9-2012

Towards Large-Scale Validation of Protein Flexibility Using Rigidity Analysis

Filip Jagodzinski
University of Massachusetts Amherst, jagodzinski@gmail.com

TOWARDS LARGE-SCALE VALIDATION OF PROTEIN FLEXIBILITY USING RIGIDITY ANALYSIS

A Dissertation Presented

by

FILIP JAGODZINSKI

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2012

Department of Computer Science

TOWARDS LARGE-SCALE VALIDATION OF PROTEIN FLEXIBILITY USING RIGIDITY ANALYSIS

A Dissertation Presented

by

FILIP JAGODZINSKI

Approved as to style and content by:

Ileana Streinu, Chair

Jeanne Hardy, Member

David Jensen, Member

David Kulp, Member

Jack Wileden, Member

Lori A. Clarke, Department Chair
Department of Computer Science

For Tyler

ACKNOWLEDGMENTS

I would like to thank my committee, for providing the guidance that made all of this work feasible. Ileana, especially, for allowing me to make – and learn – from my mistakes. I hope that I can demonstrate as much patience, and instill as much of a sense of dedication in my future students, as Ileana has in me. In addition, the work presented here is interdisciplinary at the intersection of biology, chemistry, and computer science. As such Jeanne's insights along with her insisting that the dissertation be as well developed and rounded in terms of chemistry as it is in computer science, are much appreciated. Many thanks are also due for David Jensen, David Kulp, and Jack. Their guidance in how to assess and measure the quality of my computational experiments, their insisting that I address the question of, "what is the overarching science that you are doing?," and their "where is the data to support this?" questions, are all much appreciated.

And then there are those people, whom I mention by first name, who did not directly influenced my work. Nonetheless they were there when I was formulating the core of my values, and they helped me to become the person who I am today. To them, thank you. They are Tyler, Benjy, and Chad.

And to those people who I met while in Amherst. TJ, Megan Olsen and Tim Wood, Dubi and his family. Hossein Baghdadi. Pete. Leeanne for being the fantastic Grad Program manager that she is. Audrey for her hard, honest guidance on teaching style and methodology. I've learned much from all of them, and my life has been enriched as a result. Also, thank you, Brenda, for the music. The piano. It is a part of me now. There's no going back.

Several of my former professors were instrumental in my decision to pursue a graduate degree. I hope they continue to inspire others as they have inspired me. Jacqueline van Gorkom is a fantastic professor of astronomy; Eric Gotthelf afforded me my first opportunity in conducting research. I have also fond memories in working with Frank Klassner and Boots Cassel. Thank you all.

And did I mention that I like to eat? And to spend time with my family in Connecticut? Over the entirety of my tenure at UMass, Ava, Julian, and Patrica unconditionally welcomed me into their home. Thank you for giving me a place to put my feet up when things in Massachusetts got especially crazy. Dziękuję bardzo.

I want to also thank my mom. Without her, none of this would be possible. And my grandmother Krystyna. She once told me (in Polish), "Do something that will help others." That commonly-referred-to-as-a cliché statement takes on a level of credence and assumes wisdom stature when cited by her, who is now 93 years old.

And to my current lab cohabitants, John and Ashraf. Naomi, also, whom I have worked with, complained with, laughed with, cursed at many things with, etc. She is a fantastic partner-in-crime.

And a special thanks goes out to Mícheál McDonald. He said, "Don't worry so much about pleasing each [dissertation] committee member. Just do good science. Then they'll have to say yes." Bold. And, moving forward, he said, "The day that you stop being a student, is the day that you start being a bad teacher." How true. Words to live by.

ABSTRACT

TOWARDS LARGE-SCALE VALIDATION OF PROTEIN FLEXIBILITY USING RIGIDITY ANALYSIS

SEPTEMBER 2012

FILIP JAGODZINSKI

B.Sc., COLUMBIA UNIVERSITY

M.Sc., VILLANOVA UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ileana Streinu

Proteins are dynamic molecules involved in virtually every chemical process in our bodies. Understanding how they flex and bend provides fundamental insights to their functions. At the atomic level, protein motion cannot be observed using existing experimental methods. To gain insights into these motions, simulation methods are used. However such simulations are computationally expensive.

Rigidity analysis is a fast, alternative graph-based method to molecular simulations, that gives information about the flexibility properties of molecules modeled as mechanical structures. Due to the lack of convenient tools for curating protein data, the usefulness of rigidity analysis has been demonstrated on only a handful of proteins to infer several of their biophysical properties. Previous studies also relied on heuristics to determine which choice of modeling options of important stabilizing interactions allowed for extracting relevant biological observations from rigidity analysis

results. Thus there is no agreed-upon choice of modeling of stabilizing interactions that is validated with experimental data.

In this thesis we make progress towards large-scale validation of protein flexibility using rigidity analysis. We have developed the KINARI software to test the predictive power of using rigidity analysis to infer biophysical properties of proteins. We develop new tools for curating protein data files and for generating biological functional forms and crystal lattices of molecules. We show that rigidity analysis of these biological assemblies provides structural and functional information that would be missed if only the unprocessed data of protein structures were analyzed. To provide a proof-of-concept that rigidity analysis can be used to perform fast evaluation of *in silico* mutations that may not be easy to perform *in vitro*, we have developed KINARI-Mutagen. Finally, we perform a systematic study in which we vary how hydrogen bonds and hydrophobic interactions are modeled when constructing a mechanical framework of a protein. We propose a general method to evaluate how varying the modeling of these important inter-atomic interactions affects the degree to which rigidity parameters correlate with experimental stability data.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

Proteins are dynamic biological molecules that bend and flex, and interact with other molecules, in order to perform their functions. We want to understand these motions, so that we can design medicines to therapeutically regulate these biomolecules. Unfortunately, there are no existing experimental methods that allow us to observe how proteins bend and flex on the atomic level. To gain insight into these motions, researchers use computationally intensive methods that rely on numerical techniques to simulate molecular motion. Pebble game-based rigidity analysis is an alternative, computationally efficient graph-based technique that determines the rigid components of a protein. Its usefulness in inferring structural and biophysical properties has been demonstrated on several molecules, but a large-scale study correlating protein rigidity parameters against experimental data has not been performed.

The goal of this thesis is to make progress towards a large-scale validation of protein flexibility using rigidity analysis. We develop new tools for curating protein structure data, we demonstrate KINARI-Mutagen for inferring which residues of a protein are critical, and we propose a general method for correlating rigidity parameters to experimental data in the form of $\Delta\Delta$G measurements.

1.1 Background and Motivation

Proteins are long chains of amino acids that fold into complex three dimensional shapes. They perform a myriad of functions in our bodies. Some have mechanical structural roles, others are involved in the immune response, while still others help

to translate and transcribe the genetic information in our chromosomes so that new molecules can be synthesized. Proteins perform their functions by flexing, bending, and interacting with other molecules. HIV-1 Protease, for example, plays an integral role in the maturation process of HIV. The protease undergoes a conformational change that is necessary for it to perform its function. So that medicines can be designed to regulate such proteins, we need an atomic-level understanding of where proteins flex, bend, and permit motion. Unfortunately, there are no existing experimental methods that allow us to observe atomic motion.

To gain insight into the motions of proteins on an atomic level, researchers rely on computational simulation methods. One such method is Molecular Dynamics (MD), in which the trajectories of a protein's atoms are calculated using numerical methods utilizing Newton's equations of motion. MD methods unfortunately have a very serious drawback: they are computationally intensive, and require hundreds, and up to tens of thousands, of computer processors [64].

Rigidity analysis of proteins is complementary to simulation methods such as MD. Its goal is not to simulate or predict a protein's motion, but to identify a molecule's flexible and rigid clusters of atoms. The input to rigidity analysis is a protein structure file. The Protein Data Bank (PDB), is a freely-accessible repository of protein structure data, whose entries of protein structures were solved using experimental methods such as X-ray crystallography. In preparation for rigidity analysis, important stabilizing chemical interactions among the atoms in the protein are identified. Atoms and their chemical interactions are used to construct a mechanical model of the molecule. Covalent bonds between atoms in the protein are represented as hinges in the mechanical model, and other stabilizing interactions such as hydrogen bonds and hydrophobic interactions are represented as hinges or as rigid bars. A graph is constructed from the mechanical model, in which each body in the mechanical structure is associated to a node, hinges between two bodies are associated to five edges

between two nodes, and bars are associated to edges. Depending on the type and strength of the chemical interactions that exists between atoms, a single or multiple edges – up to 6 – are placed between the two nodes in the graph representing the chemical constraints in the mechanical model. Efficient algorithms based on the pebble game paradigm [32, 46](explained in Section 2.3.3) are used to analyze the rigidity of the graph. The rigidity results permit inferring the rigid and flexible regions of the mechanical model, and hence the protein.

Rigidity analysis of proteins was first implemented in MSU-First [34, 33] and the first online tool was *FlexWeb* [79]. Since the late 1990s, the usefulness of rigidity analysis was demonstrated in inferring various structural and functional properties of proteins. Among these, rigidity analysis was used to identify the stability core of Rhodopsin [69], it was used to investigate the stability differences between temperature-sensitive proteins called thermophiles [25], and locations of rigid clusters of atoms have been correlated with the dynamics of well-studied proteins such as HIV-1 Protease [79].

Although tools such as *MSU-First* and *FlexWeb* were successful in demonstrating the usefulness of rigidity analysis in inferring structural and biological properties of a handful of proteins, the method has not been validated on a large dataset of proteins. There are several reasons why this is so.

Firstly, experimental methods such as X-ray crystallography produce the asymmetric unit, which is the smallest portion of a crystallized protein on which symmetry operations can be applied to reproduce the crystal form of the molecule. The asymmetric unit most often does not represent the biological functional form of a protein. To generate the biological assembly of a molecule, symmetry, rotation, and translation operations are applied on the atomic coordinates of the atoms in the asymmetric unit. If done by hand, generating the biological assembly of a molecule is a tedious process. And, automated tools for generating the biological form of a protein are dif-

ficult to design. This is because proteins exhibit a wide range of structural features, so the process is difficult to automate. Some biological forms of proteins are made up of a single chain of amino acids, while others are made up of multiple copies of the asymmetric unit. Others still are composed of multiple copies of different amino acid chains. And, the data files of protein structures solved using experimental methods are often incomplete, and may include water molecules along with non-standard amino acids or ligands. All of these are reasons why a high-throughput rigidity analysis of large datasets of proteins has not been performed.

Secondly, in the cases where rigidity analysis was used to infer biophysical properties of proteins, the interpretation of the rigidity results relied on biological insights of the studied molecules. Requiring in-depth knowledge of each protein that is studied makes rigidity analysis of large datasets of proteins impractical. Thus, a large-scale study to correlate rigidity properties of proteins with experimental data has not been conducted.

Finally, *MSU-First* and *FlexWeb* do not provide the user easily accessible options to designate how important stabilizing interactions should be modeled in the mechanical representation of a protein. Thus these tools cannot be used to easily perform large-scale studies to infer how changing the modeling of these interactions affects the rigidity results. Also, related to this is that there is often very sparse structural experimental data about any one protein. This is because experiments performed on a physical protein are generally expensive and time-intensive. A consequence of this is that there is no agreed-upon choice of how chemical interactions should be modeled in the mechanical framework of a protein.

The central theme of this thesis is to make progress towards a large scale validation study of rigidity analysis of proteins. As such, we have developed the KINARI software, which has allowed us to address each of the three shortcomings mentioned above. Namely, we have:

1. Developed curation tools, that permit the generating of biological forms of proteins from their asymmetric units. As a result, it is now possible to perform automated rigidity experiments on large datasets of biological structures. In Chapter 4, we demonstrate these new curation tools. Moreover, we show in several case studies that rigidity analysis of protein biological assemblies and of crystal lattices provides information about the biological form of a protein, that would not have been attained if only the asymmetric unit from a protein structure file were analyzed.

2. Developed KINARI-Mutagen, to help infer the locations of critical regions of a protein, that help to maintain its stability. This tool is a new application of rigidity analysis, whose predictive power does not rely on any biological insight of the protein that is being studied.

3. Investigated possible correlations between various rigidity parameters and experimental stability measurements of a dataset of 158 proteins. We have systematically varied how important stabilizing interactions were modeled in the mechanical representations of these biomolecules. We propose and demonstrate a general method to rank the choices of modeling of stabilizing interactions, based on how the rigidity results correlate to experimental data. Such a large-scale correlation study has not been done before.

1.2 Thesis Contributions

The contributions in this thesis build upon research results of Streinu and her collaborators. Also, the problems that we are trying to address – how should constraints be modeled, can critical residues be identified, and what is the rigidity of protein biological assemblies and crystals? – are not new problems. The contributions, however, are in the form of new approaches to these problems, and in the design of tools that

enable us to make progress towards a large-scale validation of rigidity analysis of proteins.

1.2.1 KINARI: An Infrastructure for Rigidity Analysis of Proteins

The first tools that implemented rigidity analysis of biomolecules [34, 33, 79] offered few options for curating protein structure data files. Thus large-scale studies of rigidity analysis of protein biological forms could not be performed, because the structure files had to be cleaned and curated by hand. Also, the choices of modeling of important stabilizing interactions were fixed. To provide an infrastructure to easily test if and how rigidity analysis can function as a predictive tool for inferring biophysical properties of proteins, we have developed KINARI-Web. It is a general, well-tested, versatile web server for rigidity analysis of molecular structures. It provides options for streamlining the curation of input protein data and for building protein biological assemblies and crystals. It relies on a mechanical model of a protein that is customizable by the user, it performs rigidity analysis of the mechanical framework, and it includes an interactive visualizer for exploring the rigidity results.

KINARI is an on-going, collaborative project in Ileana Streinu's Linkage Laboratory. Several people have contributed over the years. Professor Streinu supervises the entire project. Naomi Fox wrote the C++ code for the pebble game algorithms. Her dissertation work focuses on improving accuracy in the representation of proteins as mechanical frameworks. Yang Li, a former undergraduate honors thesis student at Smith College, integrated into KINARI the visualizer tools, which are based on Jmol, an open-source Java 3D viewer for chemical structures. In my work I focused on developing the curation tools, which permit a user to designate which portions of a protein structure file should be retained. Along with Naomi Fox, I developed the infrastructure for calculating rigidity metrics for proteins that are analyzed by the pebble game algorithm. All of the work leading up to and including the release of

the KINARI-Web server was published in 2011 [20]. The first public release of the underlying software library [21], which implements the pebble game algorithm and provides support for several mechanical models, was made available in 2012.

1.2.2 Rigidity Analysis of Protein Biological Assemblies

The structure data in a PDB file does not always represent the biological functional form of a protein. And, a PDB file contains only the asymmetric unit of a crystal, which is the smallest repeating unit, from which the structure of the crystal can be inferred using symmetry, rotation, and transformation matrices. Because early tools for performing rigidity analysis of proteins did not provide automated tools for generating a protein biological assembly from experimental structure data, a large-scale, high throughput, study of the rigidity properties of proteins has not been performed. In collaboration with Tiffany Liu and other undergraduate students, I have developed the tools for generating biological assemblies of proteins and crystal forms of a protein from its asymmetric unit. I have also performed benchmarking testing on over 25,000 protein structures, that relied on streamlining the automated curation tools that I developed with the rigidity analysis algorithms developed in the Linkage Laboratory.

To determine if new insights into protein flexibility can be obtained by performing rigidity analysis on biological assemblies and protein crystal structures, we have generated over 900 crystal lattices of various sizes for more than 300 proteins. Initial results indicate that analyzing protein biological assemblies and crystals provides structural and functional information that would be missed if only the asymmetric unit of a protein were analyzed. The initial results were presented in 2012 at the ICCABS conference [13], and a subsequent, extended version of the study has been submitted to an invited issue of BMC Bioinformatics [35].

1.2.3 KINARI-Mutagen: Identifying Critical Residues

Predicting the effect of a single amino acid substitution on the stability of a protein structure is a fundamental task in macromolecular modeling. Not only did we want to make progress towards a tool that permits fast evaluation of the effects of mutations that may not be easy to perform *in vitro*, but we also wanted to develop a new application of rigidity analysis that was not dependent on in-depth knowledge of any one protein to help infer from the rigidity results important structural features of the biomolecule.

Towards this goal, we have developed KINARI-Mutagen, that identifies critical residues based on the degree to which an *in silico* mutation to glycine affects the protein's rigidity. We show that the residues we identify as critical in the protein Crambin correlate with residue that are conserved for several homologues of the molecule, and that they would not have been identified by other methods. We also generate 48 mutants for 14 proteins, and compare our rigidity-based results with experimental stability measurements performed on the physical mutant proteins. Our rigidity analysis graph-theoretic approach at inferring the role of residues in stabilizing a protein's structure was presented at the Computational Structural Bioinformatics Workshop [36], and subsequently published in the Journal of Bioinformatics and Computational Biology [37].

1.2.4 Correlating Rigidity Parameters with Experimental Data

In preparation for performing rigidity analysis, a mechanical framework of a molecule is constructed, in which various stabilizing interactions among atoms are modeled according to their strength. No systematic study has been conducted as to what is the most plausible, chemically validated modeling scheme. This is in part because initial tools for performing rigidity analysis of proteins did not provide options for changing how important stabilizing interactions should be modeled. All

previous implementations relied on heuristics, which allowed for extracting relevant biophysical observations, but only for a limited set of proteins. As such, there is no agreed-upon choice of modeling of important stabilizing interactions so that rigidity parameters correlate with experimental data.

