns and **(b)** W125 of Benzoylformate Decarboxylase (*1bfd*) at 6.5 ns. Both rings exist in hydrophobic pockets near residues such as Leu, Val, Ile, and Met. Each is additionally in loose communication with an Arg and the Nϵ of either a His or a Trp.

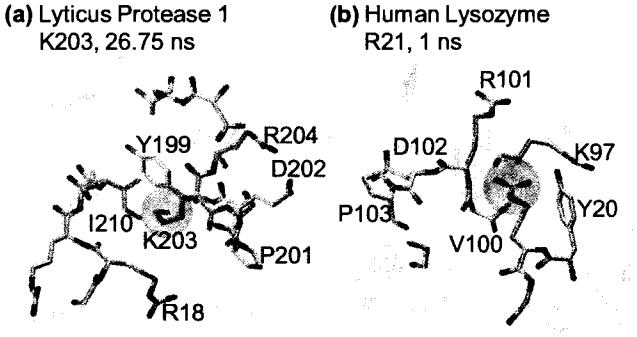


Figure 5.5: Two positively charged nodes and all residues within 10 Å of them. **(a)** Residue K203 from Lyticus Protease 1 (*1arb*) at 26.75 ns. **(b)** Residue R21 from Human Lysozyme (*1b5u*) at 1 ns. Both residues have loose but unusual communication with other positive nodes (Arg or Lys) as well as a Pro, They both exist in a relatively hydrophobic pocket in close communication with either a Val or Ile and a Tyr.

Chapter 6

GRAPH THEORETIC EVIDENCE FOR A FOUR STEP PROTEIN FOLDING/UNFOLDING PROCESS

6.1 Summary

Graphs are a powerful tool in many fields for describing the relationships that exist within complex systems. Although graphs have been used on some occasions to examine protein structure, they remain underused in the area of molecular dynamics (MD), partly due to the conflict of a discrete representation (graphs) of an inherently continuous data set (MD trajectories). Here we analyze the protein unfolding pathway of 183 proteins from our Dynameomics project by examining a graph theoretic measure, called the betweenness centrality, with graphs representing the dynamic protein structures. We find evidence that the slow separation of hydrophobic aryl rings follows protein core expansion in unfolding. We further demonstrate that the folding process is identical to the unfolding process in reverse by examining a refolding simulation of the engrailed homeodomain.

6.2 Introduction

The mechanism by which proteins fold is one of the biggest unsolved mystery in biology, partly due to the extreme difficulty studying it. Very few methods exist that allow one to study the structure of a protein as it unfolds. One such experimental method is Φ-value analysis (Fersht et al., 1992), which can reveal residues critical in the transition state (TS) structure of a protein. Molecular dynamics (MD) simulation, however, gives a much finer time and spacial reso-

lution than any experimental technique can currently offer. Although computational resources are still limited enough to make large-scale studying of the folding process difficult due to the long timescales on which folding occurs, thermal unfolding, which happens on a much shorter timescale, has been shown to be the reverse of the folding pathway (Day and Daggett, 2007; McCully et al., 2008).

The Dynameomics project (Beck et al., 2008; van der Kamp et al., 2010) is a large-scale project whose goal is to study the unfolding pathway of all proteins by simulating at least one member of every fold family. For 807 different fold families, at least one simulation has been performed for at least 31 ns at 298 K and at least two long simulation (31 ns) and three short simulations (2 ns) have been performed at 498 K. For a subset of 183 of the 498 K simulations, TSs have been identified (Jonsson et al., 2009) and verified using experimental Φ-values (Toofanny et al., 2010).

6.2.1 Protein Structure Graphs

Graphs have found limited application to protein structure research despite their intuitive application to molecules. Previous representations have tended to focus on representing each residue in the protein as a node in the graph with edges being drawn between nodes when they either have a critical number of atoms in close contact (Huan et al., 2004; Brinda and Vishveshwara, 2005), have Cα atoms within a certain distance (Webber et al., 2001; Vendruscolo et al., 2002), have correlated motions (Amadei et al., 1993), or when they have sequence similarity (Giuliani et al., 2002). These methods have been reviewed by Krishnan et al. (2008) and Böde et al. (2007).

Vendruscolo et al. (2002) have previously used protein structure graphs whose edges were defined by Cα distances to examine the TSs, as determined by Montecarlo simulation, of six proteins (Vendruscolo et al., 2001). They cal-

culated the betweenness centrality of each residue (Freeman, 1977), defined as the number of residue pairs whose shortest connecting path in the graph passes through the given residue, normalized by the total number of pairs of residues. Using this metric, they found that 2-4 residues or regions of each protein were commonly part of the protein nucleus in the TS. Additionally, they found that these residues could predict the nucleus of the native state, but their equivalents in the native state could not be used to predict the nucleus of the TS.

We have previously described a method for representing the structure of a protein as a graph by clustering atoms according to their chemical properties and covalent bonds (Benson and Daggett, 2010b). Figure 5.2 shows how a protein is divided into these nodes. These clusters of covalently linked atoms are turned into either dipolar, nonpolar, positive, negative, partially positive, or partially negative nodes depending on their chemical properties. In this paper, we further categorize nodes according to Table 6.1. A covalent edge is placed between two nodes if their constituent atoms are covalently linked, and a contact edge is placed between them if any pair of their constituent atoms is within 4.6 Å and is not occluded by another atom according to Delaunay tessellation (Delaunay, 1934). This is applied to each picosecond of a simulation giving an ordered set of graphs whose nodes and edges encapsulate the chemistry and interactions of each piece of the protein.

Using this graph definition, we extended the traditional contact measurements common in static protein structure comparisons by smoothing the contacts over time with a Gaussian. This creates a continuously changing graph whose instance at any given time represents the dynamic nature of each contact as a probability of being in contact around that time. Additionally, we defined the communication index between two nodes to be the reciprocal of the length of the shortest path between them in a reciprocal-weight probability graph. In other words, if the edges of a protein graph have weights $w_t((u,v))$

representing the probability of nodes u and v from graph G being in contact near time t, then an alternate graph, G', can be constructed with edge weights $1/w_t((u,v))$. The shortest path in G' between two nodes represents the most efficient path between the nodes in terms of the probable contact network. Supposing that $S_{a,b}(t)$ is the set of edges that form the shortest path between nodes a and b in this inverse-weight graph at time t, then the communication index between nodes a and b is given by Equation 6.1. We write the communication index of nodes a and b at time t as $\kappa_{a,b}(t)$.

$$\kappa_{a,b}(t) = \frac{1}{\sum_{e \in S_{a,b}(t)} 1/w(e)} \tag{6.1}$$

In other words, if the shortest path between nodes u and v is the path $(a_1 = u, a_2, a_3, ...a_n = v)$, then $\kappa_{(u,v)}(t) = (\Sigma_{i=2}^{n-1} 1/w((a_i, a_{i+1}))(t))^{-1}$. Figure 6.1 shows a small sample example graph with indices of connectedness and communication labeled. For further explanation of this and other graph concepts, see Benson and Daggett (Benson and Daggett, 2010b).

The betweenness centrality of a static graph is a graph theoretic concept used to examine the relative importance of a node to the overall graph. Nodes that occur in the shortest paths between many pairs of other nodes have higher betweenness than nodes that appear on few or no shortest paths. For a graph $G = (V, E)$ where $\sigma_{u,v}$ is the number of shortest paths from node u to node v (usually 1) and $\sigma_{u,v}(a)$ is the number of such shortest paths that contain the node a, then the betweenness centrality, $B(a)$, of a node a is given be Equation 6.2.

$$B(A) = \sum_{u,v \in V, u \neq v \neq a} \frac{\sigma_{u,v}(a)}{\sigma_{u,v}} \tag{6.2}$$

Here we propose an additional metric we call the impact of a node, which

serves as a dynamic local extension to the notion of betweenness. We use this metric to study and characterize the unfolding pathways of several proteins from our Dynameomics project whose TSs have been identified and verified.

The index of communication between two nodes serves as a measure of closeness within the protein graph, as adapted to a continually changing dynamic graph. To examine the importance of a single node to the overall communication within a protein, we define the k-impact of a node u at time t to be the number of pairs of nodes, not including node u, with communication indices $> k$ which, if node u is erased from the graph, have communication indices $\leq k$. In other words, this is the number of pairs of nodes which require node u to communicate above a certain level. Thus, if a node, u, is covalently bonded with two other nodes, which are restricted from contacting each other by the node u, then u will have a 1/2-impact of at least 1. This, however, almost never happens because nodes separated covalently by a single node are almost always in contact with a high probability, thus 1/2-impact values other than 0 tend to be rare. Accordingly, 2/5- and 1/3-impact values tend to be concerned primarily with nodes with which a given node is very frequently but not always in contact with. For example, Figure 6.4 shows the hydrophobic aryl node of F49 of the Engrailed homeodomain (*1enh*) and all of the residues for which it facilitates communication at a 1/3-impact level. This measure of impact generalizes the betweenness centrality to evaluate the localized part of a protein graph only. In this paper we are concerned primarily with the overall betweenness and the 1/3-impact of nodes, which can be considered to be the number of pairs of nodes that require a given node to communicate on a short (~ 0.5 ns) time scale.

6.3 Methods

6.3.1 Molecular Dynamics Simulations

All Dynameomics simulations were run using our in-house simulation package, *in lucem* Molecular Mechanics (*il*mm) (Beck et al., 2008), using the Levitt et al. (1995) force field and an explicit three-centered water model (Beck and Daggett, 2004; Levitt et al., 1997). Details of the Dynameomics protocol are given elsewhere (van der Kamp et al., 2010; Beck et al., 2008). This paper examines 31 ns each of the 183 proteins whose TSs have been identified by Jonsson et al. (2009).

6.3.2 Graph Analysis

For each simulation, graphs were established for each picosecond according to the rules described by Benson and Daggett (2010b). Nodes consisted of clusters of interdependent heavy atoms with similar chemical properties such as the hydrophobic aryl group of a Tyr residue or the Oϵ atoms of an Asp. Each node was labeled positive, negative, nonpolar, dipolar, partially positive, or partially negative according to its charges and their distributions. Highly charged atoms (those with partial charges > 0.5 or < -0.5) were not allowed to be parts of dipolar or partially charged nodes. Nodes range in size from 1-6 atoms, but the majority are 2, 3, or 4 atoms with larger and smaller nodes reserved for extreme cases such as phenyl rings and highly charged atoms. Backbones consisted of alternating dipolar nodes (each containing the C, O, and N atoms) and nonpolar nodes (each containing the Cα atom).

Nodes were covalently linked when at least one atom from one node was covalently bonded to at least one atom of another node. These bonds were maintained throughout the simulation. Nodes were considered in contact when at least one heavy atom from one node was within 4.6 Å of at least one atom from

another node and was not occluded according to Delaunay tessellation. Contact edges were calculated for every picosecond. In order to capture the dynamic nature of contact edges, each contact edge was smoothed with a Gaussian, giving, for each pair of nodes and each time, t, a probability that the pair of nodes will be in contact at a randomly selected time near t, if the time is selected according to a normal distribution centered at t and with a standard deviation equal to the Gaussian. This is essentially the probability of contact near time t. These contact probabilities make up a function $q_{(u,v)}(t) \in [0,1]$ for any pair of nodes, (u,v), which we call the index of connectedness. We chose a standard deviation of 250 ps in order to focus on an overall picture of the events near a given time. The value can be increased or decreased to focus on longer- or shorter-timescale events. Figure 6.3 shows a comparison of different values of the standard deviation using a contact between backbone nodes of A43 and K46 of the engrailed homeodomain's unfolding trajectory. We found that values below 250 ps tended to create probability functions that were not very smooth while values above 250 ps were relatively similar.

6.3.3 Unfolding

In order to characterize the unfolding pathways of our proteins, we divided each protein's dynamics up into nine regions: (1) native, (2) pre-TS, (3) TS, (4) TS + 0.5 ns, (5) TS + 1 ns, (6) TS + 5 ns, (7) TS + 10 ns, (8) TS + 15 ns, and (9) TS + 25 ns. For each of these regions we examined the betweenness and 1/3-impact of all nodes in each protein graph structure. Each node category was examined to determine its importance at different times in the unfolding pathway, and individual nodes with high 1/3-impact were examined closely during unfolding simulations to understand how structurally central nodes affected protein unfolding pathways.

6.4 Results

During native-state trajectories at 298 K, nodes with high 1/3-impact are almost exclusively limited to large hydrophobic side-chains. Ile, Trp, Phe, Tyr, Leu, and Val are the most common. Figure 6.2 shows three proteins with all nodes whose 1/3-impact were > 25 at various times in their native-state simulations highlighted. The impact values of nodes are approximately geometrically distributed and do not change considerably over the course of a 298 K simulation (Fig. 6.5a). This is true for 498 K simulations as well, but there is no significant correlation between nodes with high 1/3-impact at 298 K and nodes with high 1/3-impact at 498 K, as demonstrated by Figure 6.5b. The 1/3-impact of nodes was considerably lower at 498 K than at 298 K, while the betweenness remained comparable.

The betweenness centrality and 1/3-impact of each node category over the course of the unfolding simulation are shown in Figures 6.6 and 6.7, respectively. Backbones nodes (C, O, and N atoms) and guanidinium nodes (Arg side-chain terminus) had the most dramatic increase in betweenness of all node categories; the betweenness values of both increase steadily through the unfolding simulation and reach at least twice their initial values. Hydrophobic aryl groups and hydrophobic branches (Ile, Leu, and Val side-chains) are the only node categories that drop considerably in betweenness during unfolding, reaching approximately half their initial values. Carboxylic acids, hydrophobic linkers (side-chain carbon atoms including and immediately following the Cα), amides (Gln and Asn polar side-chain heads), and amines (Lys Nζ) steadily increase in betweenness as well though less dramatically than hydrophobic aryl and hydrophobic branching groups. Other partial rings (His side-chain, non-aromatic part of Trp side-chain, and Pro side-chain), alcohols, and thiols show almost no change or variance throughout the unfolding simulation in between-

ness.

Interestingly, all groups show an initial drop in 1/3-impact; for guanidinium groups, amines, and carboxylic acids, this is only present between the native-state and pre-TS time-points; by the TS, they have all increased in 1/3-impact. For backbones, hydrophobic linkers, alcohols and thiols, and partial rings, the decrease lasts until just after the transition state then remains steady. Hydrophobic aryl rings and branches show 1/3-impact decreases until just after the TS then shower decreases after that. Backbone atoms have the highest overall 1/3-impact at all times during the simulations.

All node categories except hydrophobic aryl rings have their largest change in betweenness immediately following their TS; this is especially noticeable in guanidinium nodes and amides, both of which increase dramatically and suddenly. Hydrophobic aryl rings and hydrophobic branch groups both have a sharp drop in betweenness from values near 3% to closer to 2% following their TS. In the case of hydrophobic aryl rings, the greatest sudden drops occur ~ 1 ns after the TS, while for hydrophobic branches it follows the TS immediately. Notably, if large/bulky hydrophobic branches (Ile and Met side-chains) are separated from the smaller hydrophobic nodes, they show an overall drop in betweenness identical to other branched side-chains rather than a similarity to the other large/bulky aryl side-chains. These drops in hydrophobic residue betweenness represent all of the immediate loss of betweenness following the transition state as categorized here.

Betweenness and 1/3-impact by node category were remarkably stable. Convergence was observed quickly, and sets of as few as three protein trajectories produce virtually identical graphs. Variance was low over time for all groups as well, with the exception of the N-terminal and C-terminal groups, which are not discussed here for this reason. Alcohols and thiols, partial rings, hydrophobic linkers, amines, and carboxylic acids had especially low variance over time

for both betweenness and 1/3-impact. Guanidinium groups had the highest variance, but the variance remained stable until $\sim$ 1 ns after the transition state and can still be confidently viewed as an upward trend.

Although the betweenness and the changes in betweenness seem initially small, only a few percent, it is worth noting that the betweenness for any protein graph should be low. This is because the vast majority of pairs of nodes can be traversed in only a few steps, either along covalent bonds (e. g. from residue i to residue $i+1$) or along a single contact. Each of these represents a shortest-path with no "between" nodes. A change in betweenness of 1% can represent a substantial change in geometry as well; if a node's betweenness increases by 1% in a small protein of 200 nodes, this represents an increase of 199 pairs of nodes whose shortest paths travel through the given node.

6.5 Discussion

The slow increase of backbone betweenness can be seen as a correlate to the expansion of the overall protein core; in a completely expanded protein, backbone atoms would be between virtually every pair of nodes. Hydrophobic linkers hold a similar but lesser position, as an expanded protein requires that more shortest paths travel through the carbon atoms that link the side-chain group to the backbone. The extremely low betweenness of all alcohols and thiols was initially surprising, and suggests that these groups do not play a major role in structure and structure formation on average. This is partially explained by their similarity to and preference for contacting water over other residues in the protein; even charged residues such as Lys will prefer to contact charges such as Asp in the absence of negatively charged ions.

The extremely high 1/3-impact of backbone nodes is also unsurprising; backbone C, O, and N atoms are very close to several graph nodes and are likely to facilitate communication between them due to their closeness. The initial drop

in backbone 1/3-impact is most likely due to a loss in hydrogen bonds. This does not necessarily lead to a drop in betweenness, however. The betweenness of a node considers the shortest path between every pair of nodes in a protein and counts those of which it is a member, even those that are quite distant; the 1/3-impact of a node only considers the shortest paths between the nodes that are in its local environment. Because the local environments of most nodes lose members as the protein expands (and nodes become further apart), the 1/3-impact must necessarily shrink because there are fewer pairs of nodes with shortest paths to consider. This explains the initial drop in 1/3-impact of all nodes. However, although a node's local environment loses members as the protein expands, this can lead to an increase in betweenness for some nodes because the shortest paths for very distant nodes are forced to travel through the few nodes that still connect in the protein's core.

Because a smaller environment generally means a smaller 1/3-impact, this indicates that carboxylic acids, guanidinium groups, and amines increase the number of nodes in their local environments during unfolding simulations in order to increase 1/3-impact. This is not particularly surprising considering that these charged groups will likely seek out other charged groups in the protein once the hydrophobic core has been disrupted. Notably, however, guanidinium groups and carboxylic acids reach levels of 1/3-impact that are observed in hydrophobic aryl groups and hydrophobic branches during native-state simulations; this suggests that the arrangements of these nodes relative to the protein's overall structure is similar to that of the hydrophobic groups' in the native state.

Jonsson et al. (Jonsson et al., 2009) report, in their study of the same 183 protein unfolding simulations examined here, that contacts decrease for hydrophobic residues in the TS, especially for Phe, Trp, and Tyr, while simultaneously increasing slightly or remaining the same for charged groups such as

Asp and Lys. It is initially puzzling, then, that we find that the betweenness of hydrophobic aryl rings and hydrophobic side-chains do not decrease substantially until after the TS. Hydrophobic aryl rings, especially, have a comparable betweenness to their native state values until ~ 1 ns after the TS. Hydrophobic branches similarly remain at their native-state levels until ~ 0.5 ns after the TS. The 1/3-impact values of both of these groups drop both before and after the TS, however.

One possible explanation of these data is that the TSs were picked too early in the simulations; this, however, contradicts the sudden change in every node category's betweenness immediately after the TS except for hydrophobic aryl rings, all of which support the hypothesis that TS were chosen correctly. It is important to keep in mind, at this point, that betweenness and 1/3-impact are not functions of contacts. In fact, a smaller number of contacts in a graph can lead to some nodes having significantly higher betweenness because fewer shortest-paths are able to travel through local contacts and instead must travel through the few nodes that make long-range contacts. The 1/3-impact values follow a similar pattern but are concerned with the smaller graph formed by the local environment of a node only. Thus, nodes will show a decrease in 1/3-impact when many contacts are lost but an increase when nearby nodes become less connected with each other.

In our simulations, we observe that this exact phenomenon drives the change in betweenness of different kinds of nodes over the course of our unfolding simulations. We observe a discrete set of events in virtually all trajectories and propose it as a 4 step method for the early thermal unfolding of proteins through the transition state. This process is outlined below, and illustrated by Figure 6.6, which uses the unfolding pathway of barnase as an example.

N. **Native.** The native structure of a protein is characterized by high levels of betweenness for hydrophobic groups, especially hydrophobic aryl rings, which occupy the core; the vast majority of shortest-paths must travel through the core via some subset of these nodes. The core is densely packed, so these hydrophobic residues have large numbers of contacts. Notably, hydrophobic branching nodes and hydrophobic aryl rings have virtually identical betweenness in the native state, though hydrophobic aryl rings have higher 1/3-impact due to their ability to occlude smaller nodes from interacting with each other directly.

1. **Expansion.** Before the protein's TS, the structure remains native-like but expands significantly, spreading residue side-chains within the core apart. These internal residues are largely composed of hydrophobic aryl groups and hydrophobic branches such as those made up of the Cγ and Cδs atoms of Leu. The nodes representing these groups have high betweenness initially due to their centrality to the 3D structure. As the protein expands, they may lose some betweenness on average, but because they remain in the core of the protein, many shortest-paths must still travel through them. The 1/3-impact drops more significantly than betweenness due to the expansion of the protein, supporting the notion that expansion is occurring and that smaller local regions of the protein are becoming more expanded during this time. Figure 6.6b shows the slight expansion of the region around Y78 in barnase; although the residues have spread apart, it maintains its neighborhood of contacts.

2. **Transition.** The protein's TS is an expanded version of the native state, and hydrophobic nodes, though still in the core, have lost contact with many of their neighbors. Hydrophilic residues, on the other hand, maintain their contacts and may seek out new contacts with each other in the

less dense solvent. In Figure 6.6c, Y78 remains in the core and in contact with I51, but its neighborhood of contacts has dwindled significantly. At this point, these nodes maintain their betweenness from the expansion state, as most shortest-paths still travel through the core, thus through them.