We seek a possible correlation between rigidity parameters and this experimental data. Towards this goal, we have used the KINARI software to systematically vary how stabilizing interactions are modeled. We propose a method to measure how rigidity metrics correlate with experimental stability data in the form of $\Delta\Delta$G measurements. Our general method is not dependent on a case-by-case analysis of the proteins that are being studied, but instead requires only experimental data, and rigidity results, for a dataset of molecules. This work has been accepted for presentation at ACM-BCB 2012, the ACM Conference on Bioinformatics, Computational Biology and Biomedicine [38].

1.3 Thesis Outline

This thesis is structured as follows. In Chapter 2, Background and Related Work, we provide a short review of relevant biology and chemistry principles, which are needed for explaining KINARI-Mutagen, and discussing biological results presented in subsequent chapters. Throughout this dissertation, rigidity results are compared and correlated with experimental data. The source of that experimental data is discussed in Section 2.1.4. In Section 2.2 we overview an experimental method and several simulation techniques for studying protein motions; we discuss the advantages and limitations of each. We then motivate the use of rigidity-based protein flexibility in Section 2.3. In Section 2.4 we present related work that relies on rigidity analysis to infer biophysical properties of proteins. Chapter 3 is a description of KINARI, the first contribution of this thesis; it is a collaborative project involving several people. There, we present our methodology of generating files of protein biological assemblies

and crystal structures, and give a system description of KINARI-Mutagen, developed to analyze mutant protein structures. In Chapter 4 we provide results of our studies in which we generated and analyzed over 900 crystal lattices of more than 300 protein structures. In Chapter 5 we discuss our results of using KINARI-Mutagen to identify critical residues that contribute to a protein's stability. In Chapter 6, we present our general method for correlating rigidity metrics with experimental data, and discuss results of systematically varying how hydrogen bonds and hydrophobic interactions are modeled in biomolecules. In Chapter 7, we conclude with a summary, and briefly discuss future work.

CHAPTER 2

BACKGROUND AND RELATED WORK

Proteins are dynamic molecules composed of amino acids. In order to understand how a protein functions, we need to know how it moves and interacts with other molecules. We cannot observe directly protein motion on an atomic level. However, we rely on experimental methods to resolve the structure of proteins, to identify the identities and locations of their atoms. Computational methods are used to simulate atomic motion.

In this chapter we review key biology and chemistry concepts relevant to this thesis. This includes a short review of protein structures, and a high-level introduction to experimental methods used to calculate a protein's stability. We discuss an experimental method, as well as several computational ones, for studying protein dynamics; we point out the strengths and limitations of each of them. We then review rigidity analysis, and motivate its use in studying proteins. We finish this chapter by highlighting a few of the research studies that have relied on rigidity analysis to infer biophysical properties of molecules.

2.1 Biology and Chemistry Primer

In order to understand a protein's function, we need to know how it interacts with other molecules, how it bends, and how it flexes. We cannot observe directly protein motion on an atomic level, so instead we rely on information from experimental methods to infer a protein's stability and possible motions. The information from measuring the thermodynamic properties of proteins is used to infer protein stability.

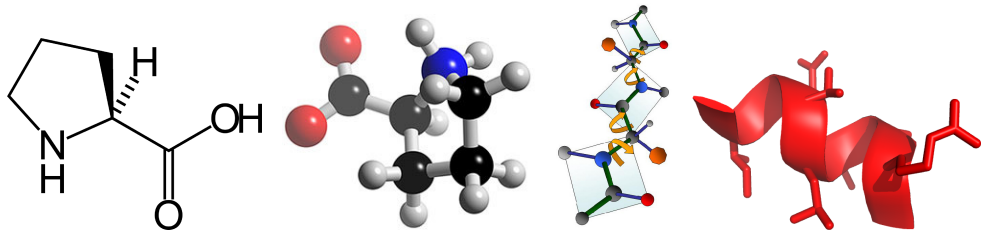
We present here a short review of protein structures and a discussion of protein stability. We review $\Delta\Delta$G, an experimentally derived stability measurement of proteins, which we rely on throughout this thesis.

2.1.1 Proteins: Polypeptide Chains of Amino Acids

Proteins are composed of amino acids joined end-to-end to form a chain. Amino acids are molecules containing an amine group, a carboxylic acid group, and a side chain that varies between the 20 different amino acids that occur in nature. Each of the 20 amino acids can be referred to by either its name, a three letter abbreviation, or a one letter abbreviation. For example, Alanine is referred to by its three letter abbreviation, **Ala**, or its single letter designation, **A**.

The chain of amino acids, which is called a polypeptide chain, folds into a three dimensional shape, the protein's tertiary structure. A single amino acid unit within a polypeptide chain is called a residue. Polypeptide chains vary in length, from as short as a few tens of residues, to as long as tens of thousands of residues. A segment of a polypeptide chain is designated by a sequence of letters, such as **AAVP**, which denotes a sequence of **A**lanine, **A**lanine, **V**aline, and **P**roline. The amino acids are held in place in a defined spatial arrangement by chemical bonds between atoms that are close in 3-dimensional space. Arrangements of segments of a polypeptide chain, or motifs, that occur frequently in nature are called secondary structures. α-helices and β-sheets are two such secondary structures. Regions of the protein that form compact, three-dimensional structures that often act independently of other regions of a protein are called domains. A protein may contain multiple domains.

Figure 2.1 shows a schematic representation of the amino acid proline, an atom model of proline, a polypeptide chain of three amino acids, and an α-helix. A schematic representation denotes the connectivity between atoms, but it does not describe the relative position of those atoms in three-dimensional space. A bar-and-



(a) (b) (c) (d)

Figure 2.1. Structure of proteins. Proline (schematic diagram shown in (a), bar-and-stick model shown in (b)). Proteins are molecules made up of amino acids that are joined end-to-end to form a polypeptide chain (c), which is the protein's primary sequence. Due to chemical interactions among the atoms, there is rotation allowed around some bonds, but not around others. Regular arrangements, or motifs, of parts of the peptide chain are called secondary structures. A cartoon drawing of an α-helix, an often-occurring secondary structure in proteins, is shown in (d).

joint model denotes the connections between atoms and their relative positions. The schematic representation of the polypeptide chain in Figure 2.1 denotes the protein's **backbone**, which is composed of alternating dark gray and blue atoms, designating the carbon and nitrogen atoms, that lie along the green bonds. An oxygen atom, red, is attached to one of the carbons, while the side-chain, of which there are 20 kinds occurring in nature, is designated by the orange hexagon. Hydrogen atoms are the small gray atoms attached to the carbons or the nitrogens. The specific orientation of atoms and their physical interactions allows for rotation along some of the bonds, but not all. It is this rotation around some bonds that allows a protein to flex, bend, and interact with other molecules.

A protein's tertiary structure is stabilized by chemical interactions that exist between atoms that are close in proximity. A chemical bond is the attraction caused by the electromagnetic force between opposing charges, either between electrons and nuclei of individual atoms, or as the result of a dipole attraction. Dipole attractions are caused by non-uniform distributions of positive and negative charges on various atoms. Examples of bonds include covalent bonds, hydrogen bonds, and disulfide

bonds. Hydrophobic interactions are another form of interaction that help to stabilize a protein's structure. The hydrophobics effect is the tendency of water molecules to exclude non-polar molecules, which leads to segregation of water and non-polar substances. Some of the bonds and interactions are strong, and are not easily broken, while others are relatively weak, and are known to continually break and re-form as the protein moves, flexes, and bends to perform its biological function. The amount of energy that it takes to break a bond is called the bond strength. Biologists measure bond strength in units of heat, measured in units of kilo-calorie (kcal). The strength of a bond or interaction is based on many factors, including the actual atoms that are involved. Sample bond strengths include 83-85 kcal/mol for a Carbon-Carbon single covalent bond, and 5-6 kcal/mol for a hydrogen bond [57] (where the involved bonding atoms are a hydrogen atom that is attached to a nitrogen, called the donor, and an oxygen, which is called the acceptor). The mole (mol) is a unit of measurement for the amount of substance or chemical amount. These two bond strengths are approximate; there is an observed range of bond strengths for a type of bond. Nonetheless there is a clear dichotomy between the strengths of different bonds.

Nearby water molecules, ions, as well as other, small compounds called ligands, interact with and affect the stability of a protein. These interactions all determine whether, and how, a protein bends, flexes, and interacts with other molecules. Figure 2.2 shows the representations of two conformations of HIV-1 Protease; the protein transitions from one conformation to another. It plays a crucial role in the maturation process of the Human Immunodeficiency Virus, HIV. The protease has two flap-like regions that close in on the interior of the protein, where a catalytic reaction occurs. It is because of this motion that the protease performs its function.

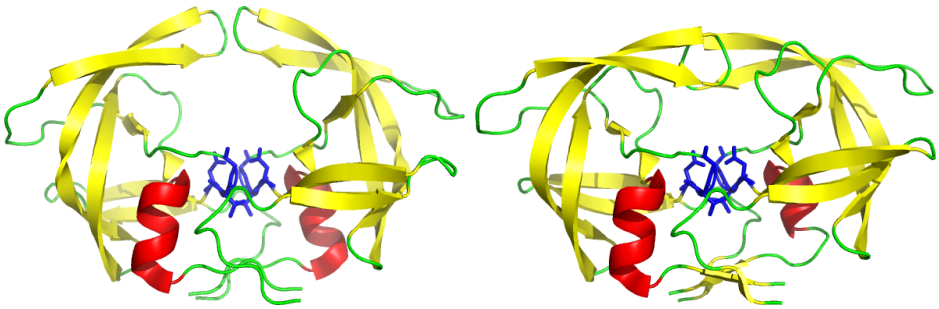


Figure 2.2. Conformations of HIV-1 Protease. Proteins flex and bend in order to perform their functions. HIV-1 Protease, which plays a crucial role in the maturation process of the virus that causes AIDS, transitions between several conformations. Movement of two flap regions renders the protein open (PDB file 1hhp, left) or closed (PDB file 1hvr, right). Images were generated by PyMol.

2.1.2 Determining Protein Structures

X-ray crystallography and protein Nuclear Magnetic Resonance (NMR) spectroscopy are two experimental methods used to resolve the structure of proteins.

In preparation for X-ray crystallography, a protein is purified and crystallized. Then, X-rays are passed through the crystal, which causes the beam to spread into many directions due to its interacting with the electrons in the atoms of the proteins in the crystal lattice. The angles and intensities of the spread-out beam generate a detectable diffraction pattern, which is used to recreate a three-dimensional model of the density of electrons in the crystal. From this electron density map, the mean positions of the atoms in the crystal are determined. 1958 was when Kendrew [39], *et al.* were the first to use X-ray crystallography to resolve the structure of a protein, sperm whale myoglobin. For that work, Perutz and Kendrew were awarded the Nobel Prize in Chemistry in 1962. Since that time, more than 72,000 structures have been resolved using X-ray crystallography.

However, X-ray crystallography does not give information about the dynamics of a protein, for one key reason: it is an averaging of the atom positions in all of the individual proteins crystallized in the crystal lattice. Proteins perform their functions

by transitioning from one conformation to another via a transition pathway that most often involves the protein assuming a high-energy, or highly unfavorable, conformation. High-energy conformers have short lifetimes, and proteins spend the majority of their times in energetically stable, or low-energy, conformations. Thus, the low probability of a protein assuming the high energy conformation among all of the proteins that are crystallized in the crystal lattice means that X-ray crystallography produces a "snapshot" structure that is the average orientation from among the ensemble of conformations of the proteins in the crystal lattice. The critical conformations, during which large dynamic motions occur that are crucial to the function of a protein, are thus "invisible" to X-ray crystallography.

In protein NMR experiments, an aqueous sample of a purified protein is placed in a magnetic field. Distinct atomic nuclei absorb electromagnetic radiation and resonate at different frequencies. The NMR resonance data is analyzed to afford information about the structure and dynamics of the molecule. One drawback of NMR is that it usually is limited to the study of small molecules, but recent advances [63] have allowed it to be used on proteins in sizes upwards of 82 kilo Daltons (kDa); the dalton is a molecular mass unit. With these advances, the dynamics of proteins such as human INPP5E (PDB file 2xsw, 357 residues) and heat shock cognate protein ATP hydrolytic activity (PDB file 1ngb, 386 residues) might be studied using NMR, but analysis of proteins any larger than these is currently beyond the reach of this method.

2.1.3 Experimental Methods for Measuring Protein Stability

Although we cannot directly observe protein motion at the atomic level, we can calculate various properties of a molecule that allow us to reason about the protein's stability and function. Thermodynamics is the study of energy conversion between heat and mechanical work. A protein performs work by virtue of its motion, and

indirectly we measure that motion by inspecting macroscopic variables such as temperature. Thermodynamic data, then, is indirect proof of the possible motions, and hence stability, of a protein.

The majority of small, single-stranded proteins exist predominantly in one of two thermodynamic states, the folded, or Native (N), state, and the Unfolded (U) state. These two states correspond to ensembles of molecular conformations. A protein may transition between the native and unfolded states according to a simple kinetic model:

$$U \rightleftharpoons N \tag{2.1}$$

where the rate from U to N is described by the rate constant k_f, and the transition from N to U is described by the rate constant k_u. The dimensionless rate constant for the equilibrium equation is given as follows:

$$K_{eq} = \frac{k_u}{k_f} = \frac{[U]_{eq}}{[N]_{eq}} \tag{2.2}$$

where the square brackets designate the concentrations of the native and unfolded proteins at equilibrium. Concentration of a protein in different states is experimentally calculated in several ways. One such method is circular dichroism (CD) spectroscopy [8], which measures the differences in absorption patterns of left and right-handed polarized light. The CD spectrum of a protein in the near ultraviolet spectral region is sensitive to certain aspects of tertiary and secondary structures. Thus, an analysis of the CD spectrum can be used to determine the presence of secondary structures. If a protein is known to have secondary structures such as α-helices, and if an analysis of the CD spectrum reveals that there are no α-helices in the sample, then the protein is assumed to be denatured, or unfolded. Using Equation 2.2, if a sample of proteins is calculated to contain a concentration of 80% folded proteins, then the ΔG value will be higher than if the concentration of folded protein is 60%. ΔG is

also referred to as the Gibbs free energy, the maximum amount of non-expansion work that can be extracted from a closed system, or the chemical potential that is minimized when a system reaches equilibrium at constant pressure and temperature.

The equilibrium constant is used to determine the conformational stability ΔG of a protein:

$$\Delta G = -RT\ln K_{eq} \tag{2.3}$$

where R is the universal Gas constant, 8.314 $\frac{\text{J}}{\text{K mol}}$, and T is the absolute temperature in Kelvin. In equation 2.3, ΔG is positive if the unfolded state is less stable (disfavored) relative to the native state. Therefore, ΔG characterizes whether a protein tends to favor the folded, native state, or whether it tends to favor the unfolded state.

If ΔG data is available for different conformations of the same protein, that information can be used to infer the relative stability of the two protein structures. The delta, or change, of the Gibbs free energy, $\Delta\Delta$G, determines how a change in the conformation or sequence of a protein affects the equilibrium constant of the protein species. For example, assume that you have two samples of the same protein, Sample 1 and Sample 2. The proteins in Sample 1 are known to perform their biological function, while the structure of the proteins in Sample 2 have been experimentally modified. Also assume that the following ΔG values have been calculated for both Sample 1 and Sample 2: $\Delta\text{G}_{\text{Sample 1}} = 4.12$ kcal/mol and $\Delta\text{G}_{\text{Sample 2}} = 3.27$ kcal/mol. The change of the Gibbs free energy, $\Delta\Delta$G, between the two proteins is defined and calculated as follows:

$$\Delta\Delta G = \Delta G_{\text{Sample 1}} - \Delta G_{\text{Sample 2}} = 0.85 \text{ kcal/mol}. \tag{2.4}$$

Therefore, given two samples, Sample 1 and Sample 2, of a protein for which there is available ΔG data, the experimentally derived $\Delta\Delta$G value can be used to determine

the relative stability of the proteins in the two samples. Using Equation 2.4, if $\Delta\Delta G$ is negative, then the proteins in Sample 1 are more stable than the proteins in Sample 2. In this example, because $\Delta\Delta G$ is positive, the proteins in Sample 2 are more stable than the proteins in Sample 1.

Throughout this thesis, the experimentally derived $\Delta\Delta G$ values for different conformations and/or mutated forms of a protein are taken as the "ground truth" values for the relative stability of two conformations of a protein, described in the next subsection.

2.1.4 The Protein Data Bank and the ProTherm Database

The Protein Data Bank [87] is a publicly accessible archive of experimentally determined structures of proteins, nucleic acids, and biomolecular complexes. More than 83,000 structures have been deposited into the PDB, of which approximately 72,000 were determined using X-ray crystallography. Proteins in the PDB are designated by four letter and/or number codes, such as 1hhp and 1hvr. Each PDB file contains the experimentally resolve identities and relative coordinates of the atoms in the asymmetric unit of a molecule. Throughout this thesis, we rely on structures from the PDB to conduct many of our computational experiments.

The *ProTherm* database [42] is a collection of numerical data of thermodynamic parameters, including $\Delta\Delta G$. Currently approximately 25,000 entries are in the database, for 733 unique proteins. Throughout this thesis, observations that are drawn regarding the relative stability of two proteins using rigidity metrics are correlated with experimentally derived $\Delta\Delta G$ values.

2.1.5 Lysozyme from Bacteriophage T4

The most-often referred to protein throughout this thesis is Lysozyme from bacteriophage T4. Lysozymes are enzymes that damage bacterial cell walls by catalyzing hydrolysis of 1,4-beta-linkages between different residues. Lysozyme is a well stud-

ied protein, and thermodynamics data for the wild type and for many of its mutated forms is available in the literature and the ProTherm[42] database. We used lysozyme thermodynamics data to evaluate and determine if the rigidity analysis results correlated with the known, calculated kinetic and stability properties of the protein. Of the more than 25,000 entries in the ProTherm database, 1,719 of them are for variants of lysozyme from bacteriophage T4, more than any other protein. It is because of this that the protein was used throughout this thesis. The rich dataset of thermodynamic properties of lysozyme from Bacteriophage T4 enables us to more easily perform correlation studies.

2.2 Methods for Studying Protein Motions

Proteins are dynamic structures. They fluctuate on the atomic level, and they transition between distinct states (Figure 2.2). We want to study their dynamics, so that we can understand how they function. Unfortunately, there are no existing experimental methods that allow us to watch, in real-time, the individual atoms moving within a protein. In this section we briefly review an experimental method, and several computational methods, used to study protein motions. Computational methods have the advantage over experimental methods in that they can describe protein dynamics completely. However, a fully accurate computational method would require a perfect force field function describing the protein-solvent system and the potential energies of all of the involved atoms. Existing force fields are used in molecular dynamics simulations. We describe molecular dynamics, the Go model, along with the *Morph Server* from the Gerstein Lab. We list the limitations of each technique. The short review of these methods is not meant to be comprehensive. Rather we review them here with the goal of placing rigidity analysis, described in Sections 2.3 and 2.4, within the context of other experimental and computational methods.

2.2.1 An Experimental Method for Visualizing Protein Motions

One recently developed exciting experimental method for studying protein dynamics is Fluorescence Resonance Energy Transfer (FRET) [14, 58]. It relies on fluorophores, which are chemical compounds that absorb certain wavelengths of visible light, and transmit or reflect others. When two fluorophores, one a donor in an electron excited state, and one an acceptor, are close in proximity, energy is transferred between the two. The efficiency of this transfer is inversely proportional to the sixth power of the distance between the donor and acceptor [44]. Measuring the efficiency of the energy transfer permits determining the distance between the two fluorophores. Diez, *et al.* [15] have used this technique to study F_0F_1-ATP synthase. They bound an acceptor fluorophore to the protein's b-subunit, and the donor to the γ subunit, and measured their intensities and their intensity ratio, over hundreds of milliseconds. This revealed three distinct states of the protein, and the donor-acceptor distances were used to identified the protein's molecular mechanism. However, one of the current limitations of single-molecule FRET is that only the change of the distance between a single pair of fluorophores is able to be measured. Thus even for this experimental method viewing the movement of individual atoms is not possible.

2.2.2 Simulating Protein Motions: Molecular Dynamics

Molecular Dynamics relies on numerical methods using Newton's equations of motion to calculate trajectories of a protein's atoms. It requires an empirical potential energy function. Its theoretical foundations were developed in the 1950s [4]. The first demonstration of an MD simulation, in 1977, was that of a 3.2 picosecond simulation of bovine pancreatic trypsin inhibitor [55]. However, many large domain motions that are important to the functioning of a protein occur on the microsecond to millisecond time ranges[31]. Even with progress in developing force fields for use in MD, protein dynamics on these long timescales are still beyond the practical reach

of MD simulations. Part of this is due to the large computational resources that are required, even for small simulations. Some of these limitations can be overcome, but doing so requires thousands of processors [64] and specialized algorithms, and even then simulations may require weeks of computation time.

2.2.3 Minimalist Models for Simulating Protein Motions

To address some of these limitations of MD, several minimalist models have been proposed. In the 1970s, Ueda *et al.* [84] and Taketomi *et al.* [75], modeled a protein as a chain of one-bead amino acids, and utilized a simplified force field of attractive and repulsive non-bonded interactions among the beads, to simulate a protein's motions. Since then, one-bead minimalist models have been developed for protein folding computational experiments [9], which require only a starting sequence, and are not dependent on the tertiary contacts that exist in the native state of the biomolecule.

2.2.4 Inferring Protein Conformations from Structure Data

In still another approach, the Yale *Morph Server* [41] produces three-dimensional animations of plausible domain motion pathways between two known conformations of a protein. Intermediate conformations are extrapolated from a trajectory between one conformation of a protein to another. Although this method is relatively quick, requiring usually no more than 30 minutes to generate a morph, there is no time-scale, kinetic, nor dynamic information that is used by the server. In recent extensions to the original *Morph Server*, energy minimization is used to calculate the intermediate frames and conformations of a protein, which produces results that are usually much better than morphs made by simple linear interpolation [17]. Still, the real intermediate conformations might be quite different from those that are hypothesized by this method, and the authors state that the morphs generate at worst a semi-plausible pathway between two submitted protein subunit conformations.

2.3 Rigidity Based Protein Flexibility

Rigidity analysis of proteins is an alternative to physics-based simulation methods, that instead analyzes a single static structure of a protein. Its goal is not to predict how a molecule bends and flexes, or to simulate a molecule's motions, but instead identify which parts of it are rigid. In this section, we give a short historical review of rigidity analysis, we describe how proteins are modeled as mechanical frameworks, and we describe pebble game rigidity analysis of these mechanical structures. The mechanical framework that is presented in this section is just one of several that have been developed. A description of another mechanical framework is available [20].

2.3.1 Rigidity of Bar-and-Joint Frameworks

The study of rigidity and flexibility of bar-and-joint frameworks was developed by 19^{th} century engineers attempting to analyze cross-bracing of steel structures. In 1864 James Clerk Maxwell [53] identified a simple counting rule to determine the rigidity of such structures. This counting rule was proven correct in 2 dimensions by Laman [45] in 1970, and subsequently was modified for the analysis of 3-dimensional structures, called body-bar-hinge framework [76]:

> **Theorem (Tay)** Let G be a graph with n vertices and m edges. G is the graph of a generic minimally rigid body-bar-hinge framework if and only if: any subset of n' vertices in G spans at most $6n$'-6 edges; and m=$6n$-6.

2.3.2 Mechanical Modeling of Proteins

In the KINARI software, body-bar-hinge 3-dimensional structures are used to model the mechanics of proteins. Atoms along with their covalently bonded neighbors form bodies. Covalent bonds between bodies are modeled as hinges, and other stabilizing interactions such as hydrogen bonds and hydrophobics are modeled as hinges or bars. In Figure 2.3 we show a schematic of a protein, and how the mechanical model is constructed.

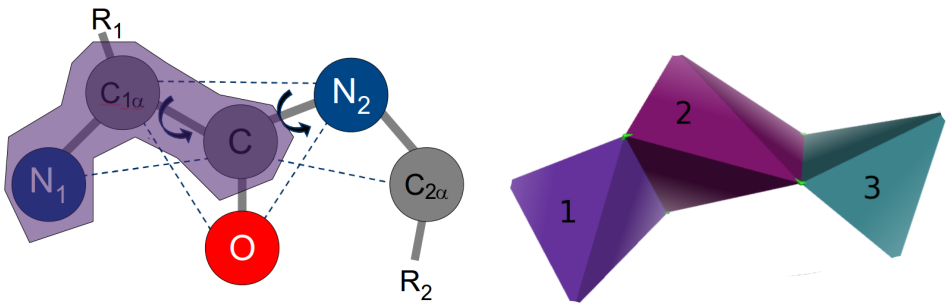


Figure 2.3. Modeling proteins as *body-bar-hinge frameworks*. Solid lines designate covalent bonds, and dashed lines represent distance constraints that arise due to angle constraints imposed by bonds; covalently bonded atoms form bodies (for example the purple region), shown on the left. R_1 and R_2 denote two residues. The *body-bar-hinge framework* for the protein (left) is shown on the right, where body 2 is composed of $C_{1\alpha}$, C, O, and N_2, while body 3 is composed of C, N_2, and $C_{2\alpha}$. A hinge between two bodies allows for a one-degree-of-freedom rotation of one body about the other, along the hinge axis.

2.3.3 Pebble Game Rigidity Analysis

Because the pebble algorithm is run on a graph, the mechanical framework must be associated to a set of nodes and edges. From the body-bar-hinge framework, the graph is constructed in the following fashion: each body is associated to a distinct node, each bar in the mechanical framework is associated to an edge between two nodes, and hinges in the mechanical framework are associated to 5 bars between the two nodes that represent the rigid bodies (Figure 2.4). Because two bodies in three-dimensional space have six non-trivial degrees of freedom between them (three translations along and three rotations around the x, y, and z axes), placing five bars between two bodies is equivalent to retaining just one of those six degrees of freedom, which represents the mechanical behavior of a hinge. An efficient pebble game algorithm [34] decomposes this graph into clusters which correspond to rigid components in the framework.

The algorithm starts with 6 pebbles on each vertex of the associated graph, and reasons about the edges one at a time, and accepts or rejects them. To be accepted, an edge must have at least 7 pebbles distributed somehow on its two endpoints. If not

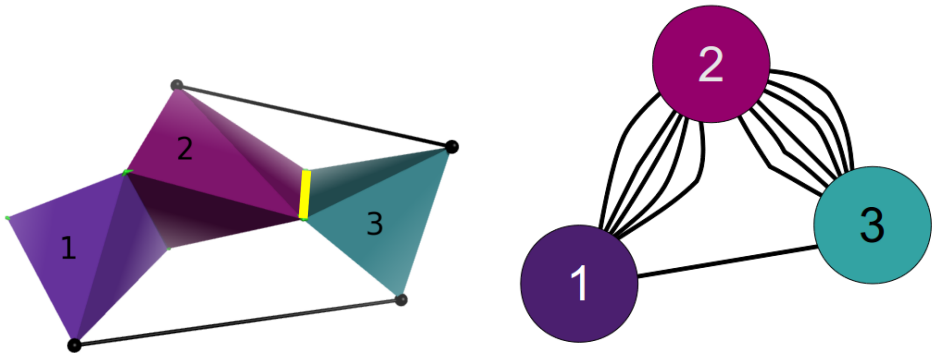


Figure 2.4. Abstract mechanical framework of a protein's structure. An abstract model illustrates how bars would be placed between bodies, if hydrogen bonds or other stabilizing interactions existed between them (left). The graph (right) is built by associating each rigid body to a node, and hinge to 5 bars, and a bar to each edge. Here bodies 2 and 3 on the left are represented as having 6 edges between the nodes in the graph that represent the two bodies, 5 edges for the hinge (depicted as yellow in (a)), and 1 for the bar.

enough pebbles are present, they are collected using a depth-first search approach. An accepted edge consumes one pebble. As more and more edges are accepted, they are combined into rigid components. The algorithm ends when all edges have been considered. A formal proof of correctness for this algorithm can be found in [46].

2.4 Rigidity Based Protein Flexibility: Related Work

Rigidity analysis of proteins has been demonstrated on a handful of proteins, and has been used to infer biophysical properties of several biomolecules. In this section we review a few of these published results.

Rigidity analysis of protein structures was first introduced in the work of Thorpe, *et al.* [79]. They studied different states of HIV-1 protease and showed that the rigid clusters in open and closed conformations of the protease are correlated with the known mechanical properties of the cantilever flaps of the molecule [80].

Rader *et al.* [69] simulated the thermal unfolding of Rhodopsin, a trans-membrane receptor, by performing a *rigidity dilution analysis* using the FIRST software [80]. In

this method, hydrogen bonds are removed from the molecular model one after another, from weakest to strongest, and rigidity analysis is performed on the model after each removal. A "folding core" is identified when there exists only one rigid cluster with at least three residues of two or more secondary structures. The computed core was correlated with experimental results, and confirmed via a visual inspection of the protein using insights of its physical properties.

In a comparative study of the rigidity analysis of 62 protein structures from six different protein families, Wells *et al.* [86] demonstrated that the main-chain rigidity of a protein is very sensitive to small structural variations. In that study, Wells concluded that the modeling of hydrogen bonds needs to be chosen carefully so that specific hypotheses about the rigidity of particular proteins can be formed.

Recently, Fox, *et al.* [22], used a benchmark dataset of 32 PDB structure files to validate the modeling in KINARI against a dataset that was analyzed using the Gerstein Lab's *RigidFinder* algorithm [1]. Fox introduced a metric called the cluster decomposition score to compare KINARI's rigidity results against Gerstein's structural predictions. They found that the sensitivity of the cluster decomposition score is dependent on the choice of the hydrogen bond energy cutoff value, which designates a threshold at which these bonds are retained in the molecular model.

Ivet Bahar, *et al.* [68] have relied on elastic network as well as constraint network models of freely rotating rods to predict protein folding nuclei. Both methods were verified against data that was attained from native state hydrogen-deuterium exchange experiments. Hydrogen-deuterium exchange gives information about the solvent accessibility of various parts of a molecule, and thus the tertiary structure of a protein. In these studies, the role of specific interactions in protein folding was also investigated.

Radestock *et al.* [25] used a dilution analysis to study the different states of thermophilic and mesophilic protein homologues. A mesophile is an organism that func-

tions best in a moderate temperature environment, while a thermophile thrives best at relatively high temperatures. In that study, macroscopic properties of the proteins were correlated with rigid cluster sizes, using Cluster Configuration Entropy (CCE) [82]. CCE is a function of the probability that a vertex in the mechanical model is part of a cluster of size s. To compute CCE, a normalized cluster number, n_s is defined as the number of clusters of size s divided by the total number of vertices in the mechanical model. The probability that a vertex belongs to an s-cluster, w_s, and the CCE value of the entire mechanical model are given as the following:

$$w_s = \frac{sn_s}{\sum_s sn_s} \tag{2.5}$$

$$CCE = -\sum_s w_s \ln w_s \tag{2.6}$$

For two conformations of a protein, the one with the higher CCE value is more disordered, and hence is more unstable. Radestock *et al.* used the CCE descriptor to show that in approximately 70% of the proteins in their dataset, the thermophilic molecules transitioned from rigid to flexible at higher temperatures than the mesophilic homologues.

AJ Rader, *et al.* have used rigidity theory to relate the constraint network of proteins to that network's deformability [70]. A protein's transition state can be determined from the inflection point in the change in the number of independent bond-rotational degrees of freedom (floppy modes) of the protein as its mean atomic coordination decreases. Rader was concerned with the calculation of a universal property of proteins that relied on a dilution analysis. Rader performed multiple rigidity analyses, as many as there were hydrogen bonds in the protein.

CHAPTER 3

KINARI: AN INFRASTRUCTURE FOR LARGE-SCALE RIGIDITY STUDIES OF PROTEINS

Rigidity analysis of proteins was initially implemented in MSU-First [34, 33] and the first online tool was *FlexWeb* [79]. Those applications had several limitations, including a need to curate protein data by hand, and the choice of modeling of important stabilizing interactions was fixed. To address several of these limitations – and to provide an infrastructure for performing large-scale validation studies of rigidity analysis of molecules – we have developed KINARI-Web. It is a second generation free online server for protein rigidity analysis, that implements a variation of the pebble game algorithm that was developed by Jacobs and Hendrickson [32, 46]. KINARI-Web is available at `http://kinari.cs.umass.edu`.

3.1 Introduction and Background

Proteins interact with organelles, other proteins, ligands, and ions. Thus, performing rigidity analysis of proteins outside of the context of their neighbors might cause important structural or functional information to be missed.

The structural data that is deposited into the PDB is the asymmetric unit of a protein's crystal, which may or may not be the same as the biological assembly (or functional form) of the protein. To help facilitate the rigidity analysis of biological assemblies of proteins, we have developed BioAssembly. It is a feature of the curation portion of the KINARI [20] software, that permits the building of biological assemblies or their sub-components, from the asymmetric unit in a PDB file. We have also

developed KINARICrystal, that allows a user to generate a structure file representing the crystal lattice of a protein, in which many instances of a molecule are nestled side-by-side, as they do in the crystal used for X-ray crystallography experiments. The KINARICrystal tool was developed in collaboration with post-baccalaureate students working with Ileana Streinu, while the BioAssembly tool was designed with the help of Tiffany Liu, an honor's thesis student at Smith College, advised by Streinu.

In addition, KINARI-Lib V1.0 [21] has been released. It is a C++ library that implements the pebble game algorithm and provides support for body-bar-hinge and bar-joint mechanisms.

3.2 System Description

KINARI-Web is comprised of two phases: (1) data input and curation, and (2) rigidity analysis and visualization.