3. **Hydrophilic Invasion.** During the $\sim$ 0.5 ns following the TS, hydrophilic residues such as Asp, Lys, Arg, and Glu begin to enter the the protein's center along with water while hydrophilic core residues continue to expand; this creates a dual environment of both hydrophobic and hydrophilic residues and can be observed via the increase in 1/3-impact of guanidinium groups, carboxylic acids, and amines, all of which become more likely to cluster in small groups. Very hydrophobic aryl groups remain in the core preferentially while most smaller hydrophobic groups such as Ala, Leu, and Ile become too separated to maintain contacts in the hydrophobic core due to the protein's continued expansion. The departure from the core of the smaller hydrophobic groups leads to a sudden drop in their betweenness, as they no longer bridge distant regions of the protein. Hydrophobic aryl groups maintain their betweenness better on average, as they have not yet been pushed out of the core. The lack of other smaller side-chains in the core prevents them from maintaining their 1/3-impact, which has dropped rapidly to a low. This drop occurs because individual contacts to one or two other nodes can maintain betweenness so long as they bridge distant regions, but not 1/3-impact, which depends only on the communication between those nearby nodes. Simultaneously, charged and hydrophilic groups such as carboxylic acids, amides (Asn and Gln side-chains) and amines see a sudden jump in betweenness as they begin to bridge distant protein regions. The combined exodus of hydrophobic

groups, continued expansion of the protein core, and contacts formed between hydrophilic residues and polar backbone atoms lead to a jump in the betweenness of the polar backbone atoms as well. Figure 6.6d shows Y78 remaining in the protein core while other hydrophobic groups begin to exit and while charged residues such as K27 and D75 begin to form the contacts that bridge distant regions of the protein.

4. **Hydrophobic Decomposition.** By $\sim$ 1 ns after the TS, the hydrophobic residues that remain in the core are insufficient to hold it together and the hydrophobic core breaks apart, as is visible in 6.6e, where Y78 no longer maintains its most distant residue contacts. The breaking of the core leads to a sharp drop in the betweenness of the hydrophobic aryl groups, which no longer connect with residues in the core. In fact, this drop brings the betweenness of hydrophobic aryl groups to a level below that of hydrophobic branches. These smaller hydrophobic groups continue to lose betweenness as well because those still remaining in the core continue to exit, but the effect, comparatively, is much smaller. Hydrophilic residues remain in contact, preserving, and eventually increasing, the betweenness of amines, amides, guanidinium groups, and carboxylic acids.

U. **Unfolding.** Following the decomposition of the hydrophobic core, the protein continues to expand and rearrange, with hydrophobic residues continuing to lose betweenness as they spread apart and hydrophilic residues continuing to gain betweenness as they become the only substantial links between the expanded backbone of the protein. The charged guanidinium Arg side-chains do especially well at this point, as they are the largest hydrophilic group, giving them a likeness to aryl rings at lower temperatures.

This 4 step model for unfolding explains all of the measurements observed here and in previous unfolding studies (Jonsson et al., 2009). When unfolding occurs in reverse as folding, we would expect these events to occur backwards. Early in the refolding process, only hydrophilic residues would be in contact. These residues would bring large hydrophobic side-chains into contact. Following this, hydrophobic residues would invade, leading to the TS, which would precede general contraction of the core. To test this theory, we examined the unfolding and refolding pathway observed in a 373 K simulation of the engrailed homeodomain (McCully et al., 2008). In this simulation, the initial (unfolding) transition state occurs at 1 ns while the second (refolding) transition state occurs at 5 ns. Figure 6.6 shows this pathway. Notably, the increased betweenness of the hydrophilic residues occurs ~ 0.5 ns after the transition state, as hydrophobic residues separate and hydrophilic residues begin to connect more strongly. In the reverse of this process, the hydrophilic residues remain in contact to a lesser extent after the TS, but they are pushed to the edge of the protein rather than remaining in a central (between) location.

6.6 Conclusions

We have shown evidence from graph theoretic measurements that protein unfolding through the transition state occurs via a common process, which involves the separation of large aryl rings after the separation of smaller hydrophobic contacts. As these aryl rings break apart, hydrophilic interactions gain dominance. This process involves four steps: expansion, transition, hydrophilic invasion, and hydrophobic decomposition. We have demonstrated with examples how this leads to unexpected changes in betweenness measures of chemical groups as compared to their contacts over time, and show that large hydrophobic aryl groups play a different role in this process than smaller hydrophobic groups. We further show that the folding process occurs through the

same mechanism in reverse.

Table 6.1: Protein structure graph node categories.

Node Category	Residues	Atoms
N-termini	All	N-terminal Ns
C-termini	All	C-terminal Cs, Os, and OTs
Backbones	All	C, O, N_{i+1}[1]
Guanidinium Groups	Arg	Nϵ, Cζ, Nη^2
Hydrophilic Aryl Rings	Phe	Cγ, Cδ, Cϵ, Cζ^2
	Trp	Cδ, Cϵ, Cη^2
	Tyr	Cγ, Cδ, Cϵ, Cζ^2
Hydrophobic Branches	Ile	Cγ, Cδ^2
	Leu	Cγ, Cδ^2
	Leu	Cα, Cβ, Cγ^2
Amides	Asn	Cγ, Oδ, Nδ
	Gln	Cδ, Oγ, Nγ
Amines	Lys	Nζ
Carboxylic Acid Os	Asp	(Oδ), (Cγ)[2][3]
	Glu	(Oϵ), (Cδ)[2][3]

[1]The backbone nodes are composed of the C and O atoms of residue i and the N atom of residue $i+1$.
[2]When more than one node is grouped in one row, each node is specified by parentheses.
[3]When multiple atoms (e. g. Cγ1, Cγ2) exist, both are specified.

Table 6.1: (continued)

Node Category	Residues	Atoms
[3] Hydrophobic Linkers	Ala	$C\alpha$, $C\beta$
	Cys	$C\alpha$, $C\beta$
	Asp	$C\alpha$, $C\beta$
	Glu	$C\alpha$, $C\beta$, $C\gamma$
	Phe	$C\alpha$, $C\beta$
	Gly	$C\alpha$
	His	$C\alpha$, $C\beta$
	Ile	$C\alpha$, $C\beta$
	Lys	($C\alpha$, $C\beta$), ($C\gamma$, $C\delta$, $C\epsilon$)[3]
	Leu	$C\alpha$, $C\beta$
	Met	($C\alpha$, $C\beta$), ($C\gamma$, $S\delta$, $C\epsilon$)[3]
	Asn	$C\alpha$, $C\beta$
	Gln	$C\alpha$, $C\beta$, $C\gamma$
	Arg	($C\alpha$, $C\beta$), ($C\gamma$, $C\delta$)[3]
	Ser	$C\alpha$, $C\beta$
	Thr	$C\alpha$, $C\beta$, $C\gamma$
	Tyr	$C\alpha$, $C\beta$
Partial Rings	His	($C\gamma$, $C\delta$), ($C\epsilon$, $N\delta$), ($N\epsilon$)[3]
	Pro	$C\alpha$, $C\beta$, $C\gamma$, $C\delta$
	Trp	($C\alpha$, $C\beta$, $C\gamma$, $C\delta$), ($N\epsilon$)[3]
Alcohols and Thiols	Cyh	$S\gamma$
	Ser	$O\gamma$
	Thr	$O\gamma$

[3]When multiple atoms (e. g. $C\gamma1$, $C\gamma2$) exist, both are specified.

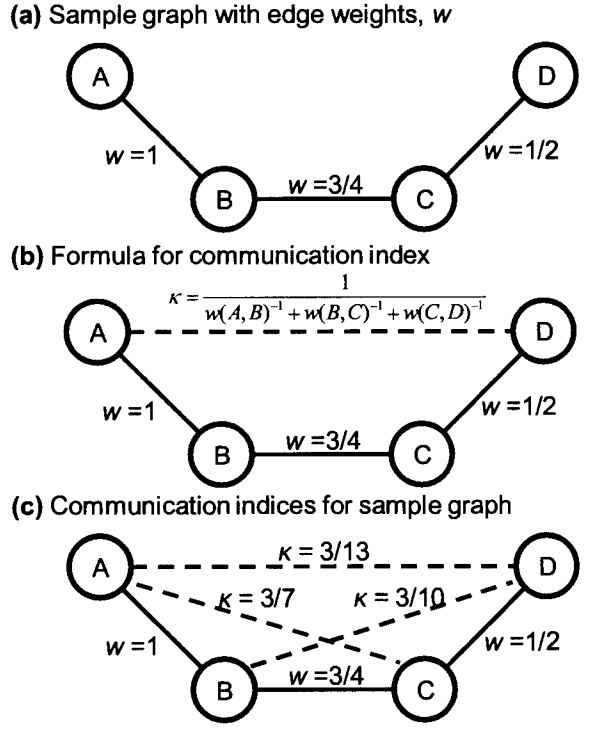


Figure 6.1: A small graph of four nodes with probabilities of contact shown on solid edges and communication indices shown on dashed edges. **(a)** The sample graph with probability of contact edge weights shown. **(b)** The same graph with an edge between nodes A and D showing the formula for that edge's communication index, κ. **(c)** The same graph with all communication indices shown. Contact probabilities are assumed to be 0 when not shown. In this graph, the node B would have a 1/2-impact of 0, a 1/3-impact of 1 (nodes A and C), a 1/4-impact of 1 (nodes A and C), and a 1/5-impact of 2 (nodes A and C and nodes A and D). Node C also has a 1/5-impact of 2 and a 1/4-impact of 1, but has a 1/3-impact of 0.

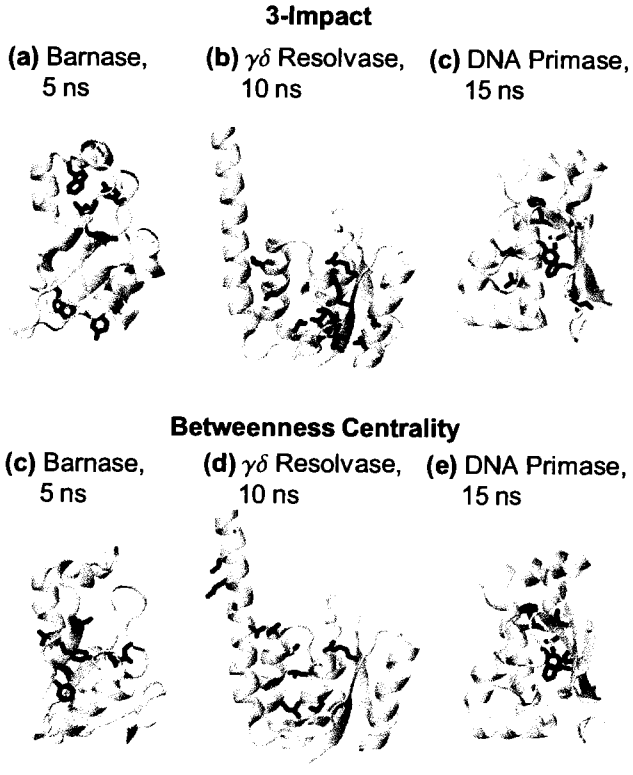


Figure 6.2: Three proteins with all residues whose 1/3-impact is > 25 and whose betweenness is > 0.8 during their 298 K simulation shown. **(a)** 1/3-impact of Barnase (*1a2p*) at 5 ns. **(b)** 1/3-impact of $\gamma\delta$ Resolvase (*1gdt*) at 10 ns. **(c)** 1/3-impact of DNA Primase (*1dde*) at 15 ns. **(d)** betweenness of Barnase (*1a2p*) at 5 ns. **(e)** betweenness of $\gamma\delta$ Resolvase (*1gdt*) at 10 ns. **(f)** betweenness of DNA Primase (*1dde*) at 15 ns. The core of barnase is less densely packed than $\gamma\delta$ resolvase or DNA primase; thus it has fewer 1/3-impact hubs than the other two. Notably, the elongated helix in $\gamma\delta$ resolvase causes several nodes along it to have a high betweenness even when they have low 1/3-impacts due to the many shortest-graph-paths that must travel along the helix.

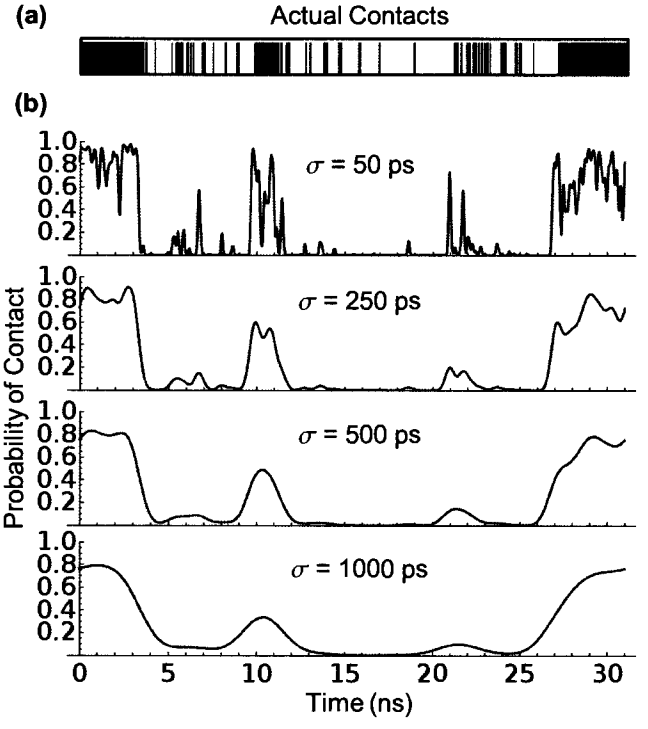


Figure 6.3: Comparison of standard deviation values for the smoothing of graph contacts over time. **(a)** Actual contacts from the unfolding simulation of the engrailed homeodomain between backbone nodes of residues A43 and K46 over time. **(b)** Probabilities of contacts between backbone nodes of residues A43 and K46 over time as smoothed by a Gaussian with a standard deviation of 50, 250, 500, or 1000 ps.

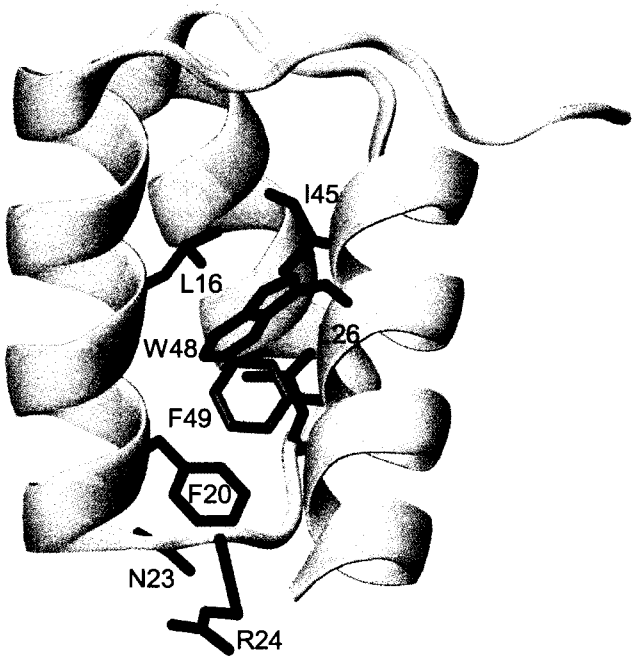


Figure 6.4: Residue F49 of the engrailed homeodomain native state simulation at 5 ns and the residues with which it facilitates communication at a 1/3-impact level. The 1/3-impact of the aromatic group of F49 (veiled in purple) is 32; from these 7 residues, 11 different nodes form 32 pairs, each of which requires this hydrophobic aryl group to communicate at near time $t = 5$ ns at the 1/3-impact level. The aromatic group of F20 and the side chain of L24 (behind F49) are one example of these pairs. Notably, not all nodes in each residue contribute to the impact of F49; only the backbone nodes of R24, for example, are involved.

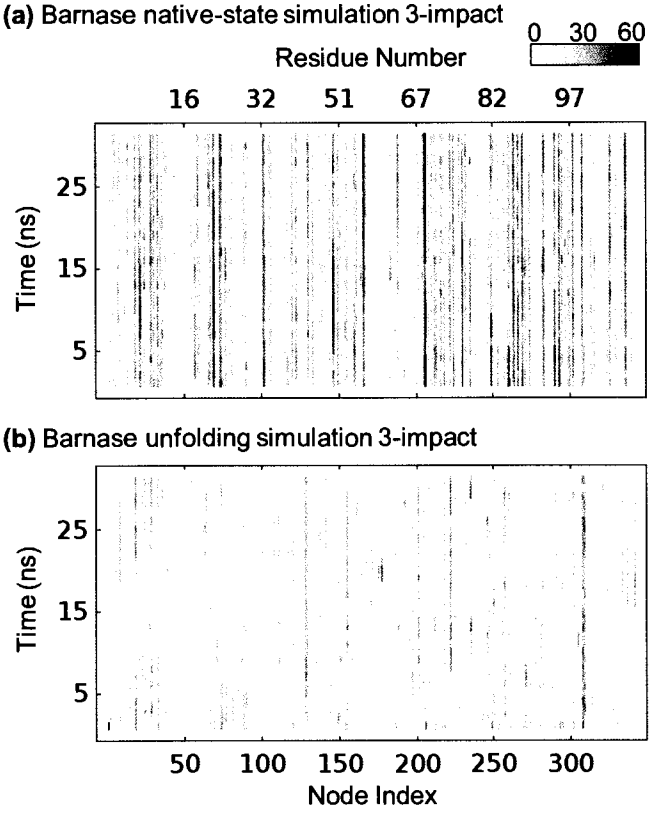


Figure 6.5: The 1/3-impact of all nodes over the course of **(a)** the native-state and **(b)** the unfolding simulation of the protein barnase (*1a2p*). 1/3-impact values ranged from 0 to 60 and changed little over the course of the native-state simulation for any individual node. The 1/3-impact drops considerably during the unfolding simulation, but is also relatively stable. The transition state for barnase occurs at 0.24 ns in this simulation.

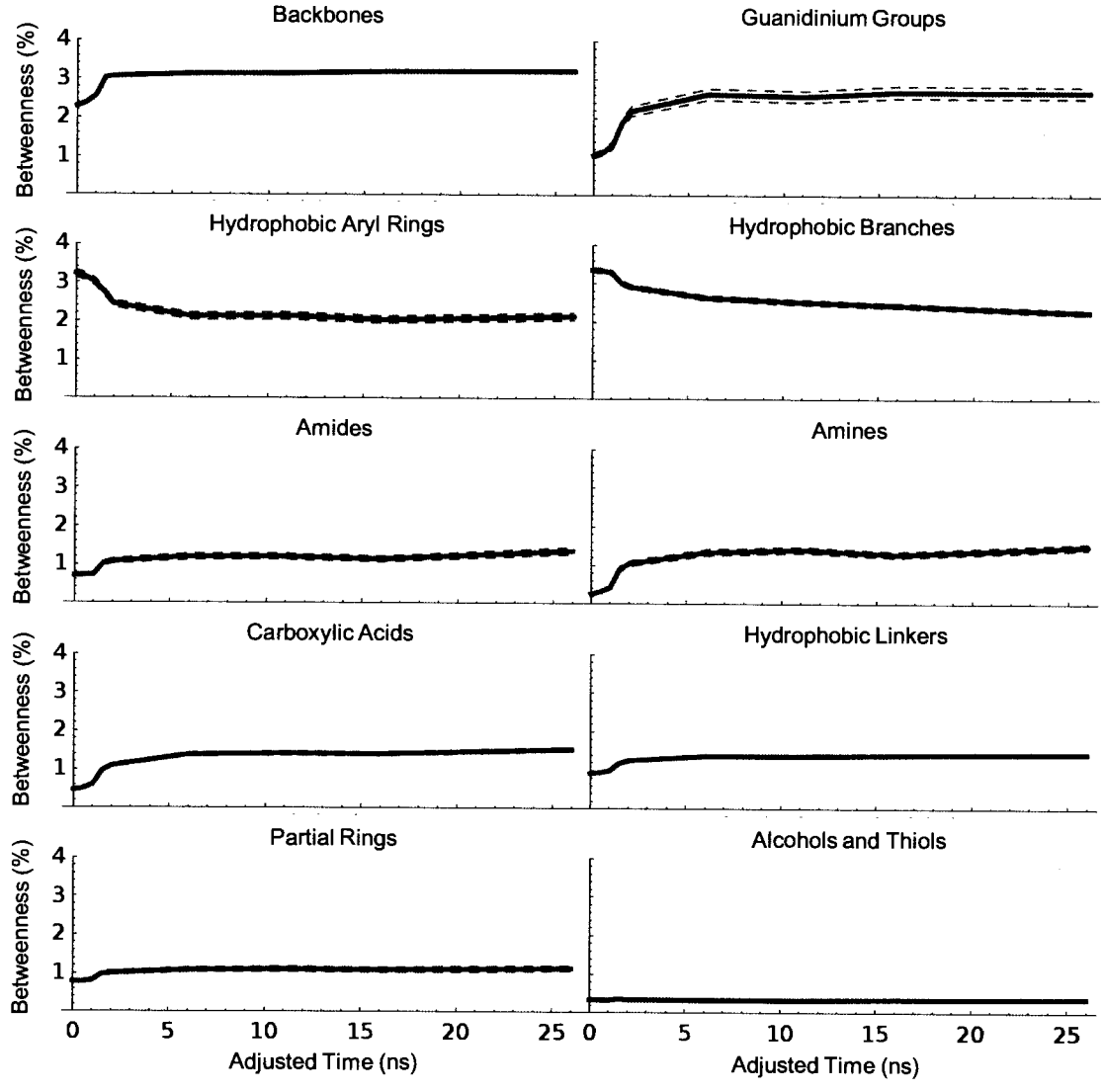


Figure 6.6: Mean betweenness centrality over the course of the 498 K simulation by node category. The mean ± standard error is plotted with dashed lines. Time is adjusted so that the transition state of each simulation occurs at 1 ns; all other time points are relative to this with the exception of time point 0, which summarizes the native state. C-terminal and N-terminal charged groups are not shown due to low sampling.

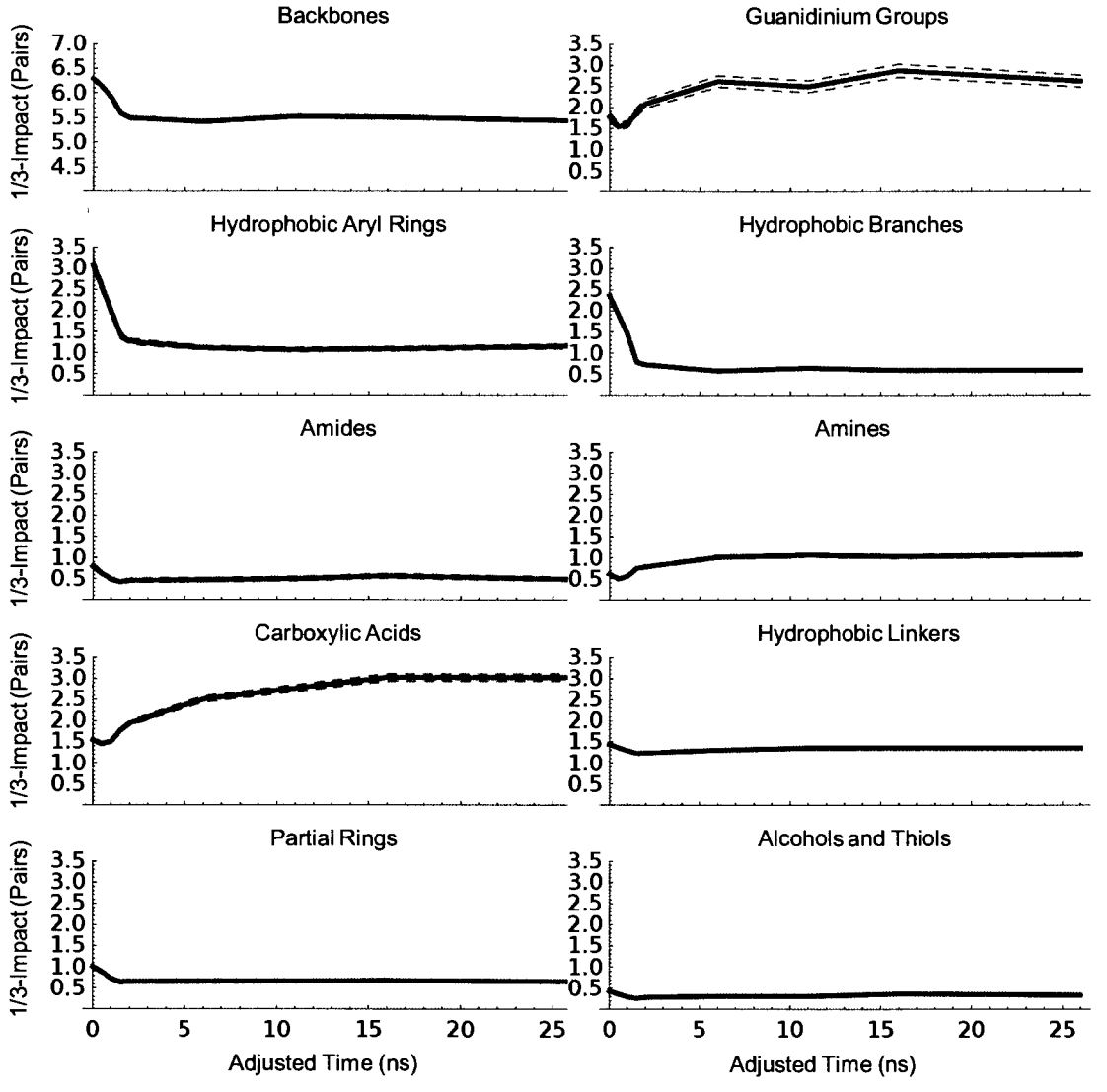


Figure 6.7: Mean 1/3-impact (in node pairs) over the course of the 498 K simulation by node category. The mean ± standard error is plotted with dashed lines. Time is adjusted so that the transition state of each simulation occurs at 1 ns; all other time points are relative to this with the exception of time point 0, which summarizes the native state. C-terminal and N-terminal charged groups are not shown due to low sampling.

Figure 6.8: Unfolding trajectory of barnase. Snapshots are taken at **(a)** 0 ns, **(b)** 0.1 ns, **(c)** the TS (0.235 ns), **(d)** 0.8 ns, **(e)** 1.2 ns, and **(f)** 2.2 ns. Residue Y78 is shown in dark colors in each frame; in all but the last frame all residues within 10 Å of Y78 are shown in light colors. During the time up to the TS, Y78 remains in contact with a large number of hydrophobic side-chains. By the TS, it has lost many of these contacts but maintains its betweenness by connecting two disparate parts of the protein through a contact with I51. This contact is maintained for some time after the TS, during which time charged groups nearby (e. g. K27 and D75) begin to enter the protein core. By 1 ns after the TS, Y78 has been ejected and no longer has a high betweenness. By 2ns after the TS, Y78 has a betweenness of less than 0.01, but a charged Lys (K27) can be seen connecting regions of the protein via a contact with the backbone of I51.

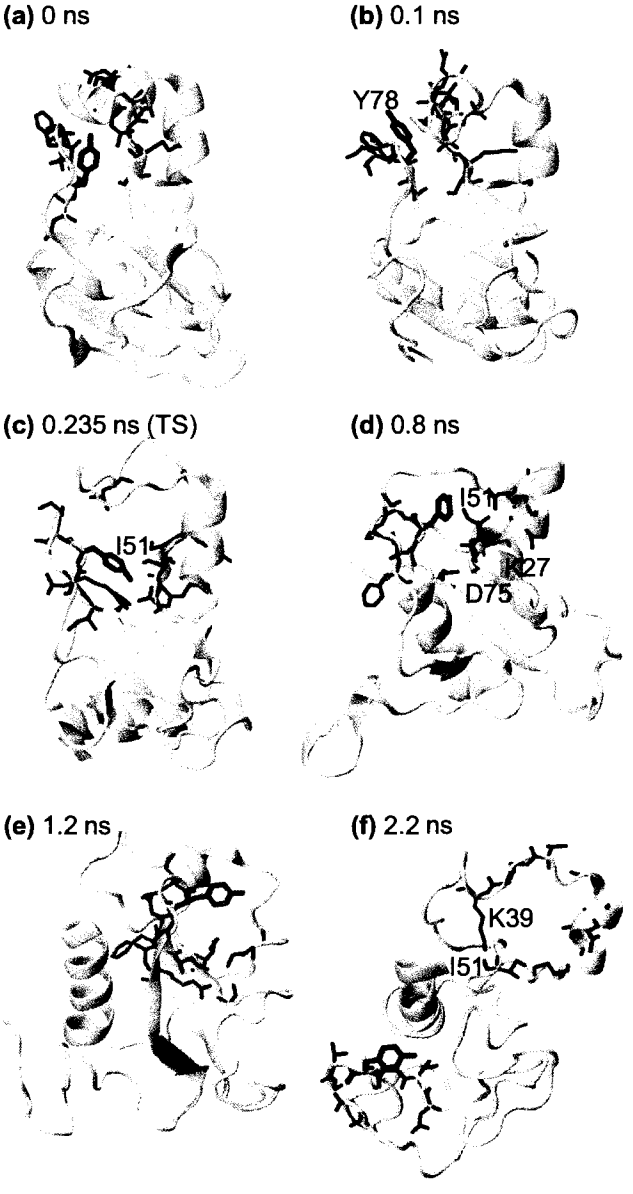
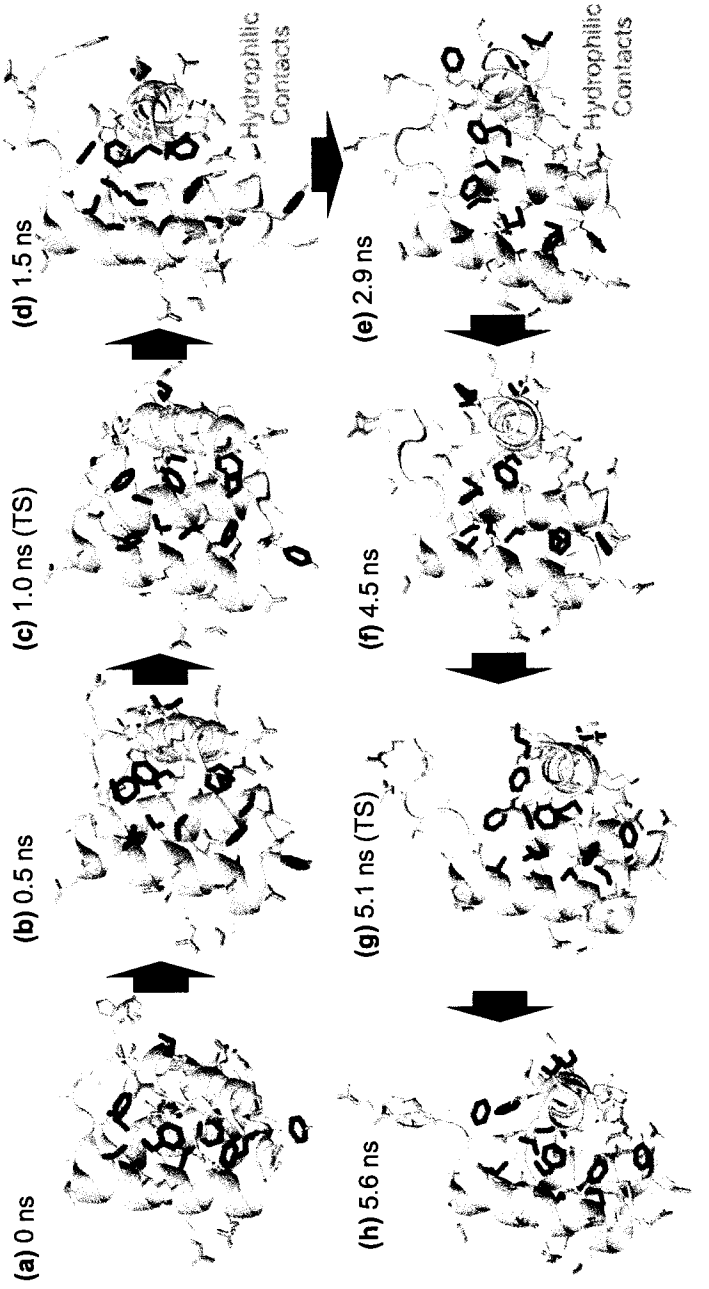

(a) 0 ns
(b) 0.1 ns
Y78
(c) 0.235 ns (TS)
I51
(d) 0.8 ns
I51
K27
D75
(e) 1.2 ns
(f) 2.2 ns
K39
I51

Figure 6.9: The unfolding and refolding pathway of the engrailed homeodomain (*1enh*) with hydrophobic branch and aryl groups colored purple and charged groups colored orange. The unfolding TS occurs at **(c)** 1.0 ns, and the refolding TS occurs at **(h)** 5.1 ns. The most unfolded state occurs at **(e)** 2.9 ns. **(a)** The native state has a clear hydrophobic core through the middle of the protein. **(b)** 0.5 ns prior to the TS, the hydrophobic core has begun to expand, but betweenness is maintained through a loosened contact network. **(c)** At the transition state, the protein core is expanded, and hydrophilic residues that can be seen drawing together. **(d)** The hydrophilic residues form a cluster of contacts at 1.5 ns, drastically increasing their betweenness, and the hydrophobic core begins to push apart, with only hydrophobic aryl groups maintaining significant hydrophobic contacts. **(e)** At the most unfolded point in the simulation, the hydrophobic residues are spread evenly throughout the protein while many hydrophilic residues remain in contact. **(f)** In the reverse of unfolding, the hydrophobic aryl rings draw back together early on; the hydrophilic residues remain in contact. **(g)** The refolding transition state sees the reformation of the hydrophobic core, eliminating most betweenness of the hydrophilic residues. **(h)** The reverse expansion state sees a slightly contracted hydrophobic core.


(a) 0 ns
(b) 0.5 ns
(c) 1.0 ns (TS)
(d) 1.5 ns
Hydrophilic Contacts
(e) 2.9 ns
Hydrophilic Contacts
(f) 4.5 ns
(g) 5.1 ns (TS)
(h) 5.6 ns

Chapter 7

A COMPARISON OF METHODS FOR THE ANALYSIS OF MOLECULAR DYNAMICS SIMULATIONS

7.1 Summary

Molecular dynamics (MD) is the only technique available for obtaining dynamic protein data at atomic spacial resolution and picosecond or finer temporal resolution. In recent years, decreases in the cost of computational resources have lead to an increase in the use of MD in biological research, both to examine phenomena that cannot be resolved experimentally and to generate hypotheses that direct further experimental research; in fact, several databases of MD simulations have arisen in recent years. MD simulations, and especially MD simulation databases, can contain massive amounts of data, and analysis of this data can be a daunting task. Here we compare several MD analysis methods to show their strengths and weaknesses using the wild-type and R282W mutant forms of the DNA-binding domain of protein p53. Our analyses indicate that the R282W mutation of p53 destabilizes the L1 Loop and loosens the H2 Helix conformation but that the loosened L1 Loop can be rescued by residue H115, preventing the R282W mutation from completely destabilizing the protein or abolishing activity.

7.2 Introduction

The interpretation and analysis of molecular dynamics (MD) simulations can be a difficult task. Choosing the wrong analysis technique to test a hypothesis will result in wasted time and inconclusive results; choosing the correct

analysis technique, on the other hand, requires a working knowledge of both the hypothesis to be tested and the strengths and weaknesses of the many techniques available. We compare a wide array of analysis methods applied to three simulations each of the protein 53 (p53) wild-type (wt) and R282W mutant proteins.

7.2.1 Model System

P53 is a cell cycle regulator that functions as a tumor suppressor by activating DNA repair, pausing growth during DNA repair, and inducing apoptosis if DNA is sufficiently damaged (see, for example, Strachan and Read (1999, chap. 18)). P53 consists of seven domains including a core DNA-binding domain (DBD; residues 100-300). Mutations in the P53 gene are the most commonly found mutations in human tumor cells, with the DBD accounting for most of these cases (Olivier et al., 2002; Hamroun et al., 2006). Additionally, it has been shown that the type of mutation is linked to prognosis and has implications for treatment (Olivier et al., 2006).

One of the many dangerous p53 mutations is the R282W mutation, which is among the 5 most common p53 mutations (Joerger and Fersht, 2007). R282W is in the periphery of the DNA-binding surface on the C-terminal helix (H2) and is known to disrupt the hydrogen bonding network of the local turn-sheet-helix motif while leaving the overall structure undisturbed (Joerger et al., 2006). The minimized crystal structure the wt p53 DBD is shown in Figure 7.1 with the R282W mutant indicated in red. The large guanidinium group of R282 forms hydrogen bonds that connect the loop and turn supporting H2; these bonds are lost with the addition of the Trp residue, distorting the L1 loop slightly, including DNA-binding residue K120. The overall DNA-binding domain is maintained, however, and the mutant is active at low levels.

Interestingly, the p53 protein is unusually unstable and melts at only

slightly above body temperature (Bullock et al., 2000; Ang et al., 2006). It has been hypothesized that this instability is linked with its unusually high flexibility (Cañadillas et al., 2006), specifically of the L1 and S7-S8 loops (Fig. 7.1, which may allow p53 to perform its many diverse functions. The R282W mutation is known to decrease p53's stability further by 3 kcal/mol (Bullock et al., 2000). The polar Arg is involved in packing the H2 helix against the S2-S2' β-turn, which is slightly disrupted in crystal structures with the Trp mutant. The R282W mutation does not disrupt the overall structure of the DNA binding region, however.

7.2.2 Traditional Analyses

The most traditional MD analysis techniques include the root mean square deviation (RMSD) and the root mean square fluctuation (RMSF). Both of these are measurements of the distance of an atom or set of atoms from a specific reference over time. Each is expressed by Equation 7.1, which describes the RMS value D in terms of a reference structure m and a trajectory structure q, each of which is an $n \times 3$ matrix containing 3D atomic coordinates (row vectors $m_1, m_2, ...m_n$) for n atoms. In the case of RMSD, the reference matrix m is usually the initial structure of the simulation or a minimized version of the crystal structure. For RMSF, the reference is the mean structure.

$$D(\mathbf{m}, \mathbf{q}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||m_i - q_i||^2} \tag{7.1}$$

The conformational genealogy (Congeneal) (Yee and Dill, 1993) is a measurement of structural relatedness similar to an entire-protein RMSD but which examines the inter-atomic distances of the Cα atoms in a protein. The Congeneal distance between two structures of equal length is the Bray-Curtis

distance between their weighted distance matrices. For two proteins of unequal length, n and m such that $n < m$, it is the minimum Bray-Curtis distance between the smaller protein's weighted distance matrix and all the contiguous $n \times n$ sub-matrices of the larger protein's weighted distance matrix. Congeneal is distinct from RMSD that it can be used on proteins of unequal size and that its weighting can cause local structural features to play a larger role in the measurement, eliminating low scores due to floppy tails or loops.

Another traditional simulation analyses is the solvent accessible surface area (SASA) (Lee and Richards, 1971), which measures the surface area of a molecule that is accessible to the solvent, generally water. The calculation of this measurement is beyond the scope of this paper but algorithms are discussed in detail by Shrake and Rupley (1973) and Weiser et al. (1999). SASA is frequently calculated for individual residues and compared over the course of a simulation to observer solvent exposure events.

Protein dihedral or torsion angle analyses are also common in MD literature. Dihedral analyses examine rotational angles throughout the protein structure, especially the (Φ, Ψ) angles along the protein backbone, which can be used to identify secondary structure arrangements. The definition of protein secondary structure (DSSP) (Kabsch and Sander, 1983) also identifies secondary structure patterns in proteins by examining hydrogen bond patterns based on a purely electrostatic definition. Both of these methods can be used to identify secondary structure; the major difference between them is the focus of DSSP on hydrogen bonds compared to (Φ, Ψ)'s focus on rotational angles. Although both methods have slight biases, we focus only on DSSP due to their overall similarity.

A final class of traditional analyses are contact-based analysis. These tend to fall into two categories: fine detail structural analysis (FDSA) and contact maps. FDSA generally consists of examining trajectories on the basis of inter-

atomic contacts over time. For example, one could examine all the contacts made by a single Cγ atom of a Val residue. While this method can be extremely valuable, it is a reorganization rather than a summary of the initial simulation with only marginally less entropy. In other words, without a firm idea of what one is looking for, examining an FDSA map is little different than examining a movie of a protein simulation; thus we do not discuss it further. Contact maps or matrices, on the other hand, summarize the frequency with which any two residues are in contact over the course of the simulation, often comparing the simulation's contacts to those of a crystal structure. Such maps can be quite useful for quickly determining the major changes that have occurred over the course of a simulation.

7.2.3 Flexibility

Flexibility is an analysis, related to principal component analysis, which is applied to each atom in a simulation individually (Teodoro et al., 2003). This allows one to immediately determine and summarize the major modes of each atom of the protein over the course of the simulation while filtering out the less significant fast vibrations. Flexibility analysis has only recently been applied to MD trajectories, but it has been used to examine large datasets of protein simulations (Benson and Daggett, 2008). This study identified basic features of protein flexibility and demonstrated that proteins in the same fold families tend to have similar flexibility patterns. It also identified a number of unusually inflexible loops with structural properties mirroring traditional secondary structure and demonstrated that the most flexible sites of a protein at room temperature predict the early thermal unfolding trajectory of a protein.

Flexibility is usually visualized as either a set of displacement vectors plotted on the mean structure from a protein trajectory or as a similar set of vectors plotted onto a median structure, which is simply the protein structure occur-

ring in the simulation that has the lowest RMSD to the mean structure. When the mean structure is used, the vectors plotted are equal but opposite and represent the principal axis of the atom's motion scaled by the standard deviation of the atom along that axis. Mean structures are not always physically realistic structures, however. Because of this, we use the median and plot arrows from the atom's position in the median structure to the ends of its principal axis as measured from the mean position so as to preserve all data from the flexibility calculation.

7.2.4 Wavelet Analysis

The continuous wavelet transform is a technique that has been widely used in fields such as meteorology (Meyers et al., 1993; Torrence and Compo, 1998) and that has been previously suggested as a tool for the analysis of MD (Askar et al., 1996). Until recently, however, only discrete versions of the wavelet transform had been applied to MD (reviewed by Liò (2003)). Recently, the continuous wavelet transform, which we will refer to simply as wavelet analysis, has been shown to be quite useful in MD research due to its ability to quickly locate regions in both time and space during which nonrandom motions are occurring (Benson and Daggett, 2010a). Wavelet analysis is performed, for a single atom, by searching for instances in its trajectory over time at which its motion is similar to that of a particular wavelet function. These wavelet functions can be stretched and compressed to identify different sizes and shapes of motion in the trajectory. In this paper, we will use the Paul wavelet function (Addison et al., 2002), which excels at detecting sigmoid as well as oscillatory motions. A complete description of the calculations involved in wavelet analysis is beyond the scope of this paper, but a practical guide is given by Torrence and Compo (1998) and a guide to the application of wavelet analysis to MD, including sample codes, is given by Benson and Daggett (2010a). More detailed theoretical

treatments can be found elsewhere as well (Daubechies, 1992; Meyers et al., 1993; Addison et al., 2002).

One interesting feature of wavelet analysis is that it lends itself to easy comparison between sets of simulations of a protein or between simulations of variants of a protein. Each region of time that matches a particular wavelet shape matches can be assigned a p-value measuring its significance compared to random noise. These p-values can be combined across simulations to determine, within statistical significance, if the propensity of an atom to move in certain ways is different between protein variants. This technique is illustrated by Benson et al. (2010).

7.2.5 Graph Theory

Graph theoretic techniques involve simplifying a protein structure into a mathematically tractable and discrete representation called a graph. Graphs are simply collections of nodes (or vertices) connected to each other by edges. Generally, nodes represent physical pieces of the protein such as individual residues while edges represent relationships such as closeness in space or contacts. Labels can be given to nodes (e.g., residue type) and to edges (e.g., distance or number of pairs of atoms in contact) to allow the graphs to capture more information. Graphs have the immediate advantage that, while they can capture much of the critical information about a protein structure, they are computationally easier to manage than coordinates and support a large array of easily calculated and well studied mathematical metrics.

Graph theoretic approaches to analyzing proteins have made various appearances in the literature, but few have been applied to MD trajectories. One rudimentary approach to analyzing a simulation with graphs is simply to visualize the graphs. This technique has shown some promise in examining structural differences in simulations of dimers (Swint-Kruse, 2004), where a simple

graph of the protein contacts was sufficient to identify interesting differences between monomeric contacts. A similar approach has been used to examine single nucleotide polymorphism (SNP) variants of a protein. Schmidlin et al. (2009) plotted and examined contacts between residues in superoxide dismutase that differed by a certain threshold in the wt and mutant simulations in order to identify changes in the contact network.

Recently, protein structure graph analysis has been applied to MD trajectories by breaking a protein into nodes of chemically similar atoms close in space (Benson and Daggett, 2010b). The graph's edges, which were formed by contacts at each frame of the simulation according to Delaunay tessellation (Delaunay, 1934) and a distance cutoff, were then smoothed in time with a Gaussian. These smoothed edges had weights at every time-point in the range of $[0, 1]$, which represented the probability of two nodes being in contact near that time. These edge weights were used to characterize the chemical environment of individual nodes by counting the number of contacts at different probability levels; these counds form a vector that serves as a dynamic environment index. These indices could be used either as a method of searching for nodes with similar chemical environments or to detect sudden changes in a node's environment. Additionally, Benson and Daggett (2010b) propose a distance metric, graph communication, C, which is given in Equation 7.2 where $S_{a,b}$ is the set of edges on the shortest path from node a to node b and where $w(u, v)$ is the weight (probability of contact) of two nodes, u and v.

$$C(a, b) = \sum_{(u,v) \in S_{a,b}} \frac{1}{w(u, v)} \tag{7.2}$$

This communication distance measures the propensity for two nodes to influence each other either through direct contact or via an intermediate. A communication distance of 1 would indicate that two graph nodes were constantly

in contact while a communication distance of 4 could indicate either that two nodes were in contact $\sim 25\%$ of the time or that they were each in contact with another occluding node $\sim 50\%$ of the time. In this fashion, graph communication can capture not only loss of contact but also intermediate rearrangements that decrease communication between otherwise disconnected nodes.