Curation is composed of four steps. In the first step, a protein structure file is either uploaded to the server, or the user designates a four-character protein code, and KINARI fetches that structure from the PDB. KINARI-Web lists the models, chains, ligands, water molecules, etc. that are included in the uploaded PDB file, and the user selects the ones to be retained. In the second step of curation, hydrogen atoms are added using the *Reduce* software. In the third step, stabilizing interactions such as single and double covalent bonds, resonance bonds in peptide units, and disulfide bonds are calculated. Hydrogen bonds are determined using the *HBPlus* software, and hydrophobic interactions are identified using the algorithm described in in the ASU-FIRST User Guide [11]. Each covalent, resonance and disulfide bond is assigned an energy, in kcal/mol, that is determined from a table of average energies for each bond type and pair of atoms that are involved in the interaction. The energy of a hydrogen bond is computed by the KINARI software using the Mayo energy function [54]. In the final curation step, the computed chemical interactions that exist between atoms

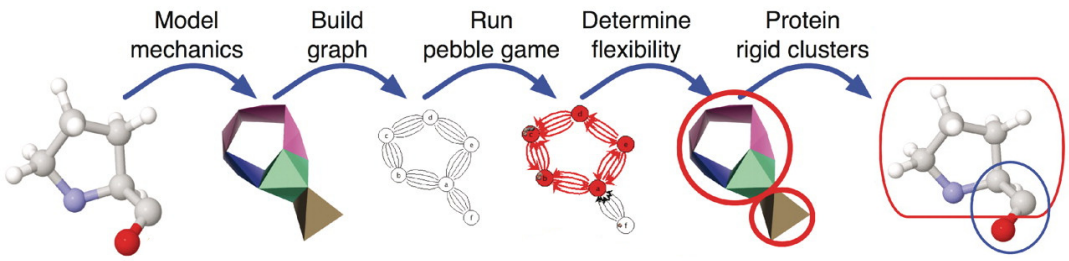


Figure 3.1. Curation, modeling, and rigidity analysis in KINARI. We model proline as a mechanical structure, from which an internal multi-graph is built. A pebble game algorithm calculates rigid components in the multi-graph, from which the protein's rigid clusters are inferred. Image adapted from [20].

in the PDB-formatted input file are presented to the user, who can designate which of them should be retained, and which should be removed. Chemical interactions not identified by KINARI are supplied as user-defined constraints.

In the second phase of KINARI-Web, a user designates how different chemical interactions should be modeled in the mechanical framework of the protein. This is a novel feature of KINARI, not available in *FlexWeb*. KINARI-Web then generates a mechanical model of the protein, and builds the associated graph, that is used as input to a pebble game algorithm. The output of the pebble game is then interpreted in terms of clusters of atoms within the protein. The entire process of building the mechanical framework, generating the association graph, and inferring the rigidity results from the output of the pebble game algorithm are shown in Figure 3.2, reproduced from [20].

After rigidity analysis has been performed, the KINARI-Web visualizer is used to explore the rigidity properties of the protein. The input biomolecule is displayed along with its calculated rigid regions. In Figure 3.2, left, we show the rigidity results of HIV-1 Protease (PDB file 1hvr). KINARI-Web's visualizer options enable a user to investigate the rigidity results. Among these options, a user can display certain clusters while hiding others, zoom in to investigate specific parts of the protein, or

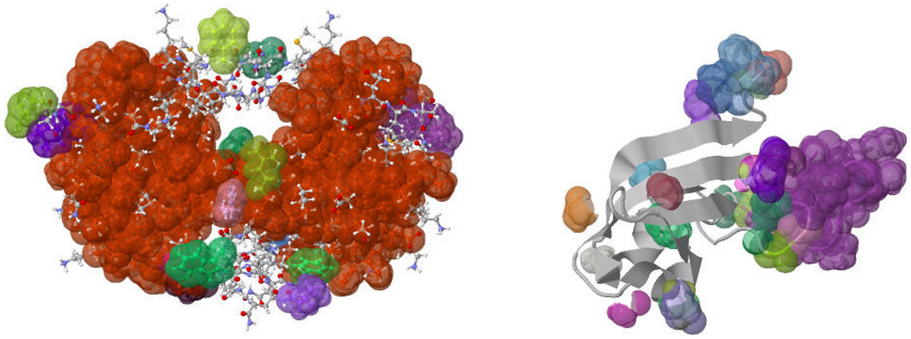


Figure 3.2. Visualizing rigid cluster of HIV-1 Protease. Groups of colored atoms indicate rigid clusters. For HIV-1 Protease (PDB file 1hvr), most of the atoms are in a dominant rigid cluster (orange). The options in KINARI-Web's visualizer enable a user to investigate the rigidity results. For the crystal structure of Diedel, a marker of the immune response of Drosophila melanogaster (PDB file 3zzo), the largest cluster has been hidden, and cartoon rendering is enabled.

display bonds that act as hinges between rigid bodies. Many of these visualizer features are novel to KINARI-Web. In Figure 3.2, right, we show the rigidity results of the crystal structure of Diedel, a marker of the immune response of *Drosophila melanogaster* (PDB file 3zzo); all rigid clusters containing more than 5 atoms are displayed, the largest rigid cluster has been hidden from view, and cartoon rendering has been enabled.

3.3 Software Profiling and Testing

Two types of software profiling were performed, to determine the limitations of KINARI-Web: (1) back-end software testing, and (2) front-end testing of the web interface. Both types of profiling were performed on the server that currently houses the KINARI webpage. The back-end software testing was performed on more than 25,000 proteins retrieved from the Protein Data Bank (PDB), and was conducted to test each of the four curation steps and the rigidity analysis phases of KINARI. The profile testing of the front-end graphical user interface was performed to evaluate the visualizer and front-end features of KINARI-Web.

For the back-end profile testing, the following five components of KINARI were tested: (1) Curation - Cleaning a PDB file, (2) Curation - Adding Hydrogen Atoms, (3) Curation - Calculating Interactions, (4) Curation - Removing Unwanted Interactions, and (5) Performing Rigidity Analysis. Several separate PDB data sets were used, each with different sized proteins. Also, the curation and modeling options were varied across several of the profile runs to test how the different components of KINARI and KINARI-Web perform under various combinations of options that a user might designate. For each data set, the PDB IDs that were used are posted on the KINARI website. The log file, as well as plots of the average run time of each curation and rigidity analysis step vs. the number of residues, are also listed there. In Figure 3.3, we show the average run time for the different phases of KINARI for proteins of various sizes.

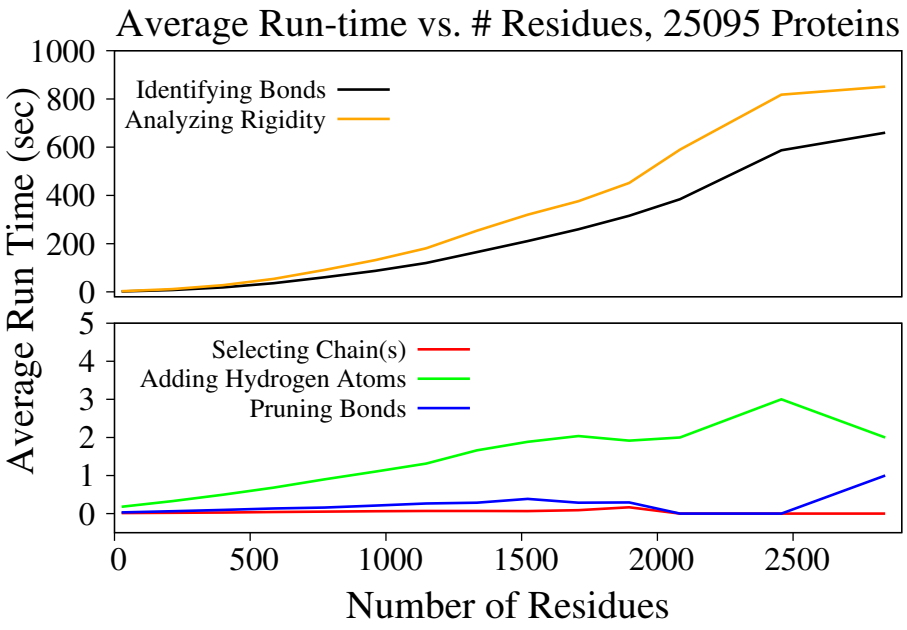


Figure 3.3. Profiling results for 25,000 proteins. Run-times were recorded for each of the four curation steps and the rigidity analysis phase, for proteins of various sizes.

For the front-end testing, each step of KINARI-Web was invoked, including downloading, cleaning/curating a PDB file, performing rigidity analysis, and generating output files necessary to render the protein and the results in the visualizer.

3.3.1 Generating Biological Assemblies

Protein structure files contain atom coordinates for the **asymmetric unit** of a protein, i.e. the minimal set of atoms necessary to reproduce the complete protein biological assembly and crystal which was analyzed with X-ray diffraction. A PDB file has information on how to create the **biomolecule**, the functional biological unit, and **unit cell**, the repeating unit of the crystal. A **unit cell vector** is a vector from the origin of the coordinate system to a lattice point of the crystal [71]. Three unit cell vectors are needed to describe a unit cell. A symmetry operation is a transformation operation (represented as a matrix) acting on a protein that produces a copy of it, possibly translated and rotated. The **space group** referred to in the PDB file is a combination of symmetry operations and a lattice specified by the unit cell vectors.

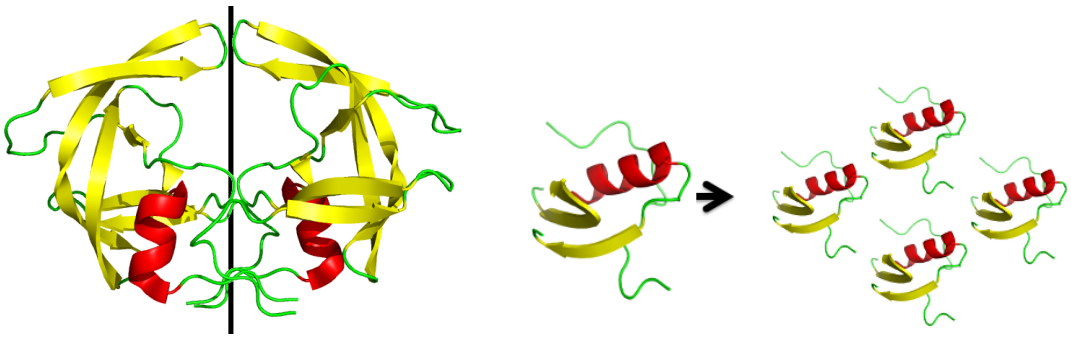
Over 87% of the protein structures deposited in the Protein Data Bank have been solved with X-ray crystallography. Protein function is correlated with flexibility, which motivates us to analyze the rigidity properties of the biological forms of proteins. However, protein flexibility studies using rigidity analysis have been performed until now primarily on *individual* asymmetric units (the smallest part of a protein that is needed to re-create the protein's biological functional form) from the data available in the PDB. For some proteins the asymmetric unit is identical to the biological assembly. However, for many proteins, especially those determined via X-ray crystallography, the asymmetric unit is different from the biological assembly. Depending on how the protein was crystallized, the relationship between the asymmetric unit and the biological assembly can vary from protein to protein. Some biological assemblies are composed of many copies of the asymmetric unit.

The PDB file contains only the atomic coordinates of the asymmetric unit, and it is natural to expect that its rigidity analysis may not always reflect the flexibility properties of the biological form. One extreme example is a viral capsid: the icosahedral type is composed of 60 repeating units, and each one may contain several monomeric units. To gain information on the flexibility of the virus would require building the entire assembly, but this would be a very large molecule, and so far, no existing software automatically performs this task.

In order to generate the biological assembly form of a protein, the asymmetric unit in a PDB file must be rotated, copied, translated, etc. The number of asymmetric units that make up a biological assembly varies from protein to protein. In some case, the asymmetric unit is the biological form, but in others, two, three, or many more asymmetric units, arranged uniquely in relation to each other, form the biological assembly. A transformation matrix in the PDB file details how chain(s) in the asymmetric unit need to be processed to form the biological form of the protein.

For the purposes of rigidity analysis, building just portions of the biological assembly may be useful. For instance, analyzing increasingly larger portions of a biological assembly may provide insight into the evolution of its flexibility as it builds up from its subunit components, or as it decomposes into its smaller subunits. From a computational point of view, generating a PDB file for the biological unit may not always be possible in the PDB format, which is what KINARI currently supports. Indeed, this format can only accommodate up to 99,999 atoms and up to 36 chains, and many large biological units such as viruses easily exceed these limits.

For generating a protein's biological assembly, we used the translation and rotation matrices included in the header of a PDB file. We applied transformation operations that were given as matrices in REMARK 350 of each structure file on each atom of the asymmetric unit. Each matrix contains a 3D rotation matrix and three translation



(a) Symmetry rotation of one of the homo-dimer halves of HIV-1 Protease about a line a symmetry

(b) Translating the protein unit in PDB file 3ovo (left) would generate a possible crystal lattice structure (right), composed of several asymmetric units

Figure 3.4. Symmetry operations in proteins. The 3 symmetry operations allowed in proteins due to chirality are rotation (a), translation (b), and screw rotation (not shown). These do not compromise the handedness of the alpha-helix.

vectors (one for each axis). The listing of secondary structures also was updated with references to the newly generated atom coordinates.

3.3.2 Building Crystal Lattices

The symmetry operations and space groups used to recreate a crystal lattice from an asymmetric unit have their foundation in mathematical crystallography. There are seven types of symmetry operations, each of which has a specific matrix representation, but only three (rotation, translation, and screw rotation) are allowed in proteins due to chirality; that is, four symmetry operations are not allowed because they would change the handedness of an alpha-helix of a protein [71] (Figure 3.4).

Matrix representations of two symmetry operations are shown in Figure 3.5. On the left, a matrix which has only ± 1 values on the diagonal and zeroes elsewhere would transform the original protein data by 180° rotations and reflections about one of the three orthogonal axes. The general structure of a transformation matrix using arbitrary angles is shown on the right.

To generate a crystal, three linearly independent translations are required. If the translations are represented by three vectors, a, b, and c, then all lattice points

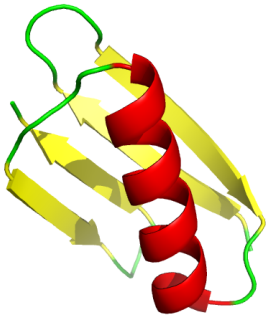
$$\begin{bmatrix} \pm 1 & 0 & 0 \\ 0 & \pm 1 & 0 \\ 0 & 0 & \pm 1 \end{bmatrix} \qquad \begin{bmatrix} cos\ \theta & -sin\ \theta & 0 \\ sin\ \theta & cos\ \theta & 0 \\ 0 & 0 & \pm 1 \end{bmatrix}$$

Figure 3.5. Matrices for generating crystal structures from an asymmetric unit. These matrices can assume one of several forms, based on the specific transformation that is required to reproduce the repeating crystal units. Depending on the combination of integer ± 1 values on the diagonal (left), these can be either a rotation by 180°or a reflection. If a symmetry operation uses rotations with angles other than 180°, the matrix has the form shown on the right, with corresponding values of *sin* and *cos* substituted.

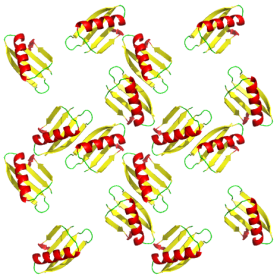
are generated by linear combinations of the vectors with integer coefficients [72]. There are 14 types of Bravais lattices, categorized into seven crystal groups: cubic, tetragonal, rhombohedral, orthorhombic, monoclinic, triclinic, and hexagonal.

A **space group** is a combination of one of the 14 lattice types and one to three symmetry operations. While there are 230 distinct space groups, proteins only crystallize into only 65 of them, due to chirality. For example, the protein in PDB file 1onj [49] crystallizes into space group P 4_1 2_1 2, with eight symmetry operations; thus it will have eight asymmetric units in the unit cell. The P and initial 4 indicate a primitive tetragonal lattice type. 4_1 indicates a four-fold screw axis: a 90^o rotation, followed by a translation of 1/4 of the c unit cell vector length. 2_1 indicates a two-fold screw axis: a 180^o rotation along with a translation of 1/2 of the a unit cell vector length. 2 indicates a 180^o rotation.

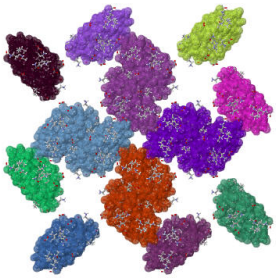
To create the asymmetric unit and crystal lattices, we applied symmetry operations that were given in REMARK 350 of each structure file on each atom of the asymmetric unit. We built the unit cell with *supercell.py* [28], a Python script. It retrieves the three unit cell vectors from the CRYST1 line of a PDB file, and generates the unit cell by applying the space group symmetry operations on the asymmetric unit. We built the crystal by translating the unit cell in the direction of the three unit cell vectors. For self-checking, we generated the crystal structures using two meth-



(a) Cartoon rendering of asymmetric unit of PDB file 2on8



(b) Crystal made up for 4 unit cells, each composed of asymmetric units.



(c) Rigidity analysis of a protein crystal. Some neighboring asymmetric units collapse into a large rigid body.