7.3 Methods

7.3.1 Protein Preparation and Simulation

Simulations were based on the 2.05 Å resolution crystal structure of the DNA-binding domain (residues 96-289) of p53 (Wang et al., 2007), PDB code *2ocj*. The R282W mutation was prepared by substitution to the wt structure and energy minimization in vacuo using the ENCAD package (Levitt, 1990). Minimization was performed using the Levitt et al. (1995) force field for 1000 steps of steepest decent minimization. These structures were solvated in a rectangular box with walls ≥ 10 Å from any protein atom with a solvent density of 0.933 g/mL, the experimental density of water at 310 K and 1 atm pressure (Kell, 1967). Solvent was additionally minimized for 1000 steps followed by 1 ps of dynamics of the solvent only and 500 more steps of solvent minimization. Following this minimization, the entire system was heated for 310 K. Simulations were performed using our in-house simulation package, in lucem molecular mechanics (*il*mm) (Beck et al., 2008) using the Levitt et al. (1995) force field and explicit three-centered flexible water molecules (Levitt et al., 1997). Three simulations of each variant (wt and R282W) protein were run at 310 K for at least 21 ns each with different random number seeds used during the assignment of initial velocities. Simulations included all hydrogen atoms and used a force-shifted non-bonded cutoff of 10 Å. The time step used was 2 fs with coordinates saved every 1 ps. Further simulation details are given elsewhere (Levitt et al.,

1995; Beck and Daggett, 2004).

7.3.2 Analysis

All analyses were performed using *il*mm. RMSD, RMSF, SASA, DSSP, contacts, correlated motion, flexibility, wavelets, and graphs were calculated. RMSD, RMSF, correlated motion, flexibility, and wavelets were calculated following the removal of rotation and translation from the system using a rigid least squares fit (Kearsley, 1989). SASA was performed using the NACCESS algorithm (Hubbard and Thornton, 1993). DSSP was calculated using the DSSP algorithm (Kabsch and Sander, 1983). Correlated motion was taken to be the average of the correlation of two atoms in the x, y, and z directions. Contacts were calculated using a C-C atom distance cutoff of 5.4 Å and a heavy-atom (C, O, N, S) distance cutoff of 4.6 Å for non-adjacent residues. Flexibility was calculated using the method outlined by Teodoro et al. (2003) and Benson and Daggett (2008). Wavelets were calculated for all Cα atoms and significance was evaluated using the noise distribution described by Benson and Daggett (2010a). Wavelet motions in the top 20% of this noise-distribution were considered significant. Graphs were generated using the node definitions described by Benson and Daggett (2010b); two nodes were considered in contact (and were linked by a contact edge) when they were within 4.6 Å and were not occluded by another atom, as determined by Delaunay tessellation (Delaunay, 1934). All analyses were performed on the first 21 ns of each simulation of both wt and R282 proteins. All plots were produced using Mathematica (Wolfram Research, 2008), and protein images were produced using Visual Molecular Dynamics (VMD) (Humphrey et al., 1996).

7.4 Results and Comparison of Analyses

Of all the analyses performed here, RMSD and RMSF are the most similar with the important distinction that RMSD shows the deviation from the minimized crystal structure while RMSF shows the deviation from the mean structure. In this sense, RMSF gives a picture of what parts of the protein are moving at any given time while RMSD gives an overall picture of how much each part of the protein has changed so far at a given time. RMSF and RMSD plots for all simulations are shown in Figures 7.2 and 7.3 respectively.

Both RMSF and RMSD tend to stay consistent per residue across each simulation and tend to be highly related between simulations. In fact, the lowest correlation of RMSF between any pair of simulations is 0.61 (between wt simulation 2 and R282W simulation 2) while the highest pairwise correlation is 0.75 (between wt simulation 1 and R282W simulation 1). This correlation is, in fact, higher than any wt-wt or mutant-mutant correlation. For RMSD, the *lowest* pairwise correlation is between wt simulation 1 and wt simulation 3 ($R = 0.31$) while the highest is between wt simulation 1 and R282W simulation 3 ($R = 0.63$).

Flexibility has a slight relationship to RMSF as well in that it measures the amount of movement along a principal axis where RMSF measures the amount of movement generally. Flexibility additionally adds a directional component to the standard RMSF values (Fig. 7.4). When examining the flexibility plots, it is immediately obvious that the L1 Loop is considerably more flexible in the wt simulations than in the R282W simulations, especially wt simulation 2, while the H2 Helix is both more flexible and more displaced on average from the S2-S2' β-sheet in the R282W simulations. Other regions show more subtle or indistinguishable differences, though there is a slightly higher flexibility of the S7-S8 loop in the wt, especially simulation 1.

The DSSP analysis gives a very clear picture of the loss and gain of secondary structure (as determined by H-bond patterns) throughout the simulation (Fig. 7.5). Most secondary structure elements are stable throughout both wt and R282W simulations, specifically S2, S2', S3, S4, S6, S7, S8, S9, S10, and H2. S1 is somewhat inconsistent in all simulations, while the L2 loop forms some helical character in wt simulations 3 and R282W simulation 1. Notably, the H1 helix is consistent in the wt simulations but inconsistent in two of the R282W simulations. S5 is consistent everywhere but in wt simulation 3.

Both SASA (Fig. 7.6) and contact (Fig. 7.7) analyses are difficult to interpret for p53 due the the lack of changes in them between any two simulations generally. SASA is nearly universally consistent throughout the simulations with deviations too small to be visible while contact maps are nearly indistinguishable from each other without much more detailed analysis. Correlated motion (Fig. 7.8), on the other hand, is difficult to compare due to the lack of similarity between any two plots. The highest correlation between correlated motion values for a pair of simulations occurs between wt simulation 3 and R282W simulation 2 ($R = 0.66$). In fact, the similarity in correlated motion between wt and mutant is generally higher than between wt and wt.

Wavelet analysis (Fig. 7.9) suggests a great deal of ordered low-frequency motion in the wt compared to the R282W simulations. Wavelet maps would immediately suggest significant motions throughout simulation 1 of the wt between 5 and 10 ns and near residues 120 and 280 (L1 loop and H2 Helix respectively) in wt simulation 2 from 5 to 15 ns. In simulation 3 of the wt, some motion is seen near residue 240 (L3 loop) from 5-15 ns. The R282W simulations show fewer significant motions according to wavelet analysis, but simulation 1 shows motion near residue 180 (H1 Helix) from 7 to 15 ns while simulation 3 shows motion scattered throughout the 5-10 ns range, especially in the N-terminus. The greatest differences in ordered motion, according to wavelet analysis, are

shown in Figure 7.10. The regions with the most significant differences include the polymorphic site (H2, L1 tip, S1), the L2 and L3 Loops, and Strand S8 and its neighbors, S3 and the S5-S6 Loop.

Graph analysis (Fig. 7.11) shows a great deal of consistency in terms of the overall communication of residue 282 with a few anomalies present. Notably, the connection between the polymorphic residue and the region near residue 125 (L1 and S2) is slightly weaker in the R282W simulations than in the wt. This is especially true of R282W simulation 2, which also shows a slightly increased closeness to residue 175 (Helix H1), increased distance from the region following residue 175 (H1-S5 Loop), and the L3 Loop near residue 240. Additionally, there is a slightly increased distance in the R282W simulations from the polymorphic residue to the S7-S8 Loop, most notably in simulation 2. Simulation 2 of the R282W variant also shows two faint horizontal bands between 3 and 10 ns during which the W282 side-chain has closer communication with the H1-S5 Loop, Helix H1, and the L3 Loop. Examination of simulation 2 shows that the larger Trp side-chain fails to hold Helix H2 near the L1 Loop, allowing it to shift toward the L3 loop. This results in a significant loss of communication with the L1 Loop and an increase in communication with the L3 Loop. Additionally, the L3 Loop swings toward Helix H2 during the window from 3-10 ns, further decreasing the communication distance between them and facilitating communication with Helix H1.

Interestingly, the RMSD of simulation 1 of the wt shows a sudden jump for several residues at 10 ns (Fig. 7.3) that is not present in RMSF except as a faint horizontal line (Fig. 7.2). In fact, no other analysis, with the exception of wavelet analysis, shows a significant change occurring at 10 ns. Wavelets show considerable low-frequency motion from 5-10 ns, all of which stops at 10ns (Fig. 7.9). Visual inspection of the trajectory reveals that a considerable amount of motion occurs between 5-10 ns in several regions of the protein,

none of which is large in and of itself but all of which add up together to cause a shift in the protein's alignment to the crystal structure, leading to a jump in RMSD following the 10 ns mark. These changes are captured by wavelet analysis, specifically near residues residue 250 (L3 Loop), 225 (S7-S8 Loop, Fig. 7.12a), 280 (Helix H2, Fig. 7.12b), 175 (H1 Helix and surrounding loops, Fig. 7.12c), and most of the N-terminus. In contrast, the region near residue 200 (S5-S6 Loop) shows no significant motion during this time region according to wavelet analysis, and is stable in the simulation as well (Fig. 7.12d), despite being solvent exposed. Notably, the overall structure of the protein is well maintained throughout this time, but individual regions shift considerably.

The L1 Loop shows very dramatic differences between wt and R282W simulations in flexibility analysis (Fig. 7.4 but relatively little difference in the RMSF (Fig. 7.2), with the R282W L1 Loop appearing to be slightly more mobile than that of the wt in RMSF analysis. It is clear from the flexibility analysis that the loop's flexibility is decreased along its principal axis in the mutant simulations. Notably, flexibility also shows that Loop L1 and Helix H2 are displaced on average in the R282W compared to the wt simulations. In fact, during wt simulations, the R282 side-chain interacts frequently with the polar backbone atoms of the L1 loop, holding its base and Helix H2 close together but allowing its tip to oscillate considerably (Fig. 7.13a). In the R282W simulations, however, this interaction is lost, thus L1 swings widely out into solvent in the first few ns of simulation (Fig. 7.13b). Interestingly, residue H115 appears to rescue L1 from being entirely disorganized by interacting with residues of Strand S2 in the R282W simulations. Although it moves a similar amount in the mutant, it does so in a less systematic fashion, leading to lower flexibility along its principal axis.

7.5 Discussion

Each analysis method examined here has considerable merit for uncovering specific types of events in MD simulations. When used properly, they have the ability to greatly decrease the amount of time and energy required to understand an MD simulation while simultaneously quantifying otherwise qualitative observations. In the case of the R282W mutation of p53, several of the analyses were extremely useful in characterizing the effects of the larger Trp side-chain on the rest of the protein.

7.5.1 Analyses

Of the 9 analysis methods compared in this paper, the least useful on this particular system were SASA, correlated motion, and contact analysis. This is largely due to the nature of the system, however. The p53 wt and R282W simulations were quite stable and contained no major opening or closing motions that would have caused a large change in SASA. This is, in and of itself, a useful observation that the SASA data (Fig. 7.6) support. SASA can be a very noisy measurement, however; thus it is frequently of greater use in quantifying hypotheses generated from other methods than it is scanning simulations and forming hypotheses. Contact analysis (Fig. 7.7) has a similar use to SASA for this particular system. Because the rearrangements that occurred in the protein were slight, contact analysis showed very few noticeable differences between simulations. This suggests that the protein maintains its overall structure well, with few events propagating beyond the local region of the mutation. However, it is worth noting that even in the regions we should most expect to see differences in contacts–the interface of H2 (near polymorphic residue 282) and Loop L1 (near residue 120), there are few distinguishable differences between simulations. This is likely due to the fact that the swapping of a Trp

for the positive Arg group, which results in a weaker connection between the L1 Loop and the H2 helix, did not eliminate the contact altogether but rather decreased its frequency. Thus, the slight change in contact propensities was overall too slight to be visible on a traditional contact map.

Graph communication analysis is related to contacts in that it uses the network of contacts in a protein to calculate a communication score that measures the distance between two parts of the protein in terms of influence. This communication propensity can be very sensitive to small changes, such as those seen in simulation 2 of the R282W variant (Fig. 7.11, due to the fact that a reduction in the probability of a single central contact can increase the communication distance of any pair of residues that are near it. Because the R282 residue was the primary link between the H2 Helix to the L1 Loop, which additionally is the primary link between regions of the protein such as the S7-S8 Loop, a decrease in the probability of this contact by the substitution of the Trp side-chain leads to an increase in the communication distance between residue 282 and the L1 and S7-S8 Loops.

The relative instability of correlated motion (Fig. 7.8) across simulations suggests that correlated motion would be more useful in a hypothesis evaluation as well. This is not particularly surprising considering that it is rare for two regions of a protein to have truly correlated or anti-correlated motion over a long period of time. The correlated motion during a small window of time, for example during a period in which an enzyme's active site is exposed, is much more likely to be useful than the correlated motion of an entire simulation, during which time the thermal vibrations of each atom tend to wash out the larger motions.

The RMSF (Fig. 7.2 and RMSD (Fig. 7.3) are useful measures for each simulation, as they tend to show differences clearly without being washed out by similarities (as is the case with SASA and contacts, for example). Although

the similarities across simulations are far more obvious than the differences via RMSF and RMSD, some differences are visible. Not surprisingly, the two measures are closely related, with RMSF often showing motion immediately before jumps in RMSD, as is the case with wt simulation 1 at 10 ns. In this fashion, both RMSD and RMSF do a good job of highlighting individual parts of a simulation that should be examined.

The most powerful technique for pointing out interesting regions of time for specific residues, however, is wavelet analysis (Fig. 7.9). Wavelet analysis pinpoints specific residues that are undergoing significant motion at precise periods of time. In the case of the motion near 10 ns in the wt simulation 1, for example, wavelet correctly identified the fact that the motion began occurring near 5 ns and finished near 10 ns (Fig. 7.12), which is not clear from either RMSF or RMSD analysis. Additionally, wavelet analysis pinpointed significant events in each simulation, many of which were directly related to the polymorphic change. The primary drawback of this type of wavelet analysis is that it does not directly compare simulations; one must inspect the events highlighted by wavelet analysis via other means. It can, however, greatly decrease the amount of energy required to identify significant events in a simulation.

Comparisons of wavelet analysis (Fig. 7.10) showed several expected results as well as a few unexpected ones. The significant differences in motion near the polymorphic site are expected, even if the polymorphic site is highly mobile in both variants, due to the changes in conformation that occur in the R282W. The many changes directly near the polymorphic site (S1, S3, L3, L2) are relatively unsurprising as well, since the polymorphic site has significant communication with these sites. The changes in the S8 Strand (residues 228-237), however, were surprising. A close examination of DSSP (Fig. 7.5) shows that S8 has a slight tendency to lose hydrogen bonds near in the wt, specifically simulations 1 and 3. Flexibility (Fig. 7.4) additionally shows that the conformation of S8 is

somewhat inconsistent in the wt simulations. In fact, this phenomenon seems to be caused by the packing of the S1 Strand, which packs more tightly agains the S2 Strand near the L1 loop in the wt simulations due to the closeness of the L1 loop to the S2 strand and the H1 Helix. This packing causes the S3 Strand to pack more closely as well, leaving extra space around the N-terminal end of Strand S8 in the wt simulations. In wt simulation 1, the S8 Strand packs tightly against S5 and the S5-S6 Loop, and, in all wt simulations, the N-terminal end of S8 changes conformation slightly to account for the shift in S3.

DSSP analysis (Fig. 7.5) can show quickly where secondary structure changes are occurring in a simulation. Because loss or gain of secondary structure is often associated with highly significant events in a trajectory, this analysis can be very valuable both for screening simulations and for hypothesis evaluation. In the case of p53, DSSP shows very clearly that the H1 Helical propensities for the R282W variant are lower than those for the wt. Additionally, it shows a significant decrease in the β-strand character of the N-terminus and L1 Loop. This is likely due to the extension and less ordered oscillation of the L1 Loop as identified by wavelet analysis and flexibility analysis.

Flexibility analysis (Fig. 7.4) is capable of summarizing an entire simulation immediately by showing both a typical structural conformation and the primary modes of that structure. Visual inspection of the flexibility plots for wt and R282W variants of p53 shows immediately that the L1 loop tends to remain closer to the H2 helix in the wt simulations, which agrees strongly with experimental results (Joerger et al., 2006). Surprisingly, however, flexibility analysis also indicates that the motion of the L1 Loop is greater along its principal axis in the wt than in the mutant simulations despite the higher RMSF of the L1 Loop in mutant simulations. Inspection of wt and R282W conformation (Fig. 7.13) indicates that this is due to the fact that the base of the L1 Loop is

held tightly in place by the positively charged R282 residue allowing the top of L1 to bend in an ordered manner. The W282 residue, however, does not hold the loop in place, allowing it to swing into solvent.

It is clear from flexibility analysis that the L1 Loop in wt simulation 2 has increased flexibility and a more distant conformation than that of the L1 Loop in simulations 1 and 3, though the conformation is not as disconnected from S2 as the conformations of the R282W simualtions. It is worth mentioning that the median structure in this simulation has an L1 conformation that is slightly farther from the mean structure than is typical, as indicated by the largely uni-directional flexibility arrows, which indirectly show the mean position. This high flexibility occurs because the L1 Loop is quite mobile during this simulation (as indicated by wavelet analysis, Fig. 7.9) and bends away from H2 around 12 ns (as indicated by a loss in graph communication, Fig. 7.11). A similar opening of the L1 Loop occurs near the beginning of simulation 1 followed by a closing event at 8.5 ns. These opening events occur in each of the R282W simulations, but without the R282 side-chain, they do not close.

7.5.2 Effects of the R282W Mutation

The overall effect of the R282W mutation, as observed in our simulations, agrees with experimental crystal structures (Joerger et al., 2006). The change from the positively charged Arg to the largely hydrophobic Trp causes the H1 Helix to disconnect from both the L1 Loop and the S2-S2' β-sheet. This leads to a significant rearrangement of active-site residues, including R280 in H1 and K120 in L1. The L1 loop, meanwhile, partially undocks from the S2 loop where R282 no longer holds it in place (Fig. 7.13).

Interestingly, the H115 residue seems to rescue the L1 Loop from becoming completely disordered in the R282W mutation by bonding to residues in the S2 Strand and holding the L1 Loop in place, albeit more loosely. This accounts

for the decreased but not abolished activity and stability of the R282W mutant and suggests that a double R282W and H115 mutant would be both less stable and less active. Interestingly, there are only two known non-silent mutations to H115, making it one of the most conserved residues in the DBD. One of these mutations is a deletion resulting in a stop at codon 116 and the other is a H115Y polymorphism (Olivier et al., 2002; Hamroun et al., 2006). The H115Y polymorphism has not been experimentally studied extensively, but one study did find that H115Y mutants of p53 lacked the ability to interfere with the protein p73, a protein with high sequence similarity to p53 also involved in transcription, while R282W p53 retained the ability to interfere with p73 (Monti et al., 2003). Because experimental results have shown that p53 is suspected to inhibit p73 via an interaction in the DBD and because a correlation exists between efficiency of p53 binding and p53's inhibition of p73 (Gaiddon et al., 2001), this suggests that the H115Y mutation may be more damaging to the L1 Loop than the R282W mutation.

One likely interpretation of these data is that the L1 Loop must stay near the H2 Helix and the S2 and S2' strands (in the loop-turn-helix motif) for p53 to remain active and stable. This is intuitive considering that the binding residue K120 is positioned at the tip of the L1 Loop and that significant loosening of the L1 Loop could easily destabilize the S1 Strand, exposing the hydrophobic core. However, our results indicate that slight displacement of the L1 Loop is acceptable, even in wt p53, so long as it does not lose complete contact with S2 and H2. These data fit well with NMR studies finding high flexibility in the L1 Loop (Cañadillas et al., 2006). It is likely, given this interpretation, that R282 is responsible for encouraging a binding-friendly structural arrangement in p53 but that H115 is responsible for keeping L1 from destabilizing the protein entirely, and that both residues work together to encourage the optimal structure.

7.6 Conclusions

Although all analyses examined here have appropriate and useful applications, we find that the combination of flexibility analysis and wavelet analysis to be an extremely powerful combination of course-grain and fine-detail analysis. This is due to the ability of flexibility ananlysis to summarize an entire simulation into a single image and the ability of wavelet analysis to highlight important events in a simulation. Using these tools as well as graph communication, DSSP, RMSF, and RMSD, we were able to show that p53 mutant R282W loosens the connections between the H2 Helix and L1 Loop, causing a rearrangment of binding residues. We also observed that the residue H115 at the base of the L1 Loop interacts with the residues in Strand S2, preventing the L1 Loop from completely losing its overall structure.

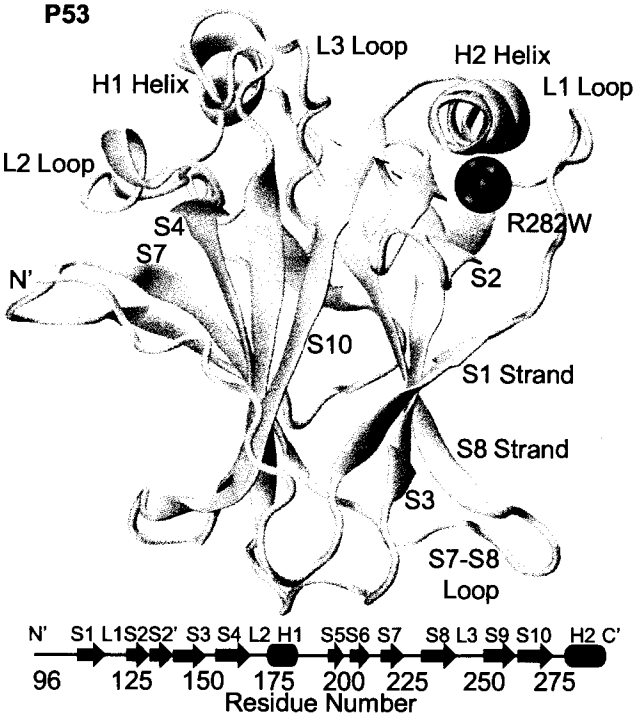


Figure 7.1: Minimized crystal structures and secondary structure map of DNA-binding domain of p53 (*2ocj*) with the polymorphic residue R282W indicated by a red sphere. DNA binding occurs on the upper surface with Helix H2 binding to the major groove. Loop L3 and Helix H1 also participate in binding and normally hold a Zinc ion.

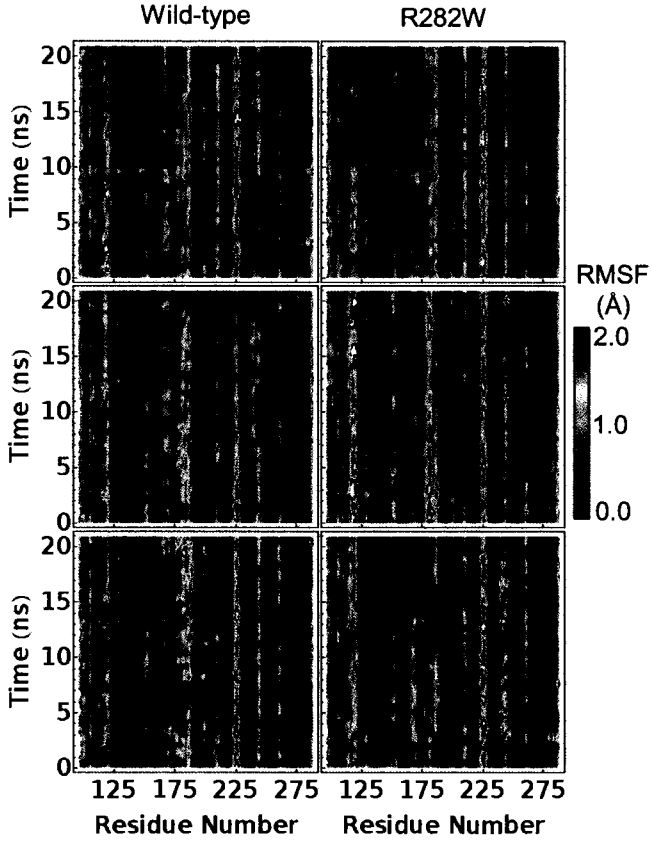


Figure 7.2: RMSF plots over time of each Cα atom of each of the wt and R282W mutant p53 simulations. Notably, a horizontal bar can be seen near 10 ns in the wt simulation 1 (top), while residue 120 has slightly higher fluctuations in the mutant.

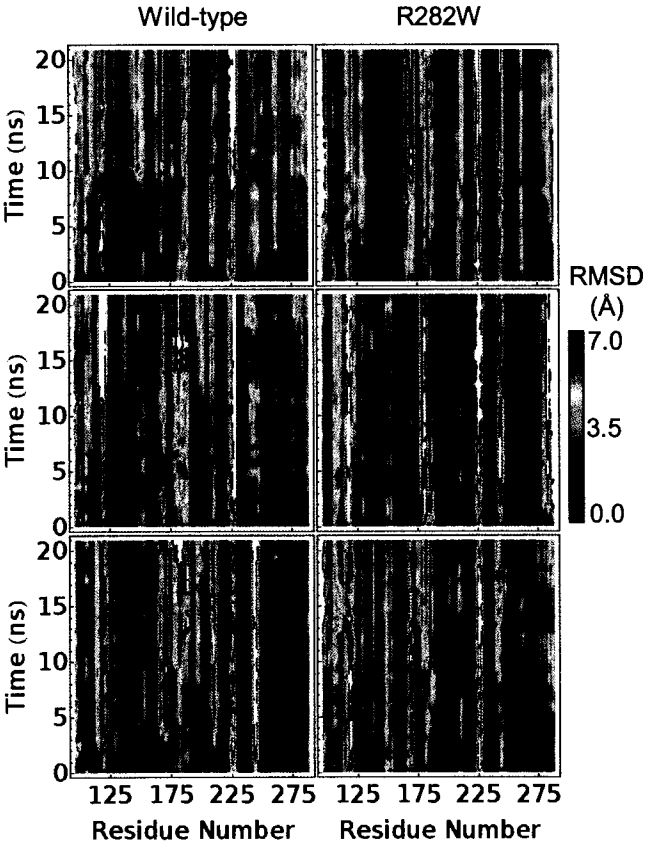


Figure 7.3: RMSD plots over time of all simulations of both wt and R282W p53. The C-terminus of the mutant can be observed to have a higher RMSD than the wt, especially in simulation 2 (center). In the wt simulation 1 (top), RMSD uniformly jumps at 10 ns.

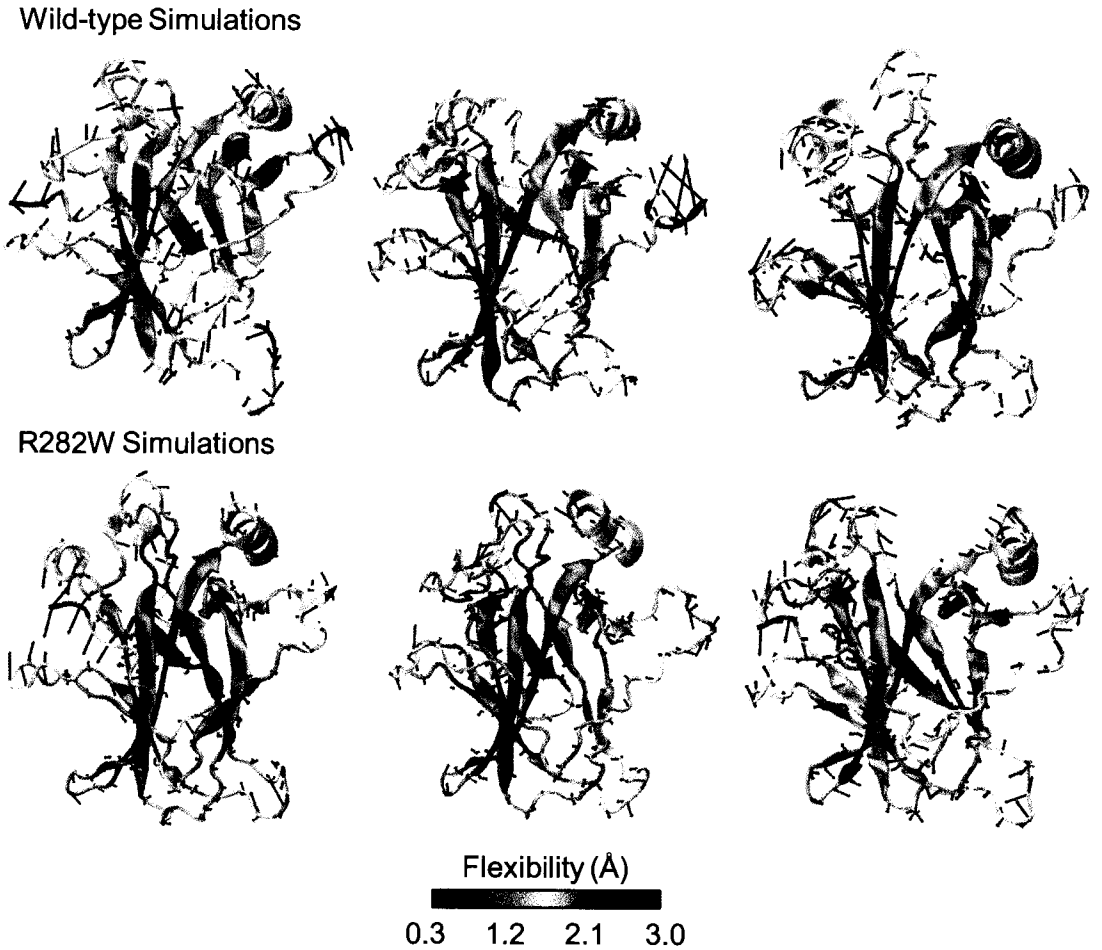


Figure 7.4: Flexibility plots of the each of the wt and R282W mutant simulations of p53. Notably, the L1 loop is considerably more flexible in the wt simulations, but the H2 helix is both more displaced and more flexible in the mutant simulations.

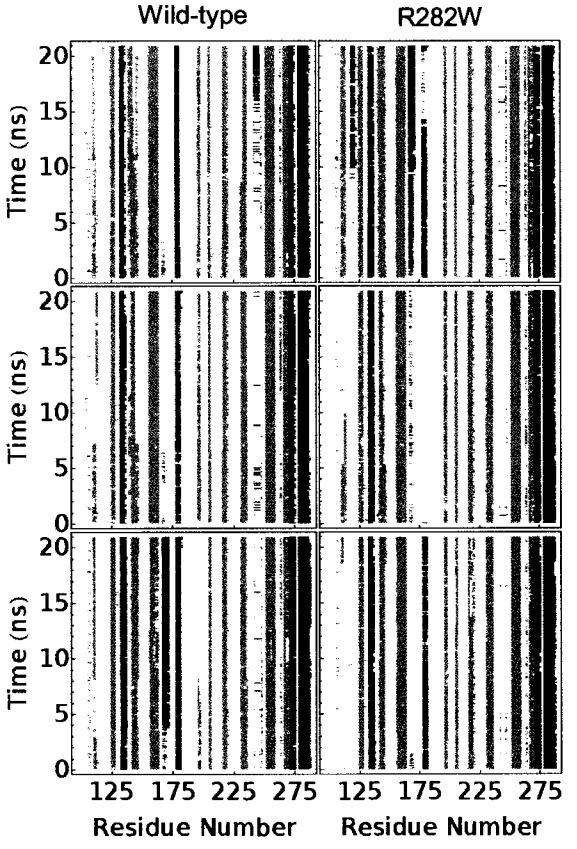


Figure 7.5: DSSP plots for all simulations of wt and R282W p53. Secondary structure consistency is relatively similar in both variants of p53, but the helix near residue 180 (H1) is slightly less stable in the mutant simulations. In mutant simulation 1 (top) of p53 only, a slight helical turn forms in the S1 loop near residue 120.

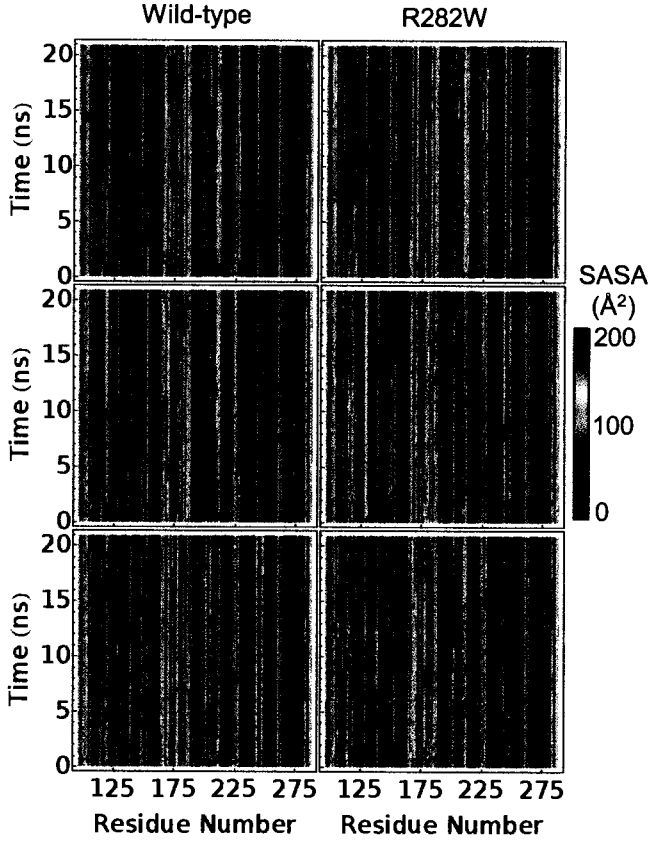


Figure 7.6: Plots of the SASA of all simulations of wt and R282W p53. Few obvious differences can be observed across simulations.

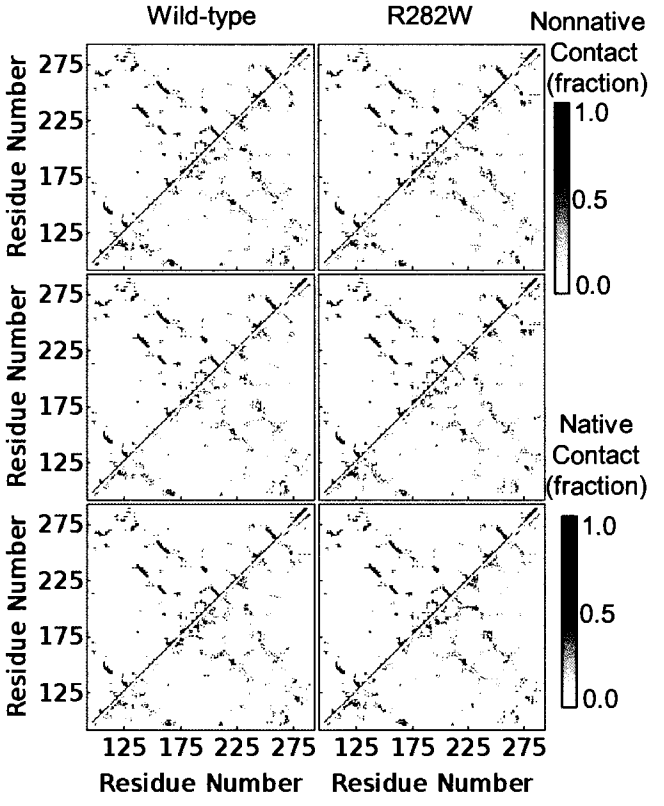


Figure 7.7: Native and non-native contacts between residues for each simulation of wt and R282W p53. All contacts are plotted as a fraction of time they occur throughout the simulation with native contacts appearing in white-to-blue in the upper left of each graph and non-native contacts appearing in white-to-red in the lower right.

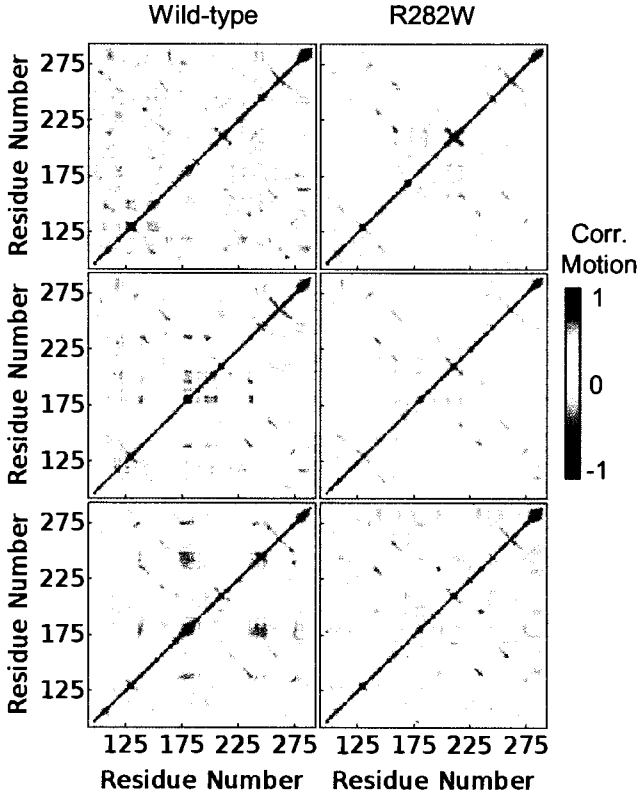


Figure 7.8: Correlated motion plots of all simulations of wt and R282W p53. Anti-correlated motion is slightly more prevalent in the wt simulations than the mutant simulations.

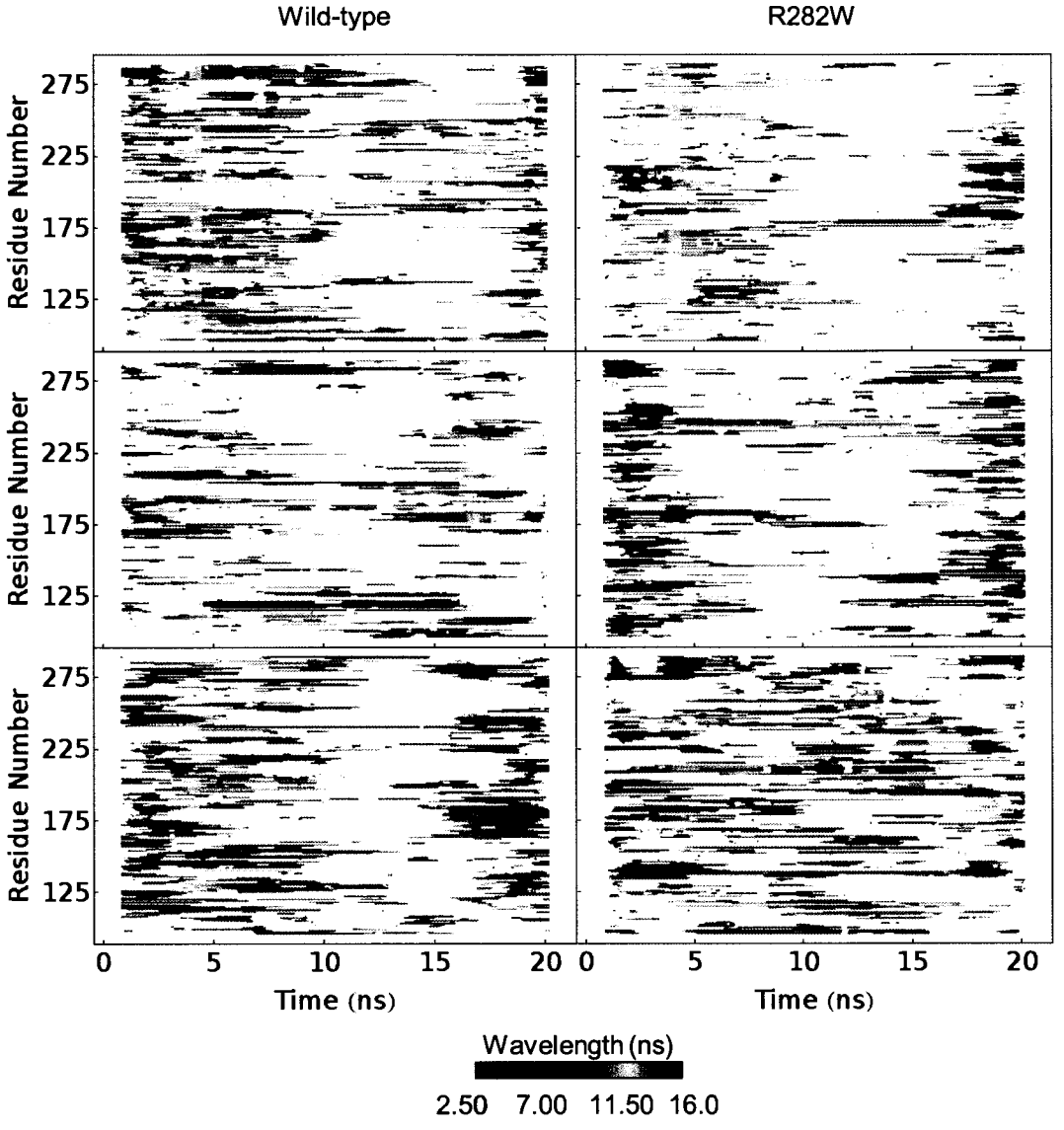


Figure 7.9: Wavelet analysis of all simulations of the wt and R282W mutant of p53. The most significant wavelet match at each time is shown for each Cα atom with white indicating no significant match. Critically, significant low frequency (high wavelength) bands of motion are seen throughout the time range of 5-10 ns in the wt simulation 1 (top) and scattered through wt simulation 2, especially near residue 120 (S1 loop). The mutant simulations show less motion overall, but simulation 1 shows bands of motion near residue 180 (Helix H1).

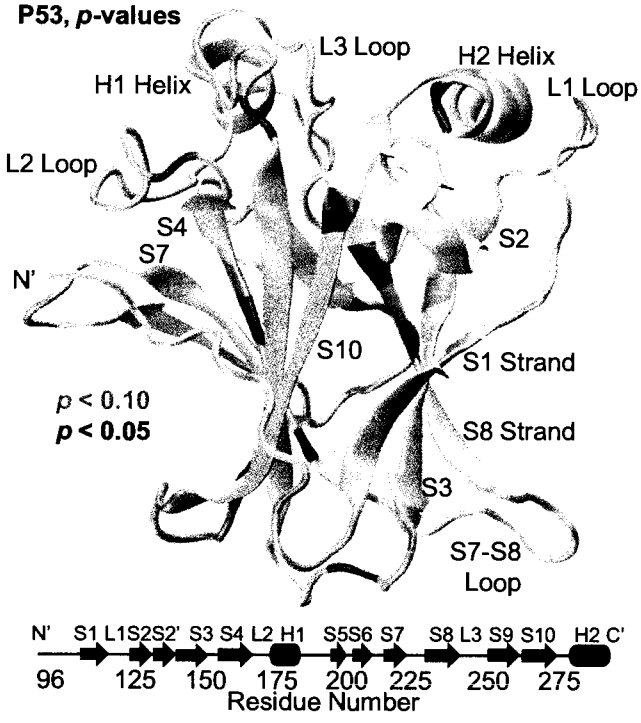


Figure 7.10: Significant differences in ordered motion between the wt and R282W p53 simulations, as determined by wavelet analysis. The greatest differences in motion tend to occur in the loops and the strands near the polymorphic site (S1 and S10) with a few significant differences in strands S8 and S4 as well.