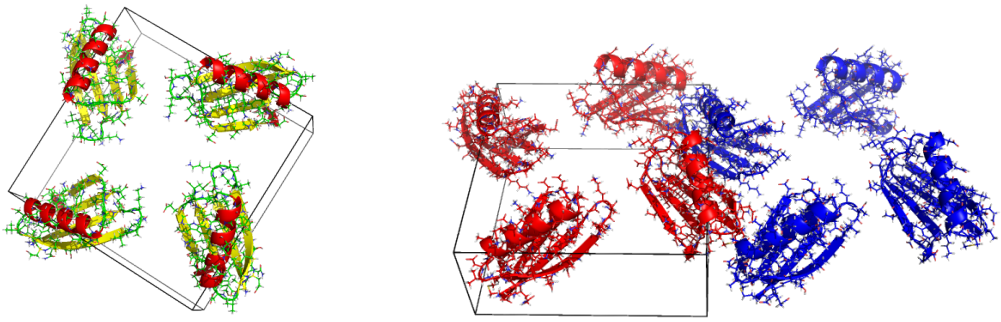
Figure 3.6. Rigidity analysis of a crystal lattice of the B domain of the streptococcal protein G. For the gbeta1 domain (PDB file 2on8), the cartoon rendering of the asymmetric unit is shown in (a). The crystal lattice of 2×2×1 unit cells of the gbeta1 domain as generated by KINARICrystal is shown in (b). Its flexibility properties are visualized in (c). Different colors designate distinct rigid clusters.

ods: a custom-built interface to the python script *supercell.py* [28], and an in-house implementation. An example, illustrating the Immunoglobulin G-binding protein G (PDB file 2on8) is shown in Figure 3.7. Figure 3.6 shows the asymmetric unit and small crystal of protein 2on8; different colors indicate different rigid clusters.

3.3.3 Generating *in silico* Mutant Protein Structures

Predicting the effect of a single amino acid substitution on the stability of a protein structure is a fundamental task in macromolecular modeling. We have extended KINARI to generate mutant protein structures and analyze their rigidity. We present here the first release of this new tool, KINARI-Mutagen. Its ultimate goal is to identify destabilizing mutations. This first version performs an *in silico* mutation to a glycine, which we call an *excision*.

The KINARI Mutation Engine performs a simple computational mutation, where a residue is converted to a glycine. For the purpose of performing the rigidity analysis,



(a) Generated unit cell, colored by secondary structure

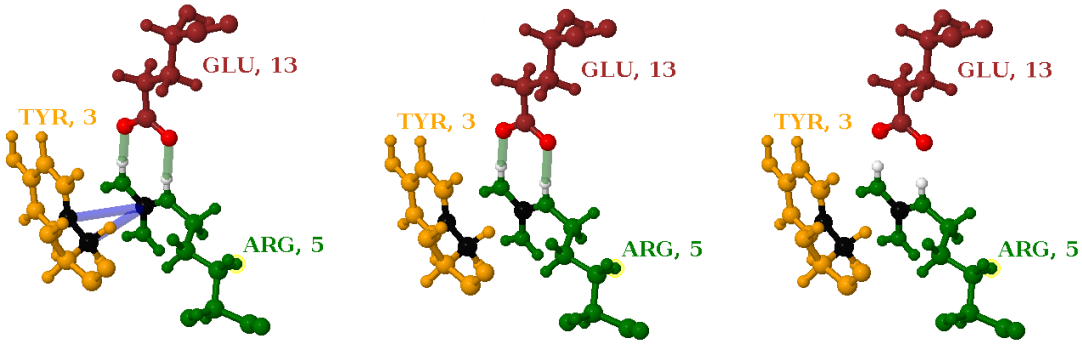
(b) Generated crystal lattice, made up of 2 unit cells (red and blue); bounding box for right unit cell not shown

Figure 3.7. Generating crystal lattices from asymmetric units. From the B domain of the streptococcal protein G (PDB file 2on8), we generated the unit cell (a) from the asymmetric unit. In (b), the unit cell is translated using the unit cell vectors to generate a 2×1×1 crystal lattice.

it is not necessary to alter the positions of, or remove, atoms. Instead, it suffices to remove the side-chain's hydrogen bonds and hydrophobic interactions from the protein's molecular framework. This functions in our model like the removal of a side-chain. Subsequent versions of the Mutation Engine will permit increasingly advanced mutation functions. Because rigidity analysis is efficient, many generated mutant protein structures can be analyzed quickly.

We demonstrate the excision process on a fragment of human α-defensin 1 (Figure 3.8). When excision is performed on residue 3, the hydrophobic interactions between it and residue 5 are removed from the molecular framework (Figure 3.8(b)). When excision is performed on residue 5 (Figure 3.8(c)), then the hydrogen bonds and hydrophobic interactions that it engages in are removed.

KINARI-Mutagen investigates how different residues affect the rigidity and stability of a protein. Analyzing a protein involves four phases: 1) downloading and curating a PDB file, 2) performing excision to generate mutants, 3) analyzing the rigidity of each mutant, and 4) aggregating the results to help the user identify critical residues. For step 1, KINARI-Mutagen provides a direct link to KINARI-Web[20],



(a) No excision. (b) Excision on Residue 3. (c) Excision on Residue 5.

Figure 3.8. Simulating mutations to glycine in KINARI-Mutagen. Side-chain hydrogen bonds and hydrophobic interactions are removed from the molecular model of a protein to simulate a mutation. In the wild-type of PDB file 2pm1 (a), two hydrogen bonds (light green bars) and two hydrophobic interactions (blue bars) exist among residues 3, 5, and 13. Excising residue 3 (b) removes the hydrophobic interactions that it forms with residue 5. Excising residue 5 (c) removes the hydrogen bonds between residue 5 and 13 and the hydrophobic interactions between residue 5 and 3.

for downloading a PDB file. Chains, ligands and water molecules in the protein can be retained or removed, and covalent and non-covalent interactions are identified.

In step 2, the Mutation Engine performs an *in silico* mutation (described in Section 3.3.3). In the third phase, the KINARI software is invoked to perform rigidity analysis on each mutant. Detailed descriptions of the rigidity calculation and modeling options are described in Section 2.3. When rigidity analysis is complete, an integrated Jmol-based visualizer is used to inspect the rigid regions of each mutant.

In the final stage of KINARI-Mutagen, the rigidity results for each of the mutants are aggregated. Information about critical residues can be inferred from several of the generated plots. Although this version of KINARI-Mutagen does not automatically predict which residues are critical, the *SASA and Size of Dominant Rigid Cluster vs Excised Residue* plot (see Figure 5.2 as an example) designates a critical residue threshold, which is the average size of the dominant rigid body for all of the analyzed mutants. Residues whose *in silico* mutation to a glycine causes the dominant rigid body to decrease in size to below this threshold, are easily identified.

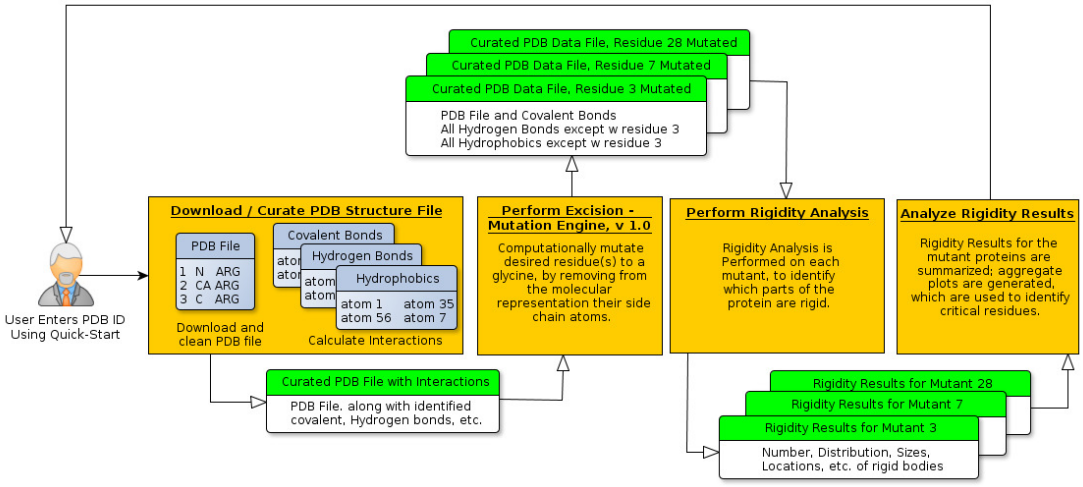


Figure 3.9. System description of KINARI-Mutagen. A PDB file is downloaded and curated, excision is performed to generate mutants, each mutant is analyzed, and rigidity results are aggregated. The generated plots and metrics provide information about which residues are critical in maintaining the protein's rigidity. Shown here is the procedure for generating and analyzing the mutation of residues 3, 7 and 28.

In addition, KINARI-Mutagen uses the *SurfRace* program[83] to calculate the Solvent Accessible Surface Area (SASA)[47] of each residue. A residue that is not exposed to the solvent has a low SASA value, measured in Å^2. Residues closer to the surface of a protein have higher SASA values, and completely buried residues have a SASA value of 0. Residues on the surface of a protein are not expected to help maintain a protein's stability[2]. Thus, we included the SASA calculation in the *SASA and Size of Dominant Rigid Cluster versus Excised Residue* plot to permit us to easily determine if KINARI-Mutagen can identify critical residues on the surface of a protein that may not be easily identified using other methods.

CHAPTER 4

ANALYZING PROTEIN BIOLOGICAL ASSEMBLIES AND CRYSTALS

The goal in developing KINARICrystal and BioAssembly was to determine if, and how, the interactions among neighboring crystal cells and protein subchains affect the flexibility of a biological assembly or of the molecule in its crystallized state. In this chapter, we demonstrate that new insights into protein flexibility are obtained by performing rigidity analysis on biological assemblies and protein crystal lattices. In the analysis of over 900 crystal lattices, we have identified two types of behaviors: some crystals aggregate into rigid bodies that span multiple unit cells/asymmetric units, while in other cases, the rigidity properties of the asymmetric units are retained, because the rigid bodies did not combine. We also identified two interesting cases where rigidity analysis correlated with the functional behavior of the protein.

4.1 Introduction and Motivation

Within the lattice of a crystal, the interactions among neighboring cells and subchains affect the flexibility of a biological assembly or of the molecule in its crystallized state. Zhang, *et al.* [90], in a comparison of 25 crystal forms of T4 lysozyme, revealed that crystal contacts perturb a protein's backbone structure by 0.2 to 0.5Å. Also, flexibility studies using rigidity analysis have been performed until now primarily on *individual* asymmetric units (the smallest part of a protein that is needed to re-create the protein's biological functional form) from the data available in the PDB. To determine if rigidity analysis could provide additional information about the effects of

Table 4.1. Experimental setup for generating crystal lattices

Step	Description
Selection and Curation	324 PDB files selected for analysis, having 50-150 residues in asymmetric unit)
Building of Crystal Lattice	*supercell.py* and custom scripts were used to apply symmetry and translation operations on each protein's asymmetric unit
Performing Rigidity Analysis	The KINARI software was used to calculate the rigid regions of the asymmetric unit and crystal lattices of each protein
Identifying Rigid Clusters	The rigidity results were analyzed to determine if interactions of unit cells led to larger rigid clusters.

crystal packing, and to determine if, and how, the interactions among neighboring cells and subchains affect the flexibility of a biological assembly or of the molecule in its crystallized state as measured using rigidity analysis, we analyzed a dataset of more than 900 crystal structures of more than 300 proteins.

4.2 Summary of Methods

Our computational setup involves the following: we parse the input PDB file, build the biological unit and desired crystal structures, then we use our KINARI-Web software to place hydrogen atoms, identify chemical interactions, and perform rigidity analysis, which outputs the rigidity clusters. To perform the experiments, we selected a dataset based primarily on protein size, for reasons having to do with the limitations of the current implementation. However, in our experiments we demonstrate that our software is able to handle relatively large protein structures.

Table 4.1 summarizes the steps of our experimental setup and methods for generating and analyzing crystals build from PDB structure files. The advanced search feature of the PDB was used to select proteins that had between 50 and 150 residues, whose structure was determined using X-Ray crystallography. Only proteins with no DNA nor RNA were selected.

We tested our method on 982 crystals lattices of various sizes, build from 324 protein structure files retrieved from the PDB. We found that the rigidity results vary among different proteins, an indication that there is additional information to be extracted from rigidity analysis of crystals and biological assemblies, as opposed to just a single asymmetric unit. In some cases, the biological unit, analyzed in isolation, exhibits significantly more flexibility than its crystal counterpart, while in others, the rigidity properties appear to be stable in the two forms.

For these 324 proteins, we built the unit cell, as well as $2\times1\times1$, and $2\times2\times1$ crystal lattices from the asymmetric unit data in each PDB file. The rigidity results were analyzed to reveal trends in the rigidity properties of the crystal lattices.

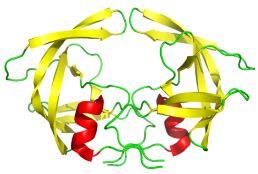
The rigidity analysis of a protein can find a *dominant* rigid cluster, whose size is substantially larger than any other rigid cluster, or several *significant clusters* of comparable sizes. Cluster sizes below a certain threshold (referred to as *insignificant*), are not taken into account in our analysis. They typically belong to flexible regions.

4.3 Results

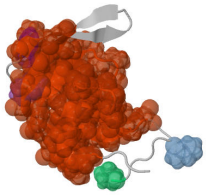
Here we present three detailed case studies of the rigidity analysis of biological assemblies, which highlight why it is important to analyze a protein in its functional form as opposed to just its asymmetric unit. Then, we include three case studies of proteins analyzed in crystal form, and identify a significant, small, or no effect in rigidity. Finally, we show a survey of 982 crystal lattice structures of various sizes, generated from 324 protein asymmetric units.

4.3.1 Merging of Rigid Clusters in a Biological Assembly

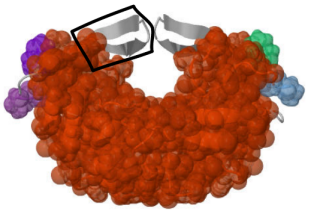
As a first proof-of-concept step to demonstrate the importance of analyzing the biological assembly versus just a protein's asymmetric unit, we analyzed PDB structure 1hhp. It is the monomeric unit (one-half) of the dimer aspartyl protease, which



(a) Functional form



(b) Monomer



(c) Dimer

Size	AU	BU
3	23	38
4	8	12
5	199	336
6	34	44
11	3	4
12	2	2
15	1	0
16	2	2
19	1	2
710	1	0
1802	0	1

(d) rigid clusters

Figure 4.1. Schematic and rigidity results of HIV-1 Protease. The protein is a dimeric aspartyl protease (PDB file 1hhp)(a). The asymmetric unit (b) in the PDB file has one significant rigid cluster. Two of the asymmetric units make up the biological form of the protein. When the rigidity of the biological form of the protein is analyzed (c), the rigid clusters of the two individual monomers combine into one dominant rigid cluster. The black outlined region designates one of the two beta hairpin loops often referred to as flaps, which function as chemical scissors and close in on the interior of the protein to facilitate an enzymatic reaction. In (d) the number of each type of rigid cluster is listed for the asymmetric and biological units (AU = Asymmetric Unit, BU = Biological Unit)

plays a crucial function in the maturation process of HIV-1. The functional form of the protease is made up of two identical chains, each composed of 99 residues. The PDB file 1hhp contains only the asymmetric unit. Using KINARI's BioAssembly, Curation, and Rigidity Analysis tools, we compared the rigidity of the asymmetric unit in 1hhp with its biological assembly (Table 4.1(d)).

From these results, we see that the asymmetric unit of 1hhp (Figure 4.1(b)) has a dominant rigid cluster of 710 atoms, while all other clusters have 19 or fewer atoms. In the biological assembly of 1hhp, however, the two monomeric chains have a rigid cluster of approximately 1,800 atoms, more than double the size in the asymmetric unit. Analyzing the monomeric form of HIV-1 Protease would not be expected to reveal any biologically relevant information about the protein, because it only exists in the dimeric form. However, we show this analysis here to demonstrate that chemical interactions between the two monomers affect the protein's rigidity, and that analyzing only the asymmetric unit of the protein would cause important structural information to be missed.

HIV-1 protease transitions between two predominant conformations, the open and closed forms. These two forms are distinguished by the locations of two extended beta hairpin regions often referred to as flaps (the left flap is shown as the outlined black region in Figure 4.1(c)). The two flaps close in onto the interior of the protease, to help facilitate an enzymatic reaction that is integral to the protein's function [62]. The two flap regions are identified in the biological assembly (Figure 4.1(c)) as being flexible, which is consistent with the studies and simulations on the protein's flaps, which have been shown to exhibit a wide range of motion [65].

4.3.2 The Biological Assembly Of A Nucleoprotein

The Rift Valley Fever Virus (RVFV) nucleoprotein [19] (PDB file 3ouo), was chosen to highlight how separate domains of a structure contribute differently to the protein's overall rigidity. The asymmetric unit in this PDB file contains a 2-chained dimer and a 1-chained monomer; each chain has 245 residues. The two biological units for this protein are the hexamer generated with three copies of the dimer and the hexamer generated with six copies of the monomer. Each monomeric chain has an extended, N-terminal arm.