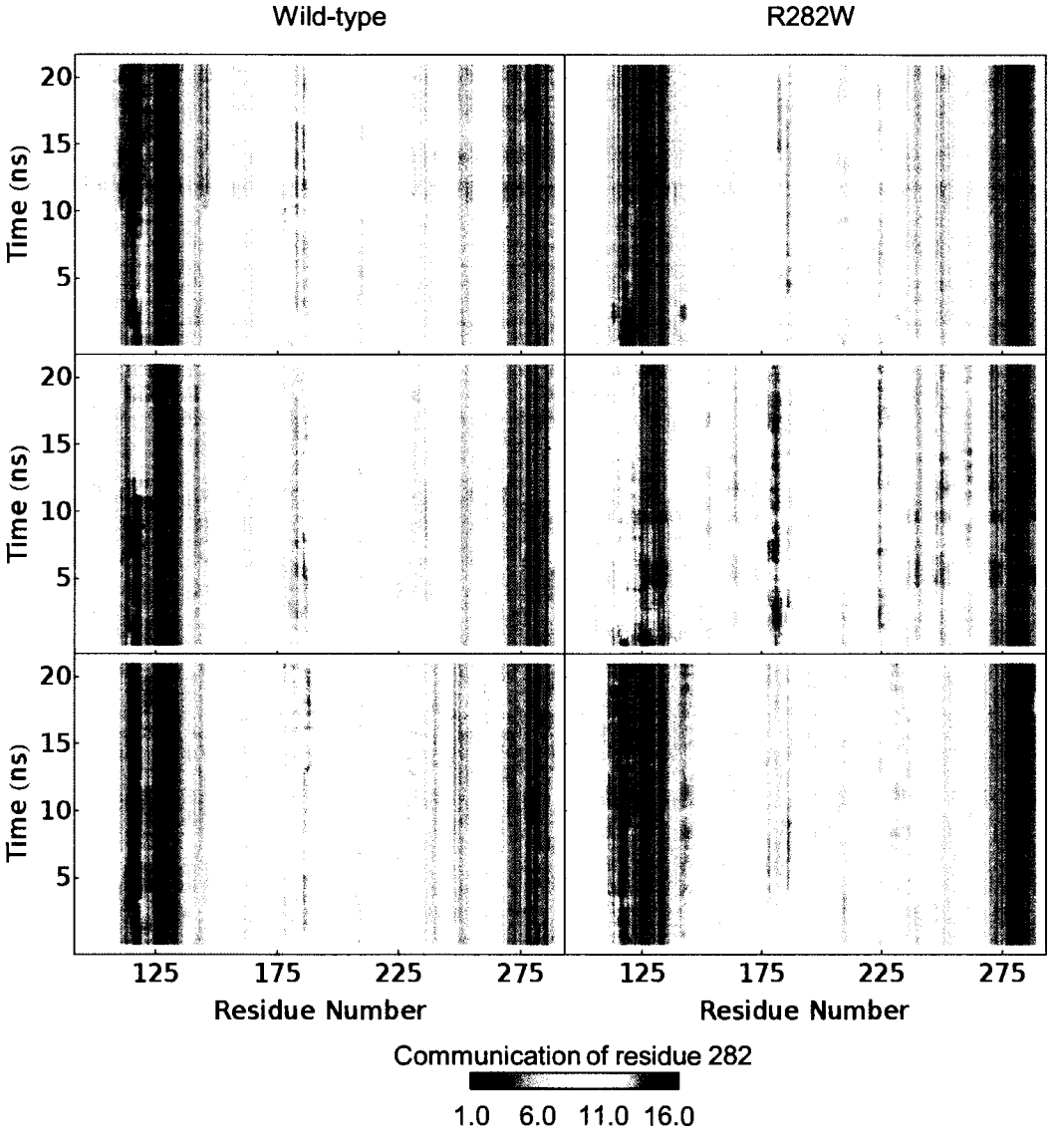


Figure 7.11: Graph communication plots of wt (R282) and R282W variants of p53. Plots show the distance from residue 282 to each other residue. Communication values are the lengths of the edges in the shortest path connecting the side-chain of residue 282 to a node of the residue in question where edges are the inverse probability that two nodes are in contact. A communication of 1 indicates that two nodes are always in contact while a communication of 4 indicates a weaker influence, for example, two side-chains that communicate via a third intermediate residue.

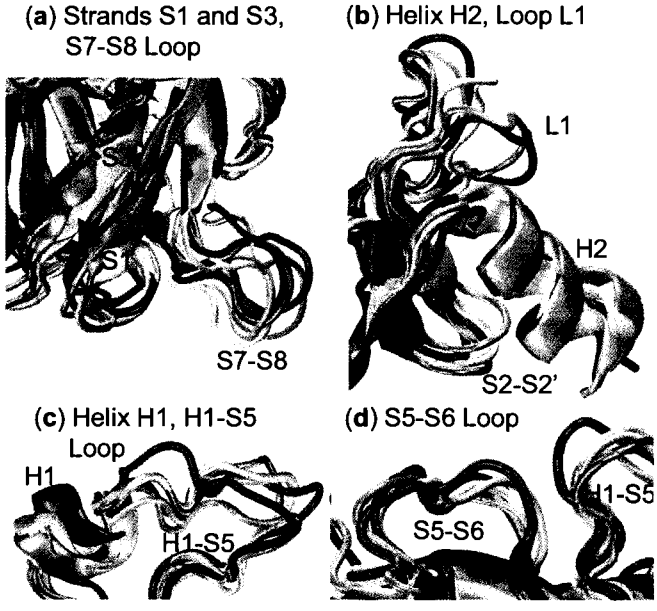


Figure 7.12: Movements observed in wt simulation 1 of p53 between 5 and 10 ns. Structures are shown for 5, 6, 7, 8, 9, and 10 ns colored blue, cyan, green, yellow, orange, and red, respectively. **(a)** Strands S1 and S3, each of which shift slightly, and the S7-S8 loop, which moves considerably throughout the 5-10 ns period. **(b)** The L1 loop and H2 Helix, each of which shifts. The L1 loop shows the most dramatic rearrangements of all structures during this time. **(c)** The H1 helix and the H1-S5 loop, each of which shift significantly from 5-10 ns. d the S5-S6 loop, which displays relatively little motion despite being solvent exposed and near the H1 helix during the 5-10 ns time range.

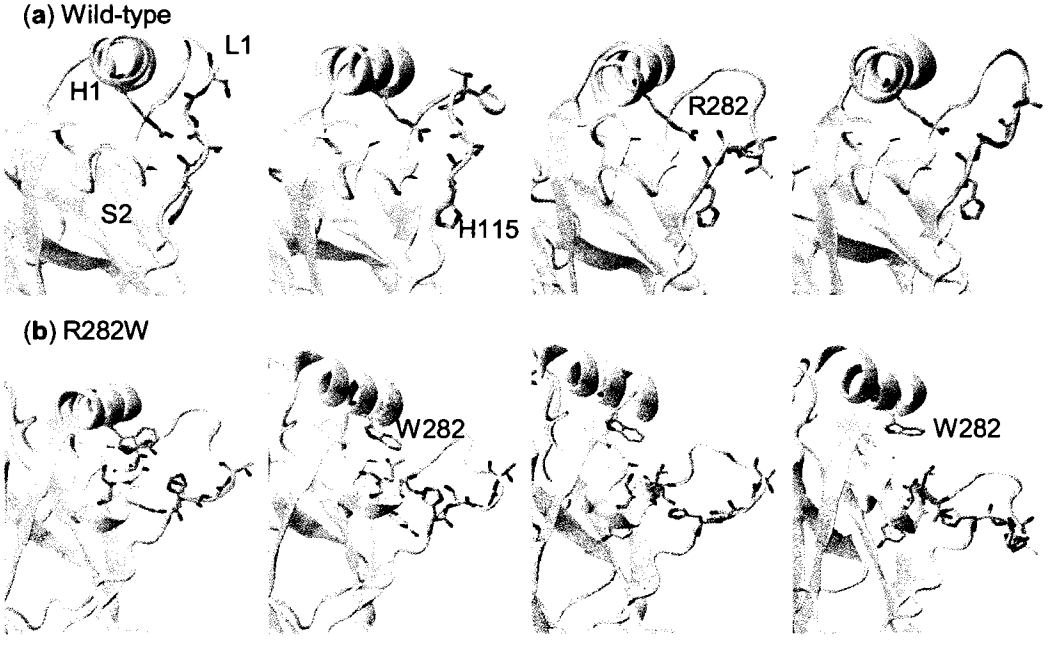


Figure 7.13: Snapshots taken sequentially from **(a)** wt and **(b)** R282W simulations of p53 featuring the H2 Helix and L1 Loop. In the wild-type, the R282 residue forms contacts with the backbone of loop L1 holding H2 and the base of L1 close together. This allows the tip of L1 to flex considerably along a single axis. The W282 residue, however, does not form these contacts, allowing H2 and L1 to separate. H115 of L1 forms contacts with the residues of S2 in this situation, preventing L1 from becoming too disorganized but allowing it to swing much more randomly into solvent.

BIBLIOGRAPHY

Abrahams, J. P., Leslie, A. G., Lutter, R. and Walker, J. E. (1994). Structure at 2.8 Å resolution of F1-ATPase from bovine heart mitochondria. Nature *370*, 621–628.

Addison, P. S., Watson, J. N. and Feng, T. (2002). Low-oscillation complex wavelets. Journal of Sound and Vibration *254*, 733–762.

Amadei, A., Linssen, A. B. and Berendsen, H. J. (1993). Essential dynamics of proteins. Proteins *17*, 412–25.

Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. and Pietrokovski, S. (2004). Network analysis of protein structures identifies functional residues. Journal of Molecular Biology *344*, 1135–1146.

Ang, H. C., Joerger, A. C., Mayer, S. and Fersht, A. R. (2006). Effects of common cancer mutations on stability and DNA binding of full-length p53 compared with isolated core domains. Journal of Biological Chemistry *281*, 21934–21941.

Arfken, G. (1985). Mathematical Methods for Physicists. 3 edition, Academic Press, Orlando, FL.

Askar, A., Cetin, A. E. and Rabitz, H. (1996). Wavelet Transform for Analysis of Molecular Dynamics. Journal of Physical Chemistry *100*, 19165–19173.

Beck, D. A. C., Alonso, D. O. V. and Daggett, V. (2004-2008). *in lucem* Molecular Mechanics. Technical report University of Washington Seattle, WA 98195.

Beck, D. A. C., Armen, R. S. and Daggett, V. (2005). Cutoff size need not strongly influence molecular dynamics results for solvated polypeptides. Biochemistry *44*, 609–16.

Beck, D. A. C. and Daggett, V. (2004). Methods for molecular dynamics simulations of protein folding/unfolding in solution. Methods *34*, 112–120.

Beck, D. A. C., Jonsson, A. L., Schaeffer, R. D., Scott, K. A., Day, R., Toofanny, R. D., Alonso, D. O. and Daggett, V. (2008). Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations. Protein: Engineering, Design and Selection *21*, 353–368.

Benson, N. C. and Daggett, V. (2008). Dynameomics: large-scale assessment of native protein flexibility. Protein Science *17*, 2038–50.

Benson, N. C. and Daggett, V. (2010a). Wavelet analysis of protein motion. Biophysical Journal *Submitted*.

Benson, N. C. and Daggett, V. (2010b). A graph theoretic approach to indexing protein dynamics. — *In Preparation*.

Benson, N. C., Rutherford, K. and Daggett, V. (2010). Understanding the Molecular Basis of Disease in Single Nucleotide Polymorphism Variants Using Wavelet Analysis. — *In Preparation*.

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

Bilder, R. M., Volavka, J., Lachman, H. M. and Grace, A. A. (2004). The catechol-O-methyltransferase polymorphism: relations to the tonic-phasic dopamine hypothesis and neuropsychiatric phenotypes. Neuropsychopharmacology *29*, 1943–61.

Black, S. D. and Mould, D. R. (1991). Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. Analytical Biochemistry *193*, 72–82.

Böde, C., Kovács, I. A., Szalay, M. S., Palotai, R., Korcsmáros, T. and Csermely, P. (2007). Network analysis of protein dynamics. FEBS Letters *581*, 2776–82.

Brinda, K. V. and Vishveshwara, S. (2005). A Network Representation of Protein Structures: Implications for Protein Stability. Biophysical Journal *89*, 4159–4170.

Bugni, J. M., Han, J., Tsai, M.-S., Hunter, D. J. and Samson, L. D. (2007). Genetic association and functional studies of major polymorphic variants of MGMT. DNA Repair *6*, 1116–1126.

Bullock, A. N., Henckel, J. and Fersht, A. R. (2000). Quantitative analysis of residual folding and DNA binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy. Oncogene *19*, 1245–1256.

Cañadillas, J. M., Tidow, H., Freund, S. M., Rutherford, T. J., Ang, H. C. and Fersht, A. R. (2006). Solution structure of p53 core domain: structural basis for its instability. Proceedings of the National Academy of Sciences of the United States of America *103*, 2109–2114.

Carugo, O. and Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Science *10*, 1470–1473.

Chiti, F. and Dobson, C. M. (2006). Protein misfolding, functional amyloid, and human disease. Annual Review of Biochemistry *75*, 333–366.

Coulthard, S. and Hogarth, L. (2005). The thiopurines: an update. Investigational New Drugs *23*, 523–32.

Cuff, A. L., Sillitoe, I., Lewis, T., Redfern, O. C., Garratt, R., Thornton, J. and Orengo, C. A. (2009). The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Research *37*, D310–D314.

Daggett, V. (2006). Protein Folding-Simulation. Chemical Reviews *106*, 1898–1916.

Daggett, V. and Fersht, A. (2003). The present view of the mechanism of protein folding. Nature Reviews Molecular Cell Biology *4*, 497–502.

Daggett, V. and Levitt, M. (1992). ten globule state from molecular dynamics simulations. Proceedings of the National Academy of Sciences of the United States of America *89*, 5142–5146.

Daggett, V., Li, A. J. and Fersht, A. R. (1998). Combined molecular dynamics and ϕ-value analysis of structure reactivity relationships in the transition state and unfolding pathway of barnase: structural basis of Hammond and anti-Hammond effects. Journal of the American Chemical Society *120*, 12740–12754.

Daubechies, I. (1992). Ten Lectures on Wavelets. 1 edition, Society for Industrial and Applied Mathematics, Philadelphia, PA.

David, C. L., Szumlanski, C. L., DeVry, C. G., Park-Hah, J. O., Clarke, S., Weinshilboum, R. M. and Aswad, D. W. (1997). Human erythrocyte protein L-isoaspartyl methyltransferase: heritability of basal activity and genetic polymorphism for thermal stability. Arch Biochem Biophys *346*, 277–86.

Davies, C., White, S. W. and Ramakrishnan, V. (1996). The crystal structure of ribosomal protein L14 reveals an important organizational component of the translational apparatus. Structure *4*, 55–66.

Dawling, S., Roodi, N., Mernaugh, R. L., Wang, X. and Parl, F. F. (2001). Catechol-O-methyltransferase (COMT)-mediated metabolism of catechol estrogens: comparison of wild-type and variant COMT isoforms. Cancer research *61*, 6716.

Day, R., Beck, D. A. C., Armen, R. S. and Daggett, V. (2003). A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. Protein Science *12*, 2150–2160.

Day, R. and Daggett, V. (2003). All-atom simulations of protein folding and unfolding. Advances in Protein Chemistry *66*, 373–403.

Day, R. and Daggett, V. (2005). Ensemble versus single-molecule protein unfolding. Proceedings of the National Academy of Sciences of the United States of America *102*, 13445–13450.

Day, R. and Daggett, V. (2007). Direct observation of microscopic reversibility in single-molecule protein folding. Journal of Molecular Biology *366*, 677–686.

Deeken, J. F., Figg, W. D., Bates, S. E. and Sparreboom, A. (2007). Toward individualized treatment: prediction of anticancer drug disposition and toxicity with pharmacogenetics. Anti-cancer drugs *18*, 111.

Delaunay, B. (1934). Sur la sphère vide. A la memoire de Georges Voronoi. Izvestiya Akademii Nauk SSSR, Otdelenie Matematicheskih i Estestvennyh Nauk *7*, 793–800.

DeVry, C. G. and Clarke, S. (1999). Polymorphic forms of the protein L-isoaspartate (D-aspartate) O-methyltransferase involved in the repair of age-damaged proteins. Journal of Human Genetics *44*, 275–88.

Dietmann, S. and Holm, L. (2001). Identification of homology in protein structure classification. Nature Sturcural Biology *8*, 953–957.

Fersht, A. R. and Daggett, V. (2002). Protein folding and unfolding at atomic resolution. Cell *108*, 573–582.

Fersht, A. R., Matouschek, A. and Serrano, L. (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. Journal of Molecular Biology *224*, 771–782.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. Sociometry *40*, 35–41.

Gaiddon, C., Lokshin, M., Ahn, J., Zhang, T. and Prives, C. (2001). A Subset of Tumor-Derived Mutant Forms of p53 Down-Regulate p63 and p73 through a Direct Interaction with the p53 Core Domain. Molecular and Cellular Biology *21*, 1874–1887.

Girard, B., Otterness, D. M., Wood, T. C., Honchel, R., Wieben, E. D. and Weinshilboum, R. M. (1994). Human histamine N-methyltransferase pharmacogenetics: cloning and expression of kidney cDNA. Mol Pharmacol *45*, 461–8.

Giuliani, A., Benigni, R., Zbilut, J. P., Webber, C. L., Sirabella, P. and Colosimo, A. (2002). Nonlinear signal analysis methods in the elucidation of protein sequence-structure relationships. Chemical Reviews *102*, 1471–92.

Glazer, D. S., Radmer, R. J. and Altman, R. B. (2009). Improving structure-based function prediction using molecular dynamics. Structure *17*, 919–929.

Goodman, J. E., Jensen, L. T., He, P. and Yager, J. D. (2002). Characterization of human soluble high and low activity catechol-O-methyltransferase catalyzed catechol estrogen methylation. Pharmacogenetics and Genomics *12*, 517.

Goupillaud, P., Grossman, A. and Morlet, J. (1984). Cycle-octave and related transforms in seismic signal analysis. Geoexploration *23*, 85–102.

Haar, A. (1910). Zur Theorie der orthogonalen Frunktionen-Systeme. Mathematische Annalen *69*, 331–371.

Hamroun, D., Kato, S., Ishioka, C., Claustres, M., Beroud, C. and Soussi, T. (2006). The UMD TP53 database and website: update and revisions. Human Mutation *27*, 14–20.

Hespenheide, G. M., Rader, A. J., Thorpe, M. F. and Kuhn, L. A. (2002). dentifying protein folding cores from the evolution of flexible regions during unfolding. Journal of Molecular Graphics and Modelling *21*, 195–207.

Holm, L. and Park, J. (2000). DaliLite workbench for protein structure comparison. Bioinformatics *16*, 566–567.

Horton, J. R., Sawada, K., Nishibori, M. and Cheng, X. (2005). Structural basis for inhibition of histamine N-methyltransferase by diverse drugs. Journal of Molecular Biology *353*, 334–44.

Horton, J. R., Sawada, K., Nishibori, M., Zhang, X. and Cheng, X. (2001). Two polymorphic forms of human histamine methyltransferase: structural, thermal, and kinetic comparisons. Structure *9*, 837–49.

Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J. and Tropsha, A. (2004). Mining protein family specific residue packing patterns from protein structure graphs. In RECOMB '04: Proceedings of the eighth annual international conference on Resaerch in computational molecular biology pp. 308–315, Association of Computing Machinery, New York, NY, USA.

Huang, C. S., Chern, H. D., Chang, K. J., Cheng, C. W., Hsu, S. M. and Shen, C. Y. (1999). Breast cancer risk associated with genotype polymorphism of the estrogen-metabolizing genes CYP17, CYP1A1, and COMT: a multigenic study on cancer susceptibility. Cancer research *59*, 4870.

Hubbard, S. and Thornton, J. M. (1993). NACCESS. Technical report Department of Biochemistry and Molecular Biology, University College London London, UK.

Humphrey, W., Dalke, A. and Schulten, K. (1996). VMD: visual molecular dynamics. Journal of Molecular Graphics *14*, 33–38.

Jmol (2010). Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/.

Joerger, A. C., Ang, H. C. and Fersht, A. R. (2006). Structural basis for understanding oncogenic p53 mutations and designing rescue drugs. Proceedings

of the National Academy of Sciences of the United States of America *103*, 15056–15061.

Joerger, A. C. and Fersht, A. R. (2007). Structure-function-rescue: the diverse nature of common p53 cancer mutants. Oncogene *26*, 2226–2242.

Jonsson, A. L., Scott, K. A. and Daggett, V. (2009). Dynameomics: a consensus view of the protein unfolding/folding transition state ensemble across a diverse set of protein folds. Biophysical Journal *97*, 2958–66.

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

Karplus, M. and Kuriyan, J. (2005). Molecular dynamics and protein function. Proceedings of the National Academy of the United States of America *102*, 6679–6685.

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

Kearsley, S. K. (1989). On the orthogonal transformation used for structural comparisons. Acta Crystallographica Section A: Foundations of Crystallography *45*, 208–210.

Kehl, C., Simms, A. M., Toofanny, R. D., Daggett, V. and Fersht, A. (2008). Dynameomics: a multi-dimensional analysis-optimized database for dynamic protein data. Protein: Design, Engineering, and Selection *21*, 379–386.

Kell, G. S. (1967). Precise representation of volume properties of water at one atmosphere. Journal of Chemical and Engineering Data *12*, 66–69.

Kondrashov, D. A., Van Wynsberghe, A. W., Bannen, R. M., Cui, Q. and G. N. Phillips, J. (2007). Protein structural variation in computational models and crystallographic data. Structure *15*, 135–136.

Krishnan, A., Zbilut, J. P., Tomita, M. and Giuliani, A. (2008). Proteins as networks: usefulness of graph theory in protein science. Current Protein Peptide Science *9*, 28–38.

Krynetski, E. Y., Schuetz, J. D., Galpin, A. J., Pui, C. H., Relling, M. V. and Evans, W. E. (1995). A single point mutation leading to loss of catalytic activity in human thiopuring S-methyltransferase. Proceedings of the National Academy of Sciences of the United States of America *92*, 949–953.

Kuntz, I. D. (1972). Protein folding. Journal of the American Chemical Society *94*, 4009–4012.

Ladurner, A. G., Itzhaki, L. S., Daggett, V. and Fersht, A. R. (1998). Synergy between simulation and experiment in describing the energy landscape of protein folding. Proceedings of the National Academy of the United States of America *95*, 8473–8478.

Lavigne, J. A., Helzlsouer, K. J., Huang, H. Y., Strickland, P. T., Bell, D. A., Selmin, O., Watson, M. A., Hoffman, S., Comstock, G. W. and Yager, J. D. (1997). An association between the allele coding for a low activity variant of catechol-O-methyltransferase and the risk for breast cancer. Cancer research *57*, 5493.

Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. Journal of Molecular Biology *55*, 379–400.

Leszczynski, J. F. and Rose, G. D. (1986). Loops in globular proteins: A novel category of secondary structure. Science *234*, 849–855.

Levitt, M. (1990). ENCAD, Energy Calculation and Dynamics. Technical report Stanford University Palo Alto, CA.

Levitt, M., Hirshberg, M., Sharon, R. and Daggett, V. (1995). Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Computer physics communications *91*, 215–231.

Levitt, M., Hirshberg, M., Sharon, R., Laidig, K. and Daggett, V. (1997). Calibration and testing of a water model for simulation of the molecular dynamics of proteins and nucleic acids in solution. J. Phys. Chem. B *101*, 5051–5061.

Li, A. J. and Daggett, V. (1994). Characterization of the transition-state of protein unfolding by use of molecular-dynamics: Chymotrypsin inhibitor 2. Proceedings of the National Academy of the United States of America *91*, 10430–10434.