We investigated the rigidity of the asymmetric unit of 3ouo, the monomeric unit of chain A, the monomeric unit of chain B, the dimer made of one copy each of chain A and chain B, the monomer made of chain C, the dimer made of two copies of chain C (Figure 4.2). Using the biological assembly information in the PDB file, we also generated the hexamer made of three copies of the A-B dimer, and the hexamer made of six copies of chain C. The tabulations of the rigid clusters for these components of the biological assembly and the two biological forms of the protein (Table A.3, Appendix A) show that as the structure becomes larger by a factor of n, the number of rigid clusters of a particular size increase by about the same factor. A closer look at Table A.3 further suggests that new rigid clusters are introduced when the hexamer

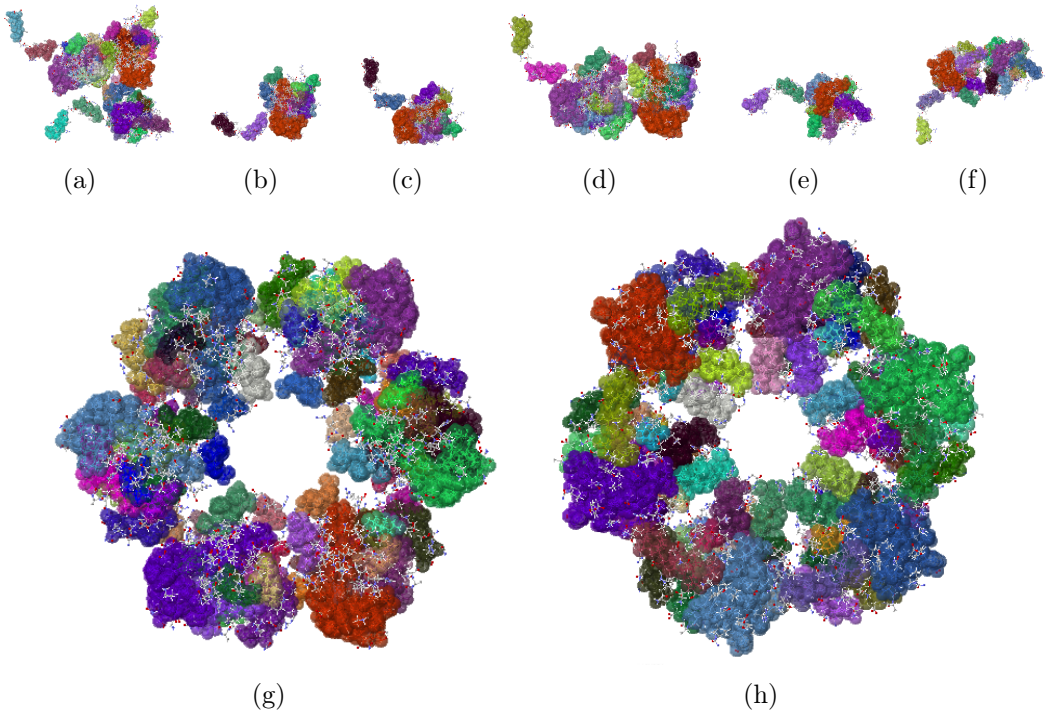


Figure 4.2. Rigidity Results of Rift Valley Virus. The asymmetric unit (PDB file 3ouo), is composed of three chains, A, B, and C. With the BioAssembly tool, we analyzed the rigidity of just chain A (b), chain B (c), the dimer made up of chains A and B (d), chain C (e), two copies of chain C (f), the hexamer made up of three copies of the dimer (g), and the hexamer made up of six copies of chain C (h).

is built from three copies of the dimer and when the hexamer is built from six copies of the monomer. In the first biological assembly, we found three new clusters with 237 atoms; in the second, we found six new rigid clusters with 118 atoms each. These rigidity results might be explained by the fact that the N-terminal arms bind to a hydrophobic pocket in the surface of the neighboring chain of the biological assembly, which is known to stabilize the hexamer structure [19].

4.3.3 Analyzing How Subunits of a Protein Affect Its Rigidity

The Vaccinia Virus D13 (PDB file 3saq) is a key structural component of the outer scaffold of viral crescents [5, 30]. The asymmetric unit contains two chains, A and B (Figure 4.3 b,d), with 576 residues each. The PDB file 3saq identifies two

biological assemblies that are generated from the two chains in the asymmetric unit. The first biological assembly (Figure 4.3c) is composed of three copies of chain A, and the second (Figure 4.3e) is composed of three copies of Chain B.

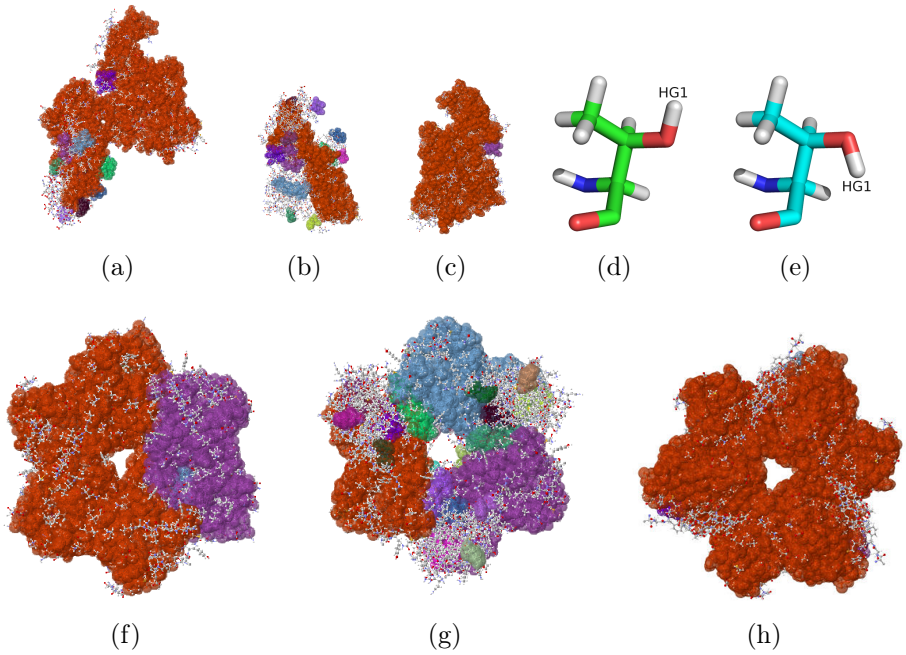


Figure 4.3. Rigidity analysis of Vaccinia Virus D13. The asymmetric unit (a) PDB file 3saq is composed of two chains, A and B. We analyzed the rigidity of just chain A (b), just chain B (c), the biological assembly made up of three copies of chain A (g),and the biological assembly made up of three copies of chain b (f). Due to the rotamer nature of certain amino acids, hydrogen atoms were placed by the *Reduce* software at different rotamer positions for certain residues, including residue 511, Threonine (residue 511 for subunits 1 and 2 are show in (d) and (e)). This caused subunit 1 to have a different number of hydrogen bonds than subunits 2 and 3, resulting in the non-symmetric rigidity of the second biological assembly. When KINARI's curation tools were used to adjust for this discrepancy of hydrogen bonds, the resulting biological assembly was symmetrically rigid (h), as expected.

The rigidity results for the two biological assemblies are surprisingly different, in that assembly 2, which has a dominant cluster composed of 9,475 atoms, is much more rigid than biological assembly 1, whose significant rigid cluster contains far fewer 2,277 atoms. To investigate this, we looked at the chemical interactions among the chains in both biological forms of the protein. Biological assembly 1 has 1,092

hydrogen bonds, and 700 hydrophobic interactions, while biological assembly 2 has 1,161 hydrogen bonds, and 805 hydrophobic interactions. This suggests that the stabilizing interactions among the three copies of chain B have a stronger effect on the rigidity of biological assembly 2 than do the chemical interactions among the three copies of chain A in biological assembly 1. The disparity in rigidity between the two biological assemblies might be explained by findings that multiple copies of both biological units form a honeycomb lattice, which is what provides structural stability for the immature virion membrane [30]. Thus both biological assemblies function cooperatively to perform their structural roles. Rigidity results for the structures generated from file 3saq are in Table A.2 in Appendix A.

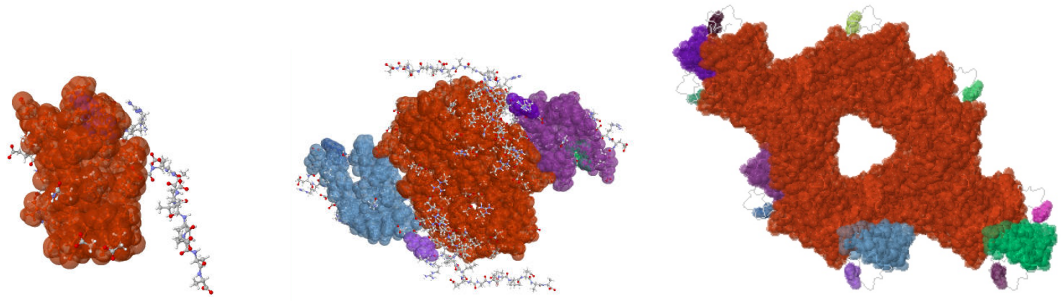
In addition, we investigated why the rigidity of the second biological assembly is non-symmetric, even though it is composed of three identical, symmetric, subunits, that are translated and rotated copies of chain B. We compared the hydrogen bonds in each of the three subunits, and found that they had 385, 386, and 384 hydrogen bonds respectively. A further inspection revealed that chain B has several amino acids, for example residue 511, Threonine, to which hydrogen atoms can be assigned in several ways. This is because threonine is one of two amino acids out of the naturally occurring twenty with two chiral centers [56], and it can exist as four possible stereoisomers (molecules that have the same molecular formula and sequence of bonded atoms, but that differ only in the three-dimensional orientations of their atoms in space). In addition, Threonine can assume one of several rotamer conformations even among a sample of the same protein [74, 50]. These help to explain why the adding of hydrogen atoms to such a residue can be done in one of several ways. The RMSD aligned, superimposed residues 511 for subunit 1 (Figure 4.3f) and subunit 2 (Figure 4.3g), have their HG1 hydrogen atoms (as were placed using the *Reduce* software) at different rotamer locations, which explains why one of these engages in a hydrogen bond, and the other does not.

To confirm that indeed the disparity of the number and placement of the hydrogen bonds among the three subunits of the second biological assembly is what causes the protein's non-symmetric rigidity, we performed a pairwise comparison of the stabilizing interactions among the three subunits. We identified where the three subunits differ in their hydrogen bonds. To investigate the effect of adding these hydrogen bonds involving the rotamer residues, we manually added 2, 1, and 3 hydrogen bonds to the first, second, and third subunits. We used the KINARI curation software (step 4) to insert the hydrogen bonds. In this case, the resulting rigidity of the biological assembly turned out to be symmetric (Figure 4.3g). Table A.2 in Appendix A shows the counts and sizes of the rigid clusters for the subunits of PDB structure 3saq.

In this case study, it is striking that the placement of such a small number of hydrogen bonds in a subunit of a biological assembly can vastly alter the rigidity of the trimeric protein. Adding 6 hydrogen bonds to an already existing 1,161 had a profound impact on the rigidity of the biological assembly. On the one hand, this suggests that the rigidity analysis of this protein is overly sensitive to these small differences in the counts and locations of hydrogen bonds. However, the fact that rigidity analysis is sensitive to the placement of these few hydrogen bonds might be taken advantage of, to help identify stabilizing interactions that are critical in stabilizing a macromolecular structure.

4.3.4 Crystal Lattice Dominant Cluster Aggregation

The putative protein from the gram-negative bacterium *Thermus thermophilus* [16] (PDB file 2yzt), crystallizes in a P 3_1 2 1 space group, which has 6 associated symmetry operations. Its small size (579 atoms) allows us to quickly analyze the asymmetric unit, unit cell (1×1×1), as well as 2×1×1 and 2×2×1 crystal lattices. The unit cell, the 2×1×1 crystal, and the 2×2×1 crystal have, respectively, 2, 4, and 8 asymmetric units.



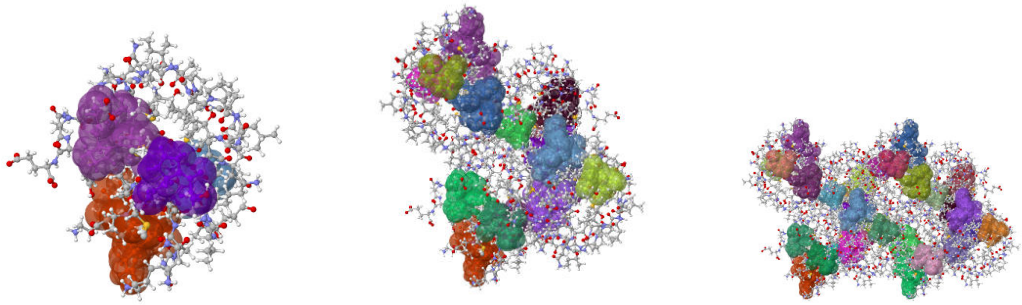
(a) Asymmetric unit (b) Unit cell made up of 2 asymmetric units (c) 2×2×1 Crystal Lattice

Figure 4.4. Aggregating of rigid clusters in unit cells of *Thermus thermophilus*. The significant rigid cluster in the asymmetric unit (a) of PDB file 2yzt has 463 atoms. In both the unit cell and the generated lattice, chemical interactions between the neighboring asymmetric units cause the rigid clusters of the individual unit cells to aggregate into a dominant one.

The asymmetric unit is a globular structure with a rigid region and a tail-like segment that remains flexible (Figure 4.4(a)). The dominant rigid cluster contains 463 atoms, and all other rigid clusters contain fewer than 26 atoms. The unit cell (the protein after the application of the 6 symmetry operations, i.e. the 1×1×1 crystal) maintains two significant rigid clusters of 463 atoms, but the other rigid clusters of the unit cell combine to form a rigid body containing 2,504 atoms, approximately 6 times the size of the significant rigid cluster in the unit cell (Figure 4.4(b)). In other words, the significant clusters being adjacent, combine to form the dominant rigid body in the crystal. For the 2×2×1 crystal, the largest body contains 14,328 atoms (Table A.1 and Figure 4.4(c)), which is significantly larger than four times the size of the significant rigid cluster in the unit cell. This indicates that the chemical interactions among the unit cells of the crystal have a significant impact on the rigidity of the entire crystal lattice. Table A.1 in Appendix A lists the rigidity results of the crystals generated from PDB file 2yzt.

4.3.5 A Significant Increase of Rigid Clusters in a Crystal Lattice

In the previous case study we found a very rigid crystal of PDB file 2yzt, with a very large dominant cluster. This is not always the case. Aggregating the asymmetric units of some proteins into a crystal lattice does not appear to greatly affect the rigidity of the resulting crystal structure. We illustrate this by analyzing one of the four types of antifreeze proteins found in marine fish living at sub-zero temperatures [40] (PDB file 1ucs). This protein crystallizes in a P 2_1 2_1 2_1 space group, which has 4 related symmetry operations. The unit cell of 1ucs is made up of 3 asymmetric units, the 2×1×1 crystal has 6 asymmetric units, and the 2×2×1 crystal has 12. In this case, the asymmetric unit does not have a dominant rigid cluster; it has four small significant clusters (Figure 4.5). Their number increases proportionally to the size of the crystal, and there is no aggregating of rigid clusters of unit cells (Table A.4 in Appendix A and Figure 4.5).



(a) Asymmetric unit (b) Unit cell of 1ucs (c) 2×1×1 Crystal Lattice

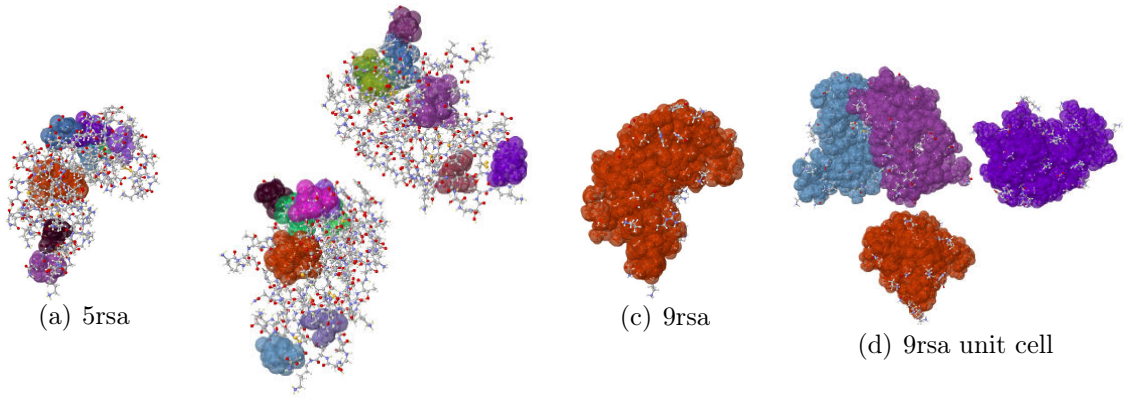
Figure 4.5. Rigid clusters of unit cells of Type III Antifreeze Protein RD1. In the crystal structure of Antarctic Eel Pout (PDB file 1ucs), the asymmetric unit (a) is composed of four significant rigid clusters (many more smaller ones are not displayed). Unlike PDB structure 2yzt (Figure 4.4), none of these is substantially larger than the others. In this case, the number of significant rigid clusters in the crystal form is more than the sum of the significant rigid clusters of the asymmetric units. This indicates interactions between the units that affect the rigidity of the crystal structure.

4.3.6 Rigidity Analysis Of Several Forms of Ribonuclease A

Our last cast study is on Ribonuclease A, which we analyze based on two PDB files (5rsa and 9rsa). In one case (5rsa), we see that aggregating the asymmetric units into a crystal has no bearing on the rigidity of the lattice. 5rsa crystallizes in a P 1 2_1 1 space group, with only 2 symmetry operations. The asymmetric unit, which contains a single instance of the protein, is made up of 2,250 atoms. It has no dominant cluster, and the significant ones are made of approximately 65 atoms. The unit cell and the two crystals we analyzed (2×1×1 and 2×2×1) all have significant clusters of about the same size (65) (Figure 4.6(a) and 4.6(b)). Unlike the previous two case studies (PDB files 1ucs and 2yzt), no clusters (significant or insignificant) are merged at the interface of the units when forming the crystals (Table A.5 in Appendix A). This may be because no hydrogen bonds form between the two asymmetric units in the unit cell (data not shown). Also, we notice that only 4 new bonds appear between the cells that make up the 2×1×1 lattice, and only 8 in the 2×2×1 lattice. However, this small number of bonds does not preclude the formation of larger clusters (see case study for 3saq). These rigidity results of the crystallized form of the protein, therefore, are in agreement with the protein's known properties, in that Bello [7], *et al.* have shown that Ribonuclease A (PDB code 5rsa) retains its function in the crystallized form.

RNase A is a widely studied protein, for which there are many structure files in the PDB. One such entry, file 9rsa, is that of a derivative, which is know to retain only 1% of its enzymatic activity [60]. We compared the rigidity results of the two forms. PDB structure 9rsa crystallizes in a P $2_1 2_1 2_1$ configuration, and the asymmetric unit contains two instances of the protein. The PDB file 9rsa contains two copies of RNase A. To make the comparison meaningful, we retained only a single instance of RNase A from file 9rsa. Interestingly, the rigidity results of the asymmetric unit (Figure 4.6(c))

of 9rsa has a dominant rigid cluster of 1,339 atoms, in stark contrast to the significant rigid clusters in 5rsa.



(a) 5rsa

(b) 5rsa unit cell

(c) 9rsa

(d) 9rsa unit cell

Figure 4.6. Comparing rigidity of two crystal forms of Ribonuclease A. For Ribonuclease A (PDB file 5rsa), which performs its function in its crystalline form [7], the asymmetric unit (a) and its unit cell (b) are largely flexible. In contrast, a derivative of the protein (PDB file 9rsa), which is known to lose virtually all activity [60], the asymmetric unit (c) and unit cell (d) are both very rigid, and contain far more atoms in the dominant rigid cluster than the structure in file 5rsa.

4.3.7 Survey of 982 Crystal Structures

In addition to the previous case studies, we surveyed a dataset of 324 proteins, in 982 biological assembly and crystal forms. The dataset contains a diverse set of proteins. As illustrated in Table 4.2, it is not biased towards proteins with only large dominant clusters. For each crystal, we tallied what percent of the structure's atoms were in the dominant cluster. We summarized the rigidity of the crystals and information concerning the dominant cluster in three groups, with dominant cluster size larger than 75%, between 50 and 75%, and between 25 and 50%. We also tabulated each crystal's number of hydrogen bonds and hydrophobic interactions. The generated crystals varied in size, the largest having 54,107 atoms (PDB file 3hon, 2×2×1 crystal). Some crystals were made up of as few as a single asymmetric unit,

Table 4.2. Summary of dataset for survey of crystal structures

# Protein structure files retrieved from PDB	324
# Crystals generated using KINARI	982
Maximum number of asymmetric units in a crystallographic unit cell	48
Largest crystal (atoms)	54,107
# Crystals with hydrogen bonds between unit cells	776
# Structures with 25% or more but fewer than 50% of atoms in dominant rigid cluster	228 (18%)
# Structures with 50% or more but fewer than 75% of atoms in dominant rigid cluster	364 (29%)
# Structures with 75% or more of atoms in dominant rigid cluster	334 (27%)

or as many as 48. The asymmetric units and generated crystals varied surprisingly in terms of how many of the structure's atoms were part of the dominant rigid body.

We performed a preliminary classification of the proteins, based on the rigidity of their crystal lattices. We summarize it in Table 4.3. For the majority of proteins, we observed a behavior which we call *dominant cluster aggregation at all levels* (**case 1**). This means that the crystal contains a dominant cluster of size more than the sum of the dominant clusters in its unit cells or asymmetric units. For 27 of the proteins (**case 2**), we observed a *dominant cluster aggregation at the unit cell level*, and no aggregation at the crystal level. **Case 3** (33 proteins), shows no change in rigidity among the asymmetric unit and any of the generated crystals. For twenty of the proteins (**case 4**) the unit cell and 2×2×1 crystal had rigid bodies that spanned several asymmetric units or unit cells, respectively, but the 2×1×1 crystal had rigid bodies that were no bigger than the rigid bodies in the unit cell. This may be because the interactions between the unit cells along one crystal lattice axis may be different. The 10 proteins in **case 5** had asymmetric units and unit cells that had the same sized dominant cluster, but for the 2×1×1 and 2×2×1 crystals there was a collapse of the dominant cluster. **Case 6** contains those proteins with a dominant cluster at the unit cell level that spans multiple asymmetric units but does not aggregate in the larger crystals.

There were some proteins which did not quite fit into any of the six cases. We observed that the asymmetric unit and unit cell of PDB file 2bf9 had no change in rigidity, but when building the 2×1×1 crystal there was a collapse of rigidity, and the 2×1×1 and 2×2×1 crystals had the same rigidity. For the analyses of PDB file 6rxn, the asymmetric unit, unit cell, and 2×1×1 crystal all had no additional collapse, but when the 2×2×1 crystal was built the size of the largest rigid cluster increased.

Although this survey is by no means comprehensive, it already displays patterns of rigidity properties for protein crystals that motivate future extensions of our computational experiments for fully understanding protein crystal lattices. Such future extensions to the software will need to take into account several structural features of the crystals lattices. For example, the functional form of a protein is stabilized in the crystal lattice not only via interactions among the biological assemblies in the different unit cells, but also through water molecules at the crystal contact locations. The molecules may play an important role in stabilizing the crystallized macromolecules. Currently, although the KINARI software does permit the inclusion of water molecules when analyzing the rigidity of a single biological assembly, the effect of water molecules at the interface of crystal units is not included in the rigidity analysis.

4.4 Conclusions

The results of a rigidity analysis of a single asymmetric unit may not always provide structural information that is relevant to the biological form of a protein. Using our KINARI software, we demonstrated that additional functional and rigidity information is gained by analyzing a protein's biological assembly and/or crystal structure.

We analyzed 982 crystal lattices, made up of unit cells composed of relatively small asymmetric units. Performing a large-scale study of protein structures with

Table 4.3. Classification of protein crystals according to their rigidity

	Case	# Proteins	% of Dataset
1.	Dominant cluster aggregation at all levels	192	59
2.	Dominant cluster aggregation at the unit cell level	27	8
3.	No combining of rigid bodies in unit cell nor in larger crystals	33	10
4.	Rigid bodies of asymmetric units combined in unit cell and 2×2×1 crystal, but not for 2×1×1 crystal	20	6
5.	Size of dominant cluster in asymmetric unit and unit cell was the same, but there was aggregation of dominant body in 2×1×1 and 2×2×1 crystals	10	4
6.	Dominant cluster at unit cell that spans multiple asymmetric units but does not aggregate in crystals	24	8
7.	Other; unclassified.	18	5

larger asymmetric units would be computationally expensive (due to the size of the molecules involved). Overcoming this limitation will require novel mathematical and computational extensions to our software.

For the analysis of larger assemblies of asymmetric units, we found that relying on "black box" software has to be taken with a grain of salt. In the case of X-ray resolved structures, the PDB files do not contain hydrogen atoms, and these have to be placed with software (such as *Reduce*). Conversely, using the KINARI curation feature allows one to formulate and verify hypotheses concerning the molecular model, when the placement of atoms or stabilizing interactions needs to be disambiguated.

In summary, this work shows that rigidity analysis of protein crystals and biological assemblies is feasible, and can now be easily performed using the KINARI software. Moreover, this permits fast evaluation of the rigidity properties of the biological form of a protein, and can be used to test hypotheses regarding the role that different subunits play in contributing to the rigidity of a biomolecule.

CHAPTER 5

PREDICTING THE EFFECT OF MUTATIONS ON PROTEIN STABILITY

Predicting the effect of a single amino acid substitution on the stability of a protein structure is a fundamental task in macromolecular modeling. It has relevance to drug design and understanding of disease-causing protein variants. We use KINARI-Mutagen to identify critical residues, and we show that our predictions correlate with destabilizing mutations to glycine.

5.1 Motivation and Introduction

A mutation in a protein's amino acid sequence can have deleterious effects on its stability and function. A number of diseases result from single point mutations. Hence knowing a mutation's effect can guide the design of drugs aimed at combating those disorders. To predict and better understand the roles of mutations, the genetic information that codes for the amino acid sequence of a protein can be altered, and the expressed mutant proteins analyzed to infer the impact of the specific mutation. Such studies are aided by several widely-used molecular biology techniques, such as site-directed mutagenesis[29]. Unfortunately, such experiments are often labor and time intensive. The possible number of mutants that can be made from even the smallest proteins makes exhaustive mutagenesis studies impractical. For example, 20^{100} mutants could in principle be engineered for a 100-residue protein using the 20 naturally occurring amino acids.

Here we focus on rigidity analysis as implemented in our software KINARI[20]. The premise is that the protein's function is directly correlated with its distribution and sizes of rigid clusters, and destabilizing any of them will have an observable effect.

We used KINARI-Mutagen to answer two types of questions: 1) will mutating a residue in a protein destabilize it, or 2) given a protein, which residues could destabilize a protein if mutated? We demonstrate our software's usefulness in two case studies, which show that the mutated residues identified by KINARI-Mutagen as critical correlate with experimental data, and would not have been identified by other methods such as Solvent Accessible Surface Area measurements or residue ranking by contributions to stabilizing interactions.

5.2 Background and Related Work

Here we review previous work that addressed the effect of mutations on the structure of a protein. We summarize other studies whose aim was to determine the effects of mutations on a protein's stability.

5.2.1 Mutations Affect Protein Structure and Function

Deoxyribonucleic acid, DNA, contains the instructions on how amino acids should be joined during protein synthesis to make a protein. If there is an error in the process, the resulting amino acid sequence may differ from the most common sequence of amino acids, which is designated the **wild-type** version of that protein. A protein with mutations is called a **mutant**. Mutant proteins contribute to many genetic diseases. For example, single point mutations in the cystic fibrosis transmembrane conductance regulator protein lead to development of cystic fibrosis. In the protein α-galactosidase there are over 190 single point mutations that lead to development of Fabry Disease[23]. Thus understanding the effect of point mutations is of biomedical importance.

A mutation in the amino acid sequence can inhibit the protein's function. Alber *et al.*[2] have found that temperature-sensitive mutations often occur at residues which are structurally important. Similarly, a mutation at a residue location that plays a crucial role may render a protein inoperative[26]. However, because not all mutations are equally disruptive, it is important to know how a mutation will affect the protein.

5.2.2 Related Work

One way to study how mutations affect a protein's function is to locate or synthesize pieces of DNA, called templates, which contain a mutation. Many copies of the DNA template can be generated, through a process called polymerase chain reaction (PCR). The multiple copies of the mutated DNA can then be introduced into a cell's nucleus, where the DNA is transcribed, and translation at the ribosomes results in the synthesis of mutated proteins.

One way in which the role of a residue substitution can be directly studied is by mutation experiments in the physical protein. Matthews *et al.* have designed and analyzed many mutants of lysozyme from the bacteriophage T4. When core residues in lysozyme were substituted by alanine, an analysis of the crystal structures revealed that the unoccupied volume in some of the mutants underwent a collapse, while other mutants formed an empty cavity[89]. Residues of T4 lysozyme with high mobility or high solvent accessibility were shown to be much less susceptible to destabilizing substitutions. The authors concluded that residues that are held relatively rigidly within the core of the protein make the largest contribution to the protein's stability[2]. Also, studies have been performed to determine the role that disulfide bonds play in stabilizing lysozyme [52], and a host of residues have been mutated, to infer how they affect the protein's stability [51, 3, 2].

Although the studies by Matthews and others provide precise, experimentally verified insight into the role of a residue based on its mutation, such studies are

time consuming and often cost prohibitive. Moreover some mutant proteins cannot be expressed, due to dramatic destabilization caused by the mutation, and so only a small subset of all possible mutations can be studied explicitly. To address this, computational and analysis techniques have been proposed.

In computational experiments by Lee, *et al.*[48], the side-chains in each of 78 structures of mutant proteins were perturbed. A heuristic energy measure, E_{calc}, was used to predict the stability of each protein, and compared to known activity data. Gilis, *et al.*[24], estimated the folding free energy changes upon mutations using database-derived potentials, and concluded that hydrophobic interactions contribute most to the stabilizing of the protein core. Similarly, Prevost, *et al.*[66], have used molecular dynamics simulations to study the effect of mutating Barnase residue Isoleucine 96 to alanine, and predicted that the major contributions to the free energy difference arose from non-bonded interactions.

Machine learning and statistical methods have also been developed to help predict the effects of mutations. Cheng,*et al.*[12] used Support Vector Machines to predict with 84% accuracy the sign of the stability change for a protein induced by a single-site mutation. However, their online tool MUpro only outputs whether a mutation is expected to stabilize or destabilize a protein, and does not provide data that can be used to rank residue mutations based on their impact on the protein's stability. Also, data of amino acid replacements that are tolerated within families of homologous proteins has been used to devise stability scores for predicting the effect of residue substitutions[81], which has been extended and implemented into an online web server[88]. It is not clear, however, how the use of environmental substitution data to devise a score for the effect of a mutation is appropriate if no such data exists, or if a newly discovered protein has few homologues.

The hydrophobics effect is the tendency of water molecules to exclude non-polar molecules, which leads to segregation of water and non-polar substances. The hy-

drophobic effect has been shown to play an important role in the stability of proteins, in which hydrophobic atoms are buried within the interior of the protein, away from its surface. In the case of mutants of lysozyme from bacteriophage T4, amino acid substitutions decreased stability of the protein by different amounts, depending on the exact pH of the solvent. In chemistry, pH is a measure of the acidity of a solution. ΔG values for different lysozyme proteins in different pH solutions ranged from 2.7 kcal/mol for the L46A mutation, to 5.0 kcal/mol for the L99A mutation. Kilo-calorie per mole, or kcal/mol, is a derived unit of energy. From such experiments, it has been shown that changes in thermal stability associated with each of the mutations vary substantially from case to case, but can be correlated with the size of the cavity that is created by the mutation [18]. According to these studies, the larger the cavity that is created by the mutation, the more destabilizing is the replacement.

It has also been found that the energy difference between two conformations of a protein arise from bonding terms involving degrees of freedom of the mutated side chain and from non-bonded interactions of that side chain with its environment in the folded protein [66]. That set of experiments concludes that the essential effect of mutations is the difference in the stability of the folded state rather than the differential salvation of isoleucine and alanine in the unfolded state.

Scientists have also used neural networks in an attempt to predict the effects of mutations. A neural network is a framework in which different nodes - atoms, in this case - interact with each other, and a numerical analysis of the neural network is used to infer the relationships and structural dependencies between the nodes. Emidio Capriotti, *et al.* have used neural networks to predict whether a given mutation increases or decreases a protein's stability, without predicting the exact $\Delta\Delta$G value, but merely the sign of that measurement [10]. Capriotti claims that the major drawback of methods based on physically effective energy functions is that they are

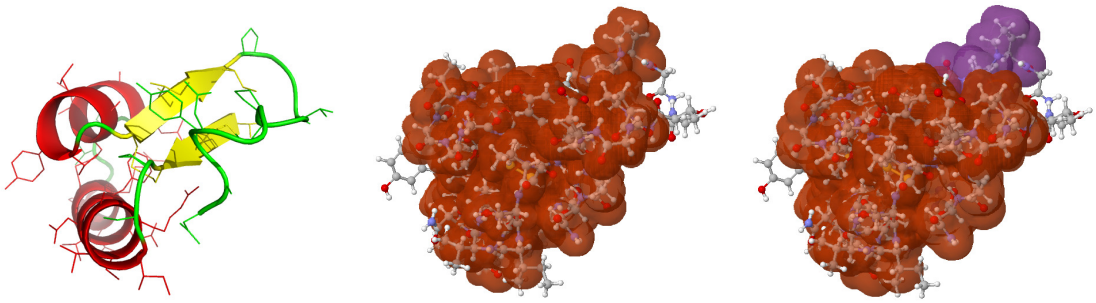
computationally intensive, and so their usage is nearly prohibitive for applications that involve an analysis of a large number of proteins.