Li, R. and Woodward, C. (1999). The hydrogen exchange core and protein folding. Protein Science *8*, 1571–1590.

Li, W., Kamtekar, S., Xiong, Y., Sarkis, G. J., Grindley, N. D. F. and Steitz, T. A. (2005). Structure of a synaptic gammadelta resolvase tetramer covalently linked to two cleaved DNAs. Science *309*, 1210–1215.

Ligneau, X., Lin, J. S., Vanni-Mercier, G., Jouvet, M., Muir, J. L., Ganellin, C. R., Stark, H., Elz, S., Schunack, W. and Schwartz, J. C. (1998). Neurochemical and behavioral effects of ciproxifan, a potent histamine H3-receptor antagonist. Journal of Pharmacology and Experimental Therapeutics *287*, 658.

Liò, P. (2003). Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics *19*, 2–9.

Matsui, A., Ikeda, T., Enomoto, K., Nakashima, H., Omae, K., Watanabe, M., Hibi, T. and Kitajima, M. (2000). Progression of human breast cancers to the metastatic state is linked to genotypes of catechol-O-methyltransferase. Cancer letters *150*, 23–31.

Mayor, U., Guydosh, N. R., Johnson, C. M., Grossmann, J. G., Sato, S., Jas, G. S., Freund, S. M. V., Alonso, D. O. V., Daggett, V. and Fersht, A. R. (2003). The complete folding pathway of a protein from nanoseconds to microseconds. Nature *421*, 863–867.

McCully, M. E., Beck, D. A. C. and Daggett, V. (2008). Microscopic reversibility of protein folding in molecular dynamics simulations of the engrailed homeodomain. Biochemistry *47*, 7079–7089.

Meyers, S. D., Kelly, B. G. and O'Brien, J. J. (1993). An Introduction to Wavelet Analysis in Oceanography and Meteorology: With Application to the Dispersion of Yanai Waves. Monthly Weather Review *121*, 2858–2866.

Microsoft (2008). SQL Server 2008 Enterprise x64 Edition. Microsoft Corporation.

Mitrunen, K., Kataja, V., Eskelinen, M., Kosma, V. M., Kang, D., Benhamou, S., Vainio, H., Uusitupa, M. and Hirvonen, A. (2002). Combined COMT and GST genotypes and hormone replacement therapy associated breast cancer risk. Pharmacogenetics and Genomics *12*, 67.

Monti, P., Campomenosi, P., Cibrilli, Y., Iannone, R., Aprile, A., Inga, A., Tada, M., Menichini, P., Abbondandolo, A. and Fronza, G. (2003). Characterization of the p53 mutants ability to inhibit p73β transactivation using a yeast-based functional assay. Oncogene *22*, 5252–5260.

Morisset, S., Rouleau, A., Ligneau, X., Gbahou, F., Tardivel-Lacombe, J., Stark, H., Schunack, W., Ganellin, C. R. and Arrang, J. M. (2000). High constitutive activity of native H3 receptors regulates histamine neurons in brain. Nature *408*, 860–864.

Murdock, S. E., Tai, K., Ng, M. H., Johnston, S., Wu, B., Fangohr, H., Essex, J. W., Jeffreys, P., Cox, S. and Sansom, M. S. P. (2005). BioSimGrid: A distributed environment for archiving and the analysis of biomolecular simulations. Abstracts of Papers of the American Chemical Society *230*, U1309–U1310.

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology *247*, 536–540.

Ng, M. H., Johnston, S., Wu, B., Murdock, S. E., Tai, K. H., Fangohr, H., Cox, S. J., Essex, J. W., Sansom, M. S. P. and Jeffreys, P. (2006). BioSimGrid: Grid-enabled biomolecular simulation data storage and analysis. Future Generation Computer Systems *22*, 657–664.

Olivier, M., Eeles, R., Hollstein, M., Khan, M. A., Harris, C. C. and Hainaut, P. (2002). The IARC TP53 database: new online mutation analysis and recommendations to users. Human Mutation *19*, 607–614.

Olivier, M., Langerod, A., Carrieri, P., Bergh, J., Klaar, S., Eyfjord, J., andC. Rodriguez, C. T., Lidereau, R., Bièche, I., Varley, J., Bignon, Y., Uhrhammer, N., Winqvist, R., Jukkola-Vuorinen, A., Niederacher, D., Kato, S., Ishioka, C., Hainaut, P. and Børresen-Dale, A. L. (2006). The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer. Clinical Cancer Research *12*, 1157–1167.

Oroszi, G., Enoch, M.-A., Chun, J., Virkkunen, M. and Goldman, D. (2005). Thr105Ile, a functional polymorphism of histamine N-methyltransferase, is associated with alcoholism in two independent populations. Alcohol Clin Exp Res *29*, 303–9.

Pahlich, S., Zakaryan, R. P. and Gehring, H. (2006). Protein arginine methylation: Cellular functions and methods of analysis. Biochim Biophys Acta *1764*, 1890–903.

Panula, P., Rinne, J., Kuokkanen, K., Eriksson, K. S., Sallmen, T., Kalimo, H. and Relja, M. (1997). Neuronal histamine deficit in Alzheimer's disease. Neuroscience *82*, 993–997.

Pauling, L. and Corey, R. B. (1951). The pleated sheet, a new layer of configuration of polypeptide chains. Proceedings of the National Academy of Sciences of the United States of America *37*, 251–256.

Preuss, C. V., Wood, T. C., Szumlanski, C. L., Raftogianis, R. B., Otterness, D. M., Girard, B., Scott, M. C. and Weinshilboum, R. M. (1998). Human histamine N-methyltransferase pharmacogenetics: common genetic polymorphisms that alter activity. Mol Pharmacol *53*, 708–17.

Price, R. A., Scott, M. C. and Weinshilboum, R. M. (1993). Genetic segregation analysis of red blood cell (RBC) histamine N-methyltransferase (HNMT) activity. Genet Epidemiol *10*, 123–31.

Reuter, M., Jeste, N., Klein, T., Hennig, J., Goldman, D., Enoch, M.-A. and Oroszi, G. (2007). Association of THR105Ile, a functional polymorphism of histamine N-methyltransferase (HNMT), with alcoholism in German Caucasians. Drug Alcohol Depend *87*, 69–75.

Rueda, M., Ferrer-Costa, C., Meyer, T., Perez, A., Camps, J., Hospital, A., Gelpi, J. L. and Orozco, M. (2007). A consensus view of protein dynamics. Proceedings of the National Academy of the United States of America *104*, 796–801.

Rutherford, K., Alphandéry, E., McMillan, A., Daggett, V. and Parson, W. W. (2008). The V108M mutation decreases the structural stability of catechol O-methyltransferase. Biochim Biophys Acta *1784*, 1098–105.

Rutherford, K., Bennion, B. J., Parson, W. W. and Daggett, V. (2006). The 108M polymorph of human catechol O-methyltransferase is prone to deformation at physiological temperatures. Biochemistry *45*, 2178–2188.

Rutherford, K. and Daggett, V. (2008). Four Human Thiopurine S-Methyltransferase Alleles Severely Affect Protein Structure and Dynamics. Journal of Molecular Biology *379*, 803–814.

Rutherford, K. and Daggett, V. (2009a). A hotspot of inactivation: The A22S and V108M polymorphisms individually destabilize the active site structure of catechol O-methyltransferase. Biochemistry *48*, 6450–60.

Rutherford, K. and Daggett, V. (2009b). The V119I polymorphism in protein L-isoaspartate O-methyltransferase alters the substrate-binding interface. Protein: Engineering, Design and Selection *22*, 713–721.

Rutherford, K. and Daggett, V. (2010). Polymorphisms and disease: hotspots of inactivation in methyltransferases. Trends in Biochemical Sciences *0*, 0–1.

Rutherford, K., Le Trong, I., Stenkamp, R. E. and Parson, W. W. (2008a). Crystal structures of human 108V and 108M catechol O-methyltransferase. Journal of Molecular Biology *380*, 120–30.

Rutherford, K., Parson, W. W. and Daggett, V. (2008b). The histamine N-methyltransferase T105I polymorphism affects active site structure and dynamics. Biochemistry *47*, 893–901.

Saito, S., Iida, A., Sekine, A., Miura, Y., Sakamoto, T., Ogawa, C., Kawauchi, S., Higuchi, S. and Nakamura, Y. (2001). Identification of 197 genetic variations in six human methyltranferase genes in the Japanese population. Journal of Human Genetics *46*, 529–37.

Santarelli, L. C., Wassef, R., Heinemann, S. H. and Hoshi, T. (2006). Three methionine residues located within the regulator of conductance for K+ (RCK) domains confer oxidative sensitivity to large-conductance Ca2+-activated K+ channels. Journal of Phhysiology *571*, 329–348.

Sazci, A., Ergul, E., Utkan, N. Z., Canturk, N. Z. and Kaya, G. (2004). Catechol-O-methyltransferase Val 108/158 Met polymorphism in premenopausal breast cancer patients. Toxicology *204*, 197–202.

Schaeffer, R. D., Fersht, A. R. and Daggett, V. (2008). Combining experiment and simulation in protein folding: closing the gap for small model systems. Current Opinion in Structural Biology *18*, 4–9.

Schaeffer, R. D., Jonsson, A. L., Simms, A. M. and Daggett, V. (2010). Generation of a consensus protein domain dictionary. — *In Preparation*.

Schmidlin, T., Kennedy, B. K. and Daggett, V. (2009). Structural Changes to Monomeric CuZn Superoxide Dismutase Caused by the Familial Amyotrophic Lateral Sclerosis-Associated Mutation A4V. Biophysical Journal *97*, 1709–1718.

Scott, M. C., Van Loon, J. A. and Weinshilboum, R. M. (1988). Pharmacogenetics of N-methylation: heritability of human erythrocyte histamine N-methyltransferase activity. Clin Pharmacol Ther *43*, 256–62.

Shen, Y. and Bax, A. (2007). Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. Journal of Biomolecular NMR *38*, 289–302.

Shield, A. J., Thomae, B. A., Eckloff, B. W., Wieben, E. D. and Weinshilboum, R. M. (2004). Human catechol O-methyltransferase genetic variation: gene resequencing and functional characterization of variant allozymes. Molecular psychiatry *9*, 151–160.

Shrake, A. and Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. Journal of Molecular Biology *79*, 351–371.

Silva, C. G., Ostropytskyy, V., Loureiro-Ferreira, N., Berrar, D., Swain, M., Dubitzky, W. and Brito, R. M. (2006). P-found: The protein folding and unfolding simulation repository. In Proceedings of the 2006 IEEE Symposion on Computation Intelligence in Bioinformatics and Computational Biology pp. 101–108,, Toronto, ON.

Simms, A. M., Toofanny, R. D., Kehl, C., Benson, N. C. and Daggett, V. (2008). Dynameomics: design of a computational lab workflow and scientific data repository for protein simulations. Protein: Design, Engineering, and Selection *21*, 369–377.

Skinner, M. M., Puvathingal, J. M., Walter, R. L. and Friedman, A. M. (2000). Crystal structure of protein isoaspartyl methyltransferase: a catalyst for protein repair. Structure *8*, 1189–201.

Stadtman, E. R., Moskovitz, J. and Levine, R. L. (2003). Oxidation of methionine residues of proteins: biological consequences. Antioxidants and Redox Signaling *5*, 577–582.

Strachan, T. and Read, A. P. (1999). Human Molecular Genetics. 2 edition, Wiley-Liss, New York.

Swint-Kruse, L. (2004). Using networks to identify fine structural differences between functionally distinct protein states. Biochemistry *43*, 10886–95.

Tai, H. L., Fessing, M. Y., Bonten, E. J., Yanishevsky, Y., d'Azzo, A., Krynetski, E. Y. and Evans, W. E. (1999). Enhanced proteasomal degradation of mutant human thiopurine S-methyltransferase (TPMT) in mammalian cells: mechanism for TPMT protein deficiency inherited by TPMT*2, TPMT*3A, TPMT*3B or TPMT*3C. Pharmacogenetics *9*, 641–650.

Tai, H. L., Krynetski, E. Y., Schuetz, E. G., Yanishevski, Y. and Evans, W. E. (1997). Enhanced proteolysis of thiopurine S-methyltransferase (TPMT) encoded by mutant alleles in humans (TPMT*3A, TPMT*2): Mechanisms for the genetic polymorphism of TPMT activity. Proceedings of the National Academy of Sciences of the United States of America *94*, 6444.

Tan, W., Qi, J., Xing, D. Y., Miao, X. P., Pan, K. F., Zhang, L. and Lin, D. X. (2003). Relation between single nucleotide polymorphism in estrogen-metabolizing genes COMT, CYP17 and breast cancer risk among Chinese women. Zhonghua zhong liu za zhi [Chinese journal of oncology] *25*, 453.

Teodoro, M. L., Phillips, Jr., G. N. and Kavraki, L. E. (2003). Understanding protein flexibility through dimensionality reduction. Journal of Computational Biology *10*, 617–634.

Thompson, P. A., Shields, P. G., Freudenheim, J. L., Stone, A., Vena, J. E., Marshall, J. R., Graham, S., Laughlin, R., Nemoto, T., Kadlubar, F. F. and Ambrose, C. B. (1998). Genetic polymorphisms in catechol-O-methyltransferase, menopausal status, and breast cancer risk. Cancer research *58*, 2107.

Toofanny, R. D., Jonsson, A. L. and Daggett, V. (2010). A comprehensive multidimensional-embedded, one-dimensional reaction coordinate for protein unfolding/folding. Biophysical Journal *98*, 2671–2681.

Torrence, C. and Compo, G. P. (1998). A Practical Guide to Wavelet Analysis. Bulletin of the American Meteorological Society *79*, 61–78.

Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Wenger, R. K., Yao, H. and Markley, J. L. (2007). BioMagResBank. Nucleic Acids Research *36*, D402–D408.

van der Kamp, M. W., Schaeffer, R. D., Jonsson, A. L., Scouras, A. D., Simms, A. M., Toofanny, R. D., Benson, N. C., Anderson, P. C., Merkley, E. D., Rysavy, S., Bromley, D., Beck, D. A. C. and Daggett, V. (2010). Dynameomics: a comprehensive database of protein dynamics. Structure *18*, 423–35.

van der Kamp, M. W., Shaw, K. E., Woods, C. J. and Mulholland, A. J. (2008). Biomolecular simulation and modelling: Status, progress and prospects. Journal of the Royal Society Interface *5*, S173–S190.

Vendruscolo, M., Dokholyan, N. V., Paci, E. and Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. Physical Review E *65*, 061910.

Vendruscolo, M., Paci, E., Dobson, C. M. and Karplus, M. (2001). Three key residues form a critical contact network in a protein folding transition state. Nature *409*, 641–645.

Vidgren, J., Svensson, L. A. and Liljas, A. (1994). Crystal structure of catechol O-methyltransferase. Nature *368*, 354–8.

Wang, Y., Rosengarth, A. and Luecke, H. (2007). Structure of the human p53 core domain in the absence of DNA. Acta Crystallographica Section D: Biological Crystallography *63*, 276–281.

Webber, C. L., Giuliani, A., Zbilut, J. P. and Colosimo, A. (2001). Elucidating protein secondary structures using alpha-carbon recurrence quantifications. Proteins *44*, 292–303.

Wedren, S., Rudqvist, T. R., Granath, F., Weiderpass, E., Ingelman-Sundberg, M., Persson, I. and Magnusson, C. (2003). Catechol-O-methyltransferase gene polymorphism and post-menopausal breast cancer risk. Carcinogenesis *24*, 681.

Weinshilboum, R. M. (2006). Pharmacogenomics: catechol O-methyltransferase to thiopurine S-methyltransferase. Cellular and Molecular Neurobiology *26*, 539–61.

Weiser, J., Shenkin, P. S. and Still, W. C. (1999). Approximate solvent-accessible surface areas from tetrahedrally directed neighbor densities. Biopolymers *50*, 373–380.

Wolfram Research, I. (2005). Mathematica. 5.2 edition, Wolfram Research, Inc., Champaign, Illinois.

Wolfram Research, I. (2008). Mathematica. 7.0 edition, Wolfram Research, Inc., Champaign, Illinois.

Wu, H., Horton, J. R., Battaile, K., Allali-Hassani, A., Martin, F., Zeng, H., Loppnau, P., Vedadi, M., Bochkarev, A., Plotnikov, A. N. and Cheng, X. (2007). Structural basis of allele variation of human thiopurine-S-methyltransferase. Proteins *67*, 198–208.

Yang, D. S., Hon, W. C., Bubanko, S., Xue, Y., Seetharaman, J., Hew, C. L. and Sicheri, F. (1998). Identification of the ice-binding surface on a type III antifreeze protein with a flatness function algorithm. Biophysical Journal *74*, 2142–2151.

Yang, W. and Steitz, T. A. (1995). Crystal structure of the site-specific recombinase gamma delta resolvase complexed with a 34 bp cleavage site. Cell *82*, 193–207.

Yee, D. P. and Dill, K. A. (1993). Families and the structural relatedness among globular proteins. Protein Science *2*, 884–899.

Yim, D. S., Park, S. K., Yoo, K. Y., Yoon, K. S., Chung, H. H., Kang, H. J., Ahn, S. H., Noh, D. Y., Choe, K. J., Jang, I. J., Shin, S. G., Strickland, P. T., Hirvonen, A. and Kang, D. (2001). Relationship between the Val158Met polymorphism of catechol O-methyl transferase and breast cancer. Pharmacogenetics and Genomics *11*, 279.

Appendix A

DYNAMANAL: MOLECULAR DYNAMICS ANALYSIS FOR THE DYNAMEOMICS DATABASE

A.1 Summary

Dynamanal is a web-service designed for easy analysis of molecular dynamics trajectories associated with the Dynameomics project, www.dynameomics.org. Because of the size and complexity of the Dynameomcis warehouse and because many users are not familiar with the database schema, analysis can be very difficult and time-consuming, especially for researchers not versed in database techniques. Dynamanal is designed to facilitate quick and user-friendly analysis of these trajectories. Dynamanal is free as a web-service and is available online as a trajectory viewing tool incorporated into http://www.dynameomics.org/.

A.2 Introduction

The Dynameomics project (Beck et al., 2008; van der Kamp et al., 2010) is a large-scale molecular dynamics (MD) initiative with the goal of simulating a representative from every fold family in the Protein Data Bank (PDB) (Berman et al., 2000). This goal has recently been achieved, and, currently, the project encompasses 807 fold families and over 920 proteins, each of which has been simulated for at least 31 ns at 298 K and at least 68 ns at 498 K for a total of $\sim$150 μs. Adjunct simulations of single nucleotide polymorphism (SNP) associated proteins, amyloid proteins, and various other systems expand the size and scope considerably. Due to the size and complexity of these data, a data

warehouse was constructed (Simms et al., 2008; Kehl et al., 2008) to contain coordinate data for these simulations as well as data from a standard set of analyses of the trajectories.

Although this data warehouse is consistently and well organized, it is simultaneously difficult to navigate. This is largely because of the complex schema that is required to link together not only the several types of raw data held within it but also to connect them to the scientific work-flow and associated meta-data that generated them. In order to efficiently hold all the relevant data, the data warehouse had to be distributed into several smaller databases, each with hundreds of tables. Dynamanal is a web-service attached to the Dynameomics warehouse that not only brings together a large set of useful MD analyses, but also allows users to visualize these data in multiple ways without the need to understand the details of the database's organization, while seamlessly connecting the analyses to the underlying structures.

Dynamanal can be reached by visiting the Dynameomics website at URL www.dynameomics.org and clicking through to the details of any simulation. It has been written in Java and is compatible with any operating system or browser that is properly configured and updated. Figure A.1 shows a screenshot of Dynamanal examining the 298 K simulation of the protein Ubiquitin (*1ubq*). This instance of Dynamanal can be reached by searching for "1ubq" from the Dynameomics home-page, clicking on the "beta-Grasp (ubiquitin-like)" fold link, clicking the "Ubiquitin" target link, and finally selecting the first run of ubiquitin at 298 K.

A.3 Analysis

Dynamanal currently supports seven types of analysis, each of which can be reached by clicking the appropriate tab. These analyses are root mean square deviation (RMSD), solvent accessible surface area (SASA), Ramachan-

dran plots (Phi-Psi), Conformational Genealogy (Congeneal) (Yee and Dill, 1993), number of contacts (Contacts), radius of gyration, and the dictionary of secondary structure of proteins (DSSP) (Kabsch and Sander, 1983).

RMSD is a measure of the distance of a particular protein conformation from the starting structure. Dynamanal plots the total or normalized RMSD of all Cα atoms. Because larger proteins should naturally have higher RMSD values than smaller proteins, RMSD100 is a normalized measure of RMSD that estimates what an equivalent RMSD might be for a protein of only 100 residues (Carugo and Pongor, 2001).

SASA is a measure of the surface area of a particular residue that is accessible to solvent water molecules. It is plotted on a per-residue basis over time, and can be plotted for specific atoms in a residue, such as backbone atoms only, or polar atoms only.

Ramachandran plots show the distribution of ϕ-ψ space that a residue inhabits over time. Two views are show: traditional Ramachandran plots as well as a 2D histogram of both ϕ and ψ.

The number of contacts is a useful measurement of deviation from a crystal structure over time. It can show the total number of contacts as well as the number of native and nonnative contacts separately. The radius of gyration analysis is displayed in a similar manner to contacts; it plots the mass-weighted RMS of all atoms from the protein's center of mass. The congeneal analysis is also plotted in a similar fashion as contacts and radius of gyration. CONGENEAL stands for conformational genealogy, which is a weighted measure of the difference between the internal distances of the residues in two structures (Yee and Dill, 1993).

The final analysis, DSSP, displays secondary structure over time. DSSP uses hydrogen bonding patterns to determine secondary structure, and can identify a wide variety including 3-10 helices, π helices, bridges, and α-sheet

in addition to α-helices and β-sheets via in-house extensions to the standard DSSP (Kabsch and Sander, 1983; Pauling and Corey, 1951).

A.3.1 Interface

Dynamanal allows the user to zoom in on plots along the x-axis; for contacts, radius of gyration, RMSD, congeneal, and SASA, this axis is time, while for DSSP it is residue, and for phi-psi it is simultaneously ϕ and ψ (on the histogram). The zoom may be reset by right clicking on the plot and selecting "Reset Zoom". The right-click menu also allows the user to precisely determine the coordinates of the pixel that is right-clicked on, as well as to save the current graph as a PNG, or to download the structure corresponding to the time of the pixel as a PDB file.

Many graphs have advanced options. For example, the SASA graph allows the user to select subsets of atoms for display on the graph as well as allowing the user to display the min/max values as a dotted line. Because there are so many more time points than pixels on the graph, the plot is shown as the mean value of the times represented by a single pixel as well as the min and max values as a dotted line. At a sufficiently high zoom, raw data are displayed.

Dynamanal interfaces with Jmol (Jmol, 2010), allowing a user to easily visualize the structures corresponding to any particular feature of an analysis. To bring up a structure in Jmol, the user may either enter it in a text box beneath the Jmol window or may right-click in a graph and the structure corresponding to the time on the plot is displayed. Both of these methods will bring up the appropriate structure in the Jmol window, and the latter will mark the time in the graph with a red line.

A.4 Conclusions

Dynamanal fills an important hole in the current landscape of tools for MD simulations by allowing one lab's simulations to be analyzed and verified by other scientists. This shift in the normal operations of MD will become more and more important as MD databases become more common.

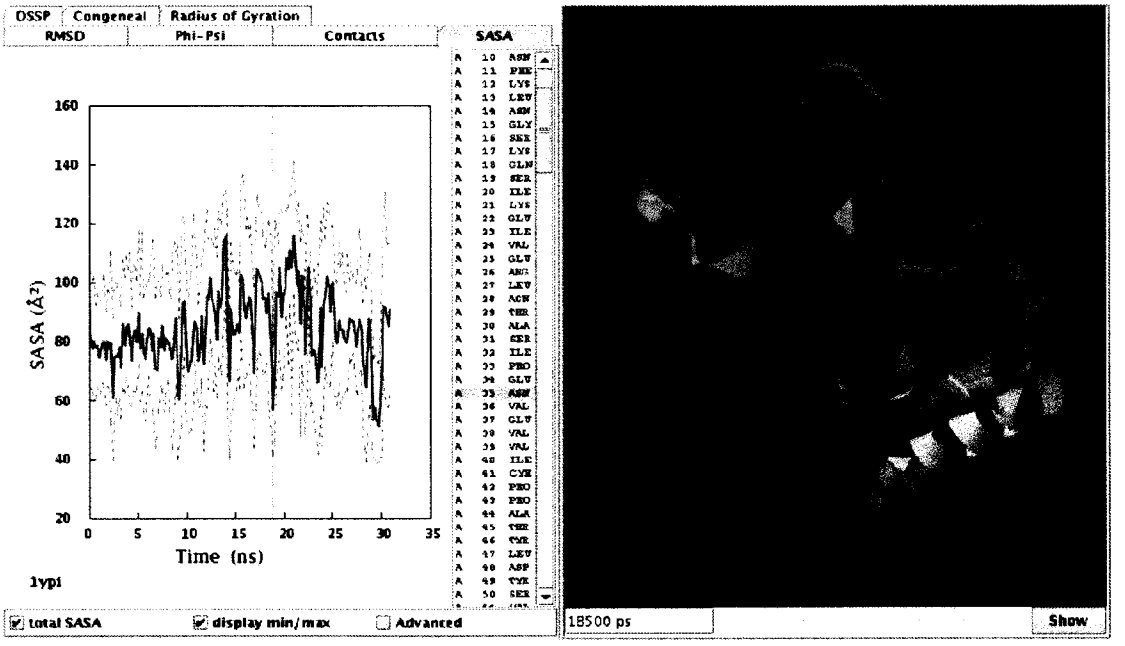


Figure A.1: A screen-shot of Dynamanal viewing analyses of the protein Ubiquitin. The red vertical line on the graph shows the time-point corresponding to the structure shown in Jmol on the right.

Appendix B

MATHEMATICA CODES FOR DATABASE ACCESS

In order to facilitate database access and user-developed queries and analyses, we have developed a Mathematica package that handles many difficult aspects of navigating the Dynameomics database automatically. The entirety of this packages is given below.

```
BeginPackage["Dynameomics`",{"DatabaseLink`","JLink`"}];

(* Errors/Messagese *)
MakeKey::unrecognized = "Specifier `1` is not recognized.";
Rows::unspecified = "Key `1` is not properly filled.";
Columns::unspecified = "Key `1` is not properly filled.";

(* All usage text here *)
DynamKeyToString::usage =
  "Turns a Dynameomics key into a string for a WHERE
   clause";
MakeKey::usage =
  "MakeKey[<rules>] returns a key that matches the
   indicated rules.  Possible rules are PDB, Run, Temp,
   Property, SimID, StructID, CID, Conditions, POD, and
   pH";
PDB::usage =
  "Rule argument to MakeKey";
```

```
Run::usage =
  "Rule argument to MakeKey";
Temp::usage =
  "Rule argument to MakeKey";
Property::usage =
  "Rule argument to MakeKey";
SimID::usage =
  "Rule argument to MakeKey";
StructID::usage =
  "Rule argument to MakeKey";
CID::usage =
  "Rule argument to MakeKey";
Conditions::usage =
  "Rule argument to MakeKey";
POD::usage =
  "Rule argument to MakeKey";
pH::usage =
  "Rule argument to MakeKey";
FillKey::usage =
  "FillKey[<key>] returns a fleshed out version of <key>
   that includes information about possible values for all
   fields.  This is useful when describing a key.";
KeyDetails::usage =
  "Describes a key.  If the key has been run through
   FillKey, it gives possible values for unspecified
   fields.";
KeyQuery::usage =
  "KeyQuery[key, prefix, suffix] runs the query specified
```

```
   by prefix and suffix, filling in the database name and
   property table from key.  For example, KeyQuery[k,
   \"select top 10 *\", \"residue_number=10 order by
   residue_number\"]";
BasicQuery::usage =
  "BasicQuery[key, string] runs the query specified by
   string, replacing any substring TABLE with the
   appropriate name of the key's table.";
Rows::usage =
  "Returns the number of rows in the table specified by a
   particular key.";
Columns::usage =
  "Returns a list of the names of the columns in the table
   specified by a particular key.";
Query::usage =
  "Query[key, columns, where] returns the results of the
   query of key's table formed over the columns specified
   in a string, columns, and the where clause specified as
   a string";
RowBuffer::usage =
  "Option to Query that specifies the number of rows to
   page";
Where::usage =
  "Option to Query that specifies the where clause of the
   query";
OrderBy::usage =
  "Option to Query that specifies how the query is sorted
   (default: step)";
```

```
GroupBy::usage =
  "Option to Query that specifies how the query is grouped
   (default: None)";

$Helix::usage = "The SQL connection to Helix";
$Turn::usage = "The SQL connection to Turn";
$Coil::usage = "The SQL connection to Coil";
$Wudang::usage = "The SQL connection to Wudang";
$Sheet::usage = "The SQL connection to Sheet";
$Strand::usage = "The SQL connection to Helix";
DynamConn::usage =
  "DynamConn[server] returns the SQL connection object for
   server if it exists.";

Begin["`Private`"];

If[!ValueQ[DynamDriver],
  DynamDriver = "jtds_sqlserver"];
(* otherwise: DynamDriver = "sqljdbc"; *)
JDBCDrivers[DynamDriver];

DynamConnect[server_String, username_String, passwd_String]
  := If[DynamDriver=="jtds_sqlserver",
        OpenSQLConnection[JDBC[DynamDriver, server],
                          Username->username,
                          Password->passwd],
        OpenSQLConnection[JDBC[DynamDriver, server],
                          Username->username,
```

```
                        Password->passwd]];

DynamDir = DynamConnect["helix","worker_bee","workerbee"];
DynamConn["HELIX"] = DynamDir;

(* the math database *)
DynamConn["WUDANG"] =
   DynamConnect["wudang","math_user","mathuser"];
DynamConn["TURN"] =
   DynamConnect["turn","worker_bee","workerbee"];
DynamConn["COIL"] =
   DynamConnect["coil","worker_bee","workerbee"];
DynamConn["STRAND"] =
   DynamConnect["strand","worker_bee","workerbee"];
DynamConn["SHEET"] =
   DynamConnect["sheet","worker_bee","workerbee"];
DynamConn["STRAND"] =
   DynamConnect["strand","worker_bee","workerbee"];

$Helix = DynamConn["HELIX"];
$Turn = DynamConn["TURN"];
$Wudang = DynamConn["WUDANG"];
$Coil = DynamConn["COIL"];
$Strand = DynamConn["STRAND"];
$Sheet = DynamConn["SHEET"];

DynamGetConn[x_String] :=
  (If[!ValueQ[DynamConn[x]],
```

```
      DynamConn[x] = DynamConnect[x, "worker_bee",
                                      "workerbee"]];
    DynamConn[x]);

DynamKeyItems =
 {{"PDB", "PDB4"},
  {"Run", "Run"},
  {"Temp", "Temp"},
  {"Property", "Property_Abbrev"},
  {"SimID", "Sim_ID"},
  {"StructID", "Struct_ID"},
  {"CID", "CID"},
  {"Conditions", "Conditions"},
  {"PID", "PID"},
  {"pH", "pH"},
  {"Server", "server_name"},
  {"DB", "database_name"},
  {"PropTable", "Property_table"}};

DynamKeyToString[k_DynamKey] :=
  StringDrop[
   StringJoin[
    Apply[
     Sequence,
     MapThread[
      If[SameQ[#2, None] || SameQ[Head[#2], List], "",
        StringJoin[
         DynamKeyItems[[#1, 2]], "=",
```

```
        If[StringQ[#2],
           "'" <> #2 <> "'", ToString[#2]] <> " AND "]] &,
      {Range[1, Length[First[k]]], First[k]}]]],
   -5];

FillKey[k_DynamKey] :=
   Module[{tmp,key},
     key = DynamKey[
       Map[
        If[Length[#]==1,First[#],#]&,
        Map[
         Union,
         Transpose[
          SQLExecute[DynamDir,
           StringJoin[
            "select distinct ",
            StringDrop[
              StringDrop[
                ToString[
                 DynamKeyItems[[All,2]]],1],-1],
            " from Directory.dbo.Master_Property_v where ",
            DynamKeyToString[k]]]]]],
       {}, {}];
      If[Total[Map[Length,First[key]]] > 0,
        (tmp="Incomplete key specification:";
         Scan[
           If[Length[key[[1,#]]]>1,
              tmp = StringJoin[
```

```
                tmp,
                "\n  ",
                DynamKeyItems[[#,1]],
                ": ",
                ToString[key[[1,#]]]]]&,
            Range[1,Length[DynamKeyItems]]];
          Print[tmp];
          DynamKey[key[[1]], {}, {}]),
         DynamKey[key[[1]], Columns[key], {}]]];

KeyDetails[k_DynamKey] :=
  Print[
    StringDrop[
     StringJoin[
       MapThread[
         (#1 <> ": " <> ToString[#2] <> "\n")&,
         {DynamKeyItems[[All,1]], k[[1]]}]],
     -1]];

MakeKey[x__Rule] :=
  Module[{key=Table[None,{Length[DynamKeyItems]}], tmp},
   Scan[
    (tmp = Position[DynamKeyItems, ToString[First[#]]];
     If[tmp=={},Message[MakeKey::unrecognized, First[#]]];
     key[[tmp[[1,1]]]] = Last[#])&,
    {x}];
   FillKey[DynamKey[key, {}, {}]]];
```

```
BasicQuery[k_DynamKey, query_String, OptionsPattern[]] :=
  (If[!ValueQ[DynamConn[k[[1,11]]]],
      DynamConn[k[[1,11]]] =
        DynamConnect[k[[1,11]],"worker_bee","workerbee"]];
   SQLExecute[
     DynamConn[k[[1,11]]],
     StringReplace[
       query,
       {"TABLE"->("[" <> k[[1,12]] <> "]." <> k[[1,13]]),
        "DB" -> ("["<>k[[1,12]]<>"]")}],
     ShowColumnHeadings->OptionValue[ShowColumnHeadings]]);
Options[BasicQuery] = {ShowColumnHeadings->False};

keyQueryStringReplacementsCols = {
  "step"->"maintable.step",
  "time_step"->"sim.time_step",
  "time"->"time=maintable.step*sim.time_Step",
  "residue_number"->"id.residue_number",
  "residue_id"->"id.residue_id",
  "atom_number"->"id.atom_number",
  "resnum"->"id.residue_number",
  "chain_id"->"id.chain_id",
  "icode"->"id.icode",
  "atom_name"->"id.atom_name",
  "atomnum"->"id.atom_number"};
keyQueryStringReplacementsBody = {
  "step"->"maintable.step",
  "time_step"->"sim.time_step",
```

```
  "time"->"maintable.step*sim.time_step",
  "residue_id"->"id.residue_id",
  "residue_number"->"id.residue_number",
  "chain_id"->"id.chain_id",
  "icode"->"id.icode",
  "atom_number"->"id.atom_number",
  "resnum"->"id.residue_number",
  "resid"->"id.residue_id",
  "atom_name"->"id.atom_name",
  "atomnum"->"id.atom_number"};
FixStringCols[s_String] :=
  Module[{tmp=StringSplit[s,"^^^"]},
    StringJoin[
      MapThread[
        If[OddQ[#1],
           StringReplace[
             #2,
             keyQueryStringReplacementsCols],
           #2]&,
        {Range[1,Length[tmp]], tmp}]]];
FixStringBody[s_String] :=
  Module[{tmp=StringSplit[s,"^^^"]},
    StringJoin[
      MapThread[
        If[OddQ[#1],
           StringReplace[
              #2,
              keyQueryStringReplacementsBody],
```

```
          #2]&,
        {Range[1,Length[tmp]], tmp}]]];


(* This needs to be fixed for the new time schema *)
KeyQuery[k_DynamKey, prefix_String,
         suffix_String, OptionsPattern[]] :=
  (If[!ValueQ[DynamConn[k[[1,11]]]],
      DynamConn[k[[1,11]]] =
        DynamConnect[k[[1,11]],"worker_bee","workerbee"]];
   SQLExecute[
     DynamConn[k[[1,11]]],
     FixStringCols[prefix] <>
      " from [" <> k[[1,12]] <> "]." <> k[[1,13]] <>
      " as maintable join [" <> k[[1,12]] <>
      "].dbo.simulation as sim on " <>
      "maintable.sim_id = sim.sim_id " <>
      " join [" <> k[[1,12]] <>
      "].dbo.ID as id on " <>
      "maintable.struct_id=id.struct_id " <>
      If[Position[k[[2]],"residue_number"] != {},
        " and maintable.residue_number=id.residue_number ",
        " "] <>
      If[Position[k[[2]],"atom_number"] != {},
        " and maintable.atom_number=id.atom_number ",
        " and id.atom_name='CA' "] <>
      "where maintable.sim_id=" <> ToString[k[[1,5]]] <>
      If[StringLength[suffix]<1,
```

```
          " ",
          " and " <> FixStringBody[suffix]],
      ShowColumnHeadings->OptionValue[ShowColumnHeadings]]);
Options[KeyQuery] = {ShowColumnHeadings->False};

Rows[k_DynamKey] :=
  (If[Or[SameQ[Last[First[k]],None],
         SameQ[Head[Last[First[k]]],List]],
      Message[Rows::unspecified, First[k]]];
   BasicQuery[k, "select count(*) from TABLE"][[1,1]]);

Columns[k_DynamKey] :=
  (If[Or[SameQ[Last[First[k]],None],
         SameQ[Head[Last[First[k]]],List]],
      Message[Columns::unspecified, First[k]]];
   Flatten[BasicQuery[k, "select top 0 * from TABLE",
                      ShowColumnHeadings->True]]);

queryStringReplacementsCols = {
  "step"->"maintable.step",
  "time_step"->"sim.time_step",
  "time"->"maintable.step*sim.time_Step",
  "residue_number"->"id.residue_number",
  "residue_id"->"id.residue_id",
  "atom_number"->"id.atom_number",
  "resnum"->"id.residue_number",
  "chain_id"->"id.chain_id",
  "icode"->"id.icode",
```

```
  "atom_name"->"id.atom_name",
  "atomnum"->"id.atom_number"};

Query[k_DynamKey,cols_String, OptionsPattern[]] :=
  Module[{tmp={}, rows=0, r=0, ps=OptionValue[RowBuffer],
          where = OptionValue[Where],
          order = OptionValue[OrderBy],
          grp = OptionValue[GroupBy]},
    rows=KeyQuery[k,"select count(*)", where][[1,1]];
    While[r < rows,
      AppendTo[tmp,
        KeyQuery[k,
          StringJoin[
            "select ^^^", cols,
            " from (select rank() OVER (order by ",
            StringReplace[order,
                          queryStringReplacementsCols],
            ")^^^ as r, ", cols],
          StringJoin[
            If[StringLength[where]>0, where, "1=1"],
            ") as a where r between ",
            ToString[r],
            " and ",
            ToString[r+ps-1],
            " order by r"]]];
        r += ps];
    Flatten[tmp,1]];
Options[Query] = {RowBuffer->25000, Where->"",
```

```
                    OrderBy->"step", GroupBy->None};

End[];
EndPackage[];
```

Appendix C

WAVELET DETAILS AND IMPLEMENTATION

The wavelet calculations were completed using Fourier transforms. The exact algorithm is given below in Mathematica code. The FourierConvolve[] function (below) performs a circular convolution of two signals using the convolution theorem. The ScaleWavelet[f, n, s] function takes a wavelet function f, an ideal length n, and a stretch s, and returns the wavelet as a signal of length n stretched by s. The function CWT[signal, f, s0, ds, S] takes the signal, wavelet function f, and variables s0, ds, and S, and returns the continuous wavelet transform of the signal with wavelets scaled according to the formula in Equation C.1.

$$s_k = s_0 2^{kds}; k \in \{0, 1, ...S-1\} \tag{C.1}$$

For the wavelets used in Chapter 3, the values $s_0 = 100$, $ds = 1/8$, and $S = 60$ were used, and all values were measured in picoseconds.

```
FourierConvolve[a_List, b_List] :=
  Times[Sqrt[Length[a]],
   InverseFourier[
    Times[
     Fourier[a, FourierParameters -> {0, -1}],
     Fourier[b, FourierParameters -> {0, -1}]],
    FourierParameters -> {0, -1}]];
ScaleWavelet[wltfunc_, n_Integer, scale_] :=
  Times[
```

```
    1/Sqrt[scale],
    Table[N[wltfunc[k/scale]],
          {k, -Floor[n/2], Floor[(n - 1)/2]}]];
CWT[signal_List, wltfunc_, s0_, ds_, S_Integer] :=
  Module[{n = Length[signal],
          scales = Table[s0*2^(k*ds), {k, 0, S - 1}]},
    Map[
     FourierConvolve[
      signal,
      RotateLeft[ScaleWavelet[wltfunc, n, #], Floor[n/2]]] &,
     scales]];
Morlet[t_] := Pi^(-1/4)*Exp[-t^2/2]*Exp[2*Pi*I*t];
Paul[t_] := 8*Sqrt[2/(35*Pi)]*(1 - I*t)^(-5);
```

For example, if the variable x contained the x-coordinate of an atom over time, then the continuous wavelet transform, using the Paul wavelet as described in Chapter 3, could be obtained with the command wlt=CWT[x, Paul, 105, 1/8, 60]. Thevalue of wlt[[1]] is then the wavelet coefficient vector with a scale of 105 ps while the value of wlt[[41]] is the wavelet coefficient vector with a scale of $105 \cdot 2^{41/8} \approx 3.66$ ns.

Once the wavelet coordinates are collected, significance testing is performed using the algorithm below. The WaveletSignificance[wlt, scales, x, pval, correction] function returns the wavelength of the frequency most strongly matched by the model described in this paper. The wlt parameter is the wavelet coordinates as generated by the CWT[] function while the scales variable should be a list of the scales calculated in CWT[]. The x parameter is the same as that passed to CWT[] while the pval is the minimum p-value acceptable for the significance test. Finally the correction parameter is the scale-to-wavelength

factor described in the paper (1.01 for Morlet, 1.389 for Paul).

```
WaveletSignificance[wlt_List, scales_List, x_List,
                    pval_, correction_] :=
  Module[
   {n = Length[wlt[[1]]],
    chival = (InverseCDF[ChiSquareDistribution[2], t] /.
              t -> (1 - pval))/2,
    res = Table[{0, 0}, {Length[wlt[[1]]]}],
    var = Variance[x],
    tmp, min},
   Scan[
    Function[{s},
     min = (0.00647*(correction*scales[[s]])^1.41344 +
                    19.7527)*chival;
     res = Table[
       (tmp = Abs[wlt[[s, k]]]^2/var;
        If[res[[k, 2]] < tmp && min < tmp,
         {correction*scales[[s]], tmp},
         res[[k]]]),
       {k, 1, n}]],
    Range[1, Length[scales]]];
   First[Transpose[res]]];
```

Note that in the case of a wavelet with no imaginary part, the fifth and sixth lines would be as follows.

```
chival = (InverseCDF[ChiSquareDistribution[1], t] /.
          t -> (1 - pval)),
```

VITA

Noah Charles Benson received his Bachelor's of Science degree from Purdue University in 2005, where he triple-majored in Mathematics, Computer Science, and Biology. In 2010 he earned his Doctor of Philosophy from the University of Washington's Division of Biomedical and Health Informatics in the Deparment of Medical Education and Biomedical Informatics. His publications include:

Benson, N. C. and Daggett, V. (2010). Graph theoretic evidence for a four step protein folding/unfolding proces In Preparation.

Benson, N. C. and Daggett, V. (2010). A graph theoretic approach to indexing protein dynamics. In Preparation.

Benson, N. C., Rutherford, K. and Daggett, V. (2010). Understanding the molecular basis of disease in single nucleotide polymorphism variants using wavelet analysis. In Preparation.

Benson, N. C. and Daggett, V. (2010). Wavelet analysis of protein motion. Biophysical Journal Submitted.

van der Kamp, M. W., Schaeffer, R. D., Jonsson, A. L., Scouras, A. D., Simms, A. M., Toofanny, R. D., **Benson, N. C.**, Anderson, P. C., Merkley, E. D., Rysavy, S., Bromley, D., Beck, D. A. C. and Daggett, V. (2010). Dynameomics: a comprehensive database of protein dynamics. Structure *18*, 423–35.

Benson, N. C. and Daggett, V. (2008). Dynameomics: large-scale assessment of native protein flexibility. Protein Science *17*, 2038–50.

Simms, A. M., Toofanny, R. D., Kehl, C., **Benson, N. C.** and Daggett, V. (2008). Dynameomics: design of a computational lab workflow and scientific data repository for protein simulations. Protein: Engineering, Design and Selection *21*, 369–377.

Benson, N. C., Whipple, M. and Kalet, I. J. (2006). A Markov model approach to predicting regional tumor spread in the lymphatic system of the head and neck. In American Medical Informatics Association Annual Symposium Proceedings pp. 31–35, American Medical Informatics Association.