In still other work, scientists have developed force fields to help predict protein stability. A force field refers to the parameter sets used to describe the potential energy of a system of particles, and they are derived from experimental work and quantum mechanical calculations. Such force fields allow us to reason about and quantify the interactions among atoms. Guerois *et al.* have developed a computer algorithm, *FOLDEF*, that aims to provide a fast and quantitative estimation of the importance of the interactions contributing to the stability of proteins and protein complexes [27]. Guerois concludes that packing density around each atom is a suitable parameter that can be used to predict the flexibility of proteins; a simple method of counting the contacts around hydrophobic residues can be quite successful at predicting $\Delta\Delta$G.

Thus, progress has been made in predicting the effects of mutations on protein stability. However, many such methods rely on computationally intensive energy calculations, or are not able to infer the role of a single amino acid in stabilizing a protein's structure. To complement these already existing methods, we seek to apply rigidity concepts to the computational prediction and analysis of the stability of mutant protein structures.

5.3 Methods and Results

In two in-depth case studies we show that the mutated residues identified by KINARI-Mutagen as critical correlate with experimental data, and would not have been identified by other methods such as Solvent Accessible Surface Area measurements or residue ranking by contributions to stabilizing interactions. We also generated 48 mutants for 14 proteins, and compare our rigidity results with experimental data.



(a) (b) (c)

Figure 5.1. Rigidity results of two *in silico* mutants of Crambin. KINARI-Mutagen was used to analyze PDB file 1crn (a). The dominant rigid cluster of the wild-type protein (b) is compared to the rigidity results of a mutant (c) generated by KINARI, to determine the effect of the mutation.

5.3.1 Case Study - Crambin

To demonstrate KINARI-Mutagen, for the first case study we generated and analyzed mutants of Crambin (PDB file 1crn, Figure 5.1(a)), a 46 amino acid plant seed protein, whose crystals diffract to ultra-high resolution[78, 77].

The cartoon representation and rigidity results for two generated mutants of Crambin are shown in Figure 5.1. The wild-type protein has a dominant rigid cluster (brown, Figure 5.1(b)). Viewing the rigidity results of a mutant can be used to infer the impact of the mutation on the protein's rigidity. When excision was performed on residue 10, an arginine, (Figure 5.1(c)), the size of the dominant rigid cluster decreased, and the number of clusters increased, when compared to the wild-type.

Several residues in the core of Crambin had a pronounced effect on the protein's predicted rigidity when they were mutated (residue 3 for example). Similarly, many residues (7, 15, and 28) that are solvent accessible, when mutated, had little effect on the dominant rigid cluster. These findings were not surprising, because residues on the surface of a protein are not expected to help maintain a protein's stability[2]. However, the software was able to identify critical residues on the surface of the protein that affected the protein's rigidity when mutated to a glycine.

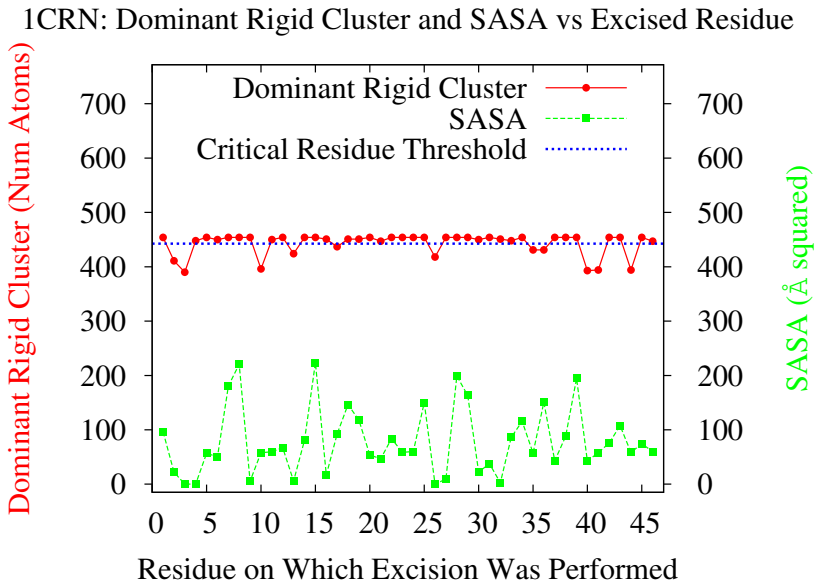


Figure 5.2. SASA and Size of Dominant Rigid Cluster plot for Crambin. Solvent exposed amino acids that play a crucial role in stabilizing the protein are identified. The blue dotted line, the critical residue threshold, designates the size of the average dominant rigid body among all of the generated mutants, and can be used to select residues whose mutation affects the protein's rigidity.

We inspected the *Dominant Rigid Cluster and SASA vs. Excised Residue* plot (Figure 5.2), to identify critical residues that could not be located by using the SASA calculations alone. Of the 11 mutants that had dominant rigid clusters below the critical residue threshold, eight of them (residue 2, 10, 17, 35, 36, 40, 41, and 44) had SASA values in the wild-type protein that were well above zero. Of these eight, 4 are known to be identical among viscotoxin A3 and α_1-purothionin[78], while another 3 of them were conserved among two of these three homologous proteins. Only residue 44, with a SASA value of 70, was not conserved among the three homologues, indicating that KINARI-Mutagen identified incorrectly residue 44 as critical.

One may hypothesize that KINARI's results may have a simpler explanation. A residue engaged in many stabilizing interactions (hydrogen bonds and hydrophobics) is likely to have an effect on the protein's stability and rigidity. To investigate this, we inspected the strengths of the hydrogen bonds of Crambin, which are calculated

by KINARI-Mutagen using an energy function[54]. Residues 46, 21, and 30, have side chains that engage in strong hydrogen bonds with energies of -5.2, -5.9, and -5.07 kcal/mol. KINARI-Mutagen did not identify them as critical, and they are not conserved among homologues of Crambin. We similarly confirmed that critical residues could not have been found by merely identifying amino acids that engage in many hydrophobic interactions. Residue 19, a proline, engages in 5 hydrophobic interactions. It is neither conserved among Crambin homologues, nor did KINARI-Mutagen identify it as critical.

KINARI-Mutagen is thus a method that supplements other approaches that study protein stability due to mutations and residue conservation. The set of critical residues identified by our method is different that the set of amino acids that are ranked by just the strength of hydrogen bonds or number of stabilizing hydrophobic interactions. Moreover, KINARI-Mutagen can identify conserved surface exposed residues that would not be detected using SASA measurements alone.

5.3.2 Case Study - Lysozyme from Bacteriophage T4

In the second case study, we evaluate whether rigidity analysis can identify destabilizing mutations. From the literature[2, 6, 59, 61] we retrieved stability data for 158 different point mutations in lysozyme from bacteriophage T4 (PDB file 2lzm for wild-type). The rigidity of the wild-type of lysozyme from bacteriophage T4 is shown in Figure 5.3. The experimentally derived value $\Delta\Delta$G, the free energy of unfolding, measures the stability of a variant against a reference protein (nearly always the wild-type protein). The lower the $\Delta\Delta$G value, the more unstable is the variant. From the available $\Delta\Delta$G dataset, we selected the 8 mutations that involved a substitution to a glycine. We compared $\Delta\Delta$G values of these mutations that had been performed in the physical protein to the rigidity calculation predictions of KINARI-Mutagen.

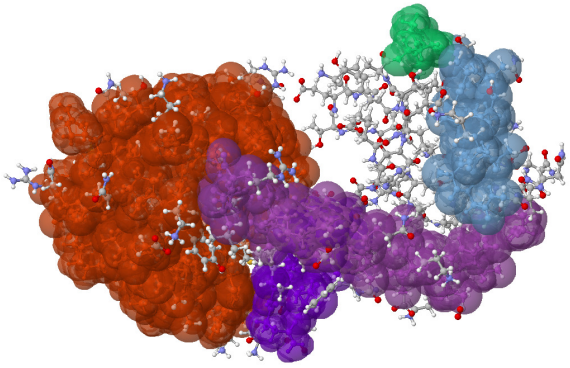


Figure 5.3. Rigidity results of wild-type Lysozyme from bacteriophage T4. The color bodies of PDB file 2lzm indicate clusters of atoms that are rigid, as identified by KINARI-Web.

Table 5.1 lists for each lysozyme mutant several rigidity measures, that we evaluated as predictors of protein stability. For the amino acid which was mutated at a particular sequence location, we list its Solvent Accessible Surface Area, the volume of the wild-type amino acid, the change of the volume of the residue when mutated to glycine, as well as the stability data from the literature ($\Delta\Delta$G), which we consider the "ground truth" stability measurement. The loss in number of hydrogen bonds and hydrophobic interactions that were caused by the mutation to glycine are listed, as well as the change of the dominant rigid cluster relative to the wild-type.

When KINARI-Mutagen was used to mutate residues 96, 105, 157, and 124, the size of the dominant rigid cluster decreased in size in parallel to a decrease in the $\Delta\Delta$G value. For example, residue 96, an arginine, when mutated to a glycine, caused the dominant rigid cluster of the mutant to decrease by 130 atoms relative to the dominant rigid cluster in the wild-type protein. When residue 96 was mutated to a glycine in the physical lysozyme, the stability of the protein decreased significantly, as indicated by the low $\Delta\Delta$G value. Similarly, for residues 105, 157, and 124, the size of the dominant rigid cluster decreased in size relative to the wild-type protein. The $\Delta\Delta$G values for mutations at residues 105, 157, and 124, indicate that their

Table 5.1. Rigidity data of 8 lysozyme mutants. For each mutant, the experimental ΔΔG value, the change in volume of the amino acid when mutated to glycine, the loss of hydrogen bonds and hydrophobic interactions, the change of the dominant rigid cluster, and the Cluster Configuration Entropy (Section 2.4) values are listed. For the wild-type of the protein, the CCE value is 0.43, and the dominant rigid cluster contains 830 atoms. Mutants are ordered by ΔΔG values, and rows shaded gray indicate amino acid mutations that KINARI-Mutagen correctly identified as destabilizing.

Sequence Number	WT Amino Acid, Volume ($Å^3$)	SASA ($Å^2$) of WT Amino Acid	AA Volume Change for mutation to GLY	ΔΔ G from literature	Hydrogen Bonds Lost in mutant	Hydrophobics Lost in mutant	Change to Dominant Rigid Cluster in mutant	CCE[25]
99	LEU, 166	0	-106	-6.3	0	0	0	0.43
96	ARG, 173	73.04	-113	-2.6	2	0	130	0.61
3	ILE, 166	43.72	-106	-2.1	0	1	0	0.43
59	THR, 116	128.75	-56	-1.6	1	0	0	0.39
105	GLN, 143	64.8	-83	-1.5	2	0	11	0.44
157	THR, 116	100.97	-56	-1.1	2	1	63	0.49
55	ASN, 114	113.45	-54	-0.6	0	0	0	0.43
124	LYS, 168	103.19	-108	-0.1	1	2	15	0.44

destabilizing effect is not as great as when a mutation is performed on residue 96. In these cases KINARI-Mutagen was able to predict a change in the protein's stability.

KINARI-Mutagen was not able in all instances to predict the effect of a mutation on the protein's stability. For residues 99, 3, 59, and 55, the lack of loss of hydrogen bonds and/or hydrophobic interactions when these residues were mutated to a glycine explains why KINARI-Mutagen's could not be used as a discerning measure of protein stability. For these mutations, the loss of hydrogen bonds and hydrophobic interactions were not as great as when mutations were performed on residues 96, 105, 157, and 124. The predictive ability of KINARI-Mutagen relies on the change in the molecular model due to a loss of these interactions. For these mutation instances, the change in the protein's stability is caused by phenomena that KINARI-Mutagen does not currently capture. We suspect that the change in volume of the wild-type amino

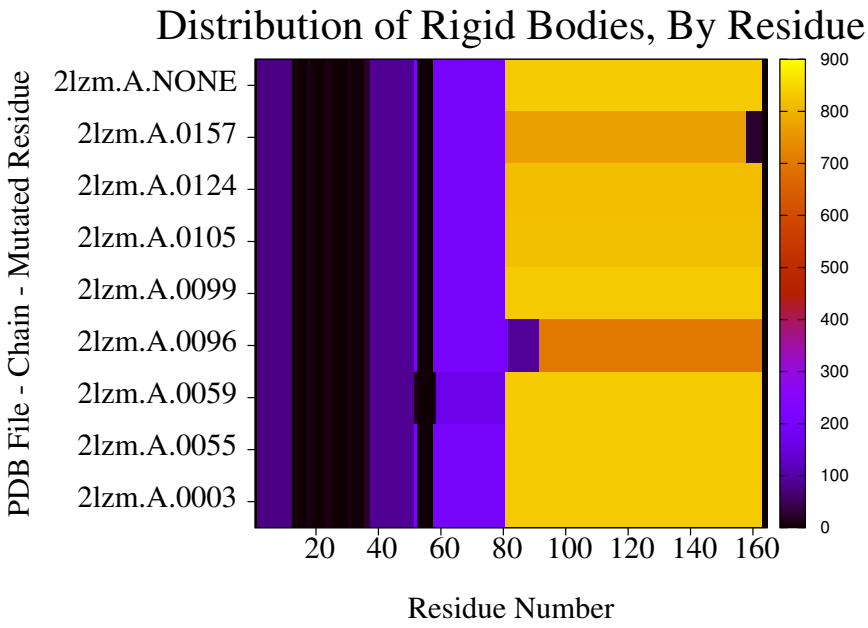


Figure 5.4. Distribution of Rigid Bodies, By Residue, Plot for Lysozyme. The left axis lists mutants that were analyzed. The vertical color legend on the right-hand side assigns colors to the rigid body sizes found among the mutants. The color at each x-y position in the plot indicates the size of the largest cluster that residue x belongs to for the mutant in row y.

acid to a glycine causes a large-enough collapse or reorientation of the protein's structure in the vicinity of the substitution, which affects the protein's stability.

Lastly, we compared KINARI-Mutagen's predictions to the Cluster Configuration Entropy (CCE) measurement (Section 2.4). The CCE values for several variants correlated well with the dominant rigid cluster metrics for those mutants. For mutants that had residue substitutions at amino acids 105, 124, 157, and 96, the CCE values were 0.43, 0.44, 0.49, and 0.61, respectively, while the change in the size of the dominant rigid cluster for those variants were 11, 15, 63, and 130.

To investigate why we did not predict the mutation of residue 59 Tyrosine being mutated to Glycine (designated as T59G) to be destabilizing, we referred to the *Distribution of Rigid Bodies, By Residue* (DRBR) plot (Figure 5.4). It was used to distinguish between mutations that have only a local effect on the rigidity of a protein and mutations that drastically affect a protein's stability. The row **2lzm.A.0059**

indicates that a mutant was generated by excising residue **59** of chain **A** of protein **2lzm**. For the residue 59 mutation, the change of the protein's rigidity is not localized to the largest rigid cluster, which explains why using the size of the dominant rigid cluster was not a good predictor of protein stability.

5.3.3 Validation - 48 Mutants

To further determine if KINARI-Mutagen could correctly identify destabilizing mutations in a wider range of proteins, we searched the ProTherm Database[42] for $\Delta\Delta$G measurements for substitutions that have been performed in the physical protein. A total of 167 entries had mutations to glycine. Of those, 48 mutants among 14 proteins had single-point substitutions. We also chose PDB files that had all core residues resolved.

We used KINARI-Mutagen to generate the 48 *in silico* mutants and analyze their rigidity. Along with the SASA value for each wild-type residue at the location where the mutation was performed, we tallied the change to the dominant rigid cluster of the protein caused by the point mutation, and the degree of hydrophobicity of each wild-type residue, using the Kyte and Doolite hydrophobicity scale[43]. The output of KINARI was also used to tally how many hydrogen bonds and hydrophobic interactions were lost due to the mutation. To facilitate analysis, the 48 mutants were grouped according to whether the substituted residue engaged in hydrogen bonds and hydrophobic interactions (Table 5.2). Detailed rigidity results for the 48 mutants are shown in Tables B.1 - B.4 in Appendix B.

KINARI-Mutagen relies on the loss of hydrogen bonds and hydrophobic interactions upon a residue's change to glycine, to predict the effects of a mutation. Thus we did not expect to accurately predict a substitution as destabilizing, if KINARI found that in the wild-type protein the amino acid engaged in neither hydrogen bonds nor hydrophobic interactions (Group 1). Group 2 has entries for which the residue of the

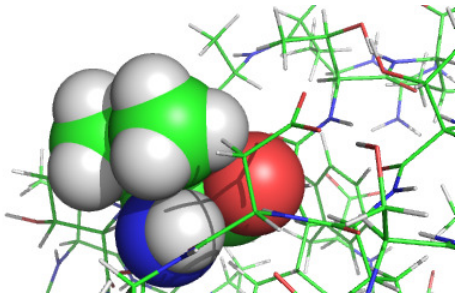
Table 5.2. Rigidity results of 48 mutants analyzed by KINARI-Mutagen. The results for the 48 mutants were grouped based on whether the mutated residue engaged in stabilizing interactions.

Group	**Description of Wild Type AA at Mutation Point**	# **Mutants**	**Identified As Destabilizing**
1	No hydrogen bonds or hydrophobics detected	13	0
2	Solvent exposed (>50%)	8	0
3	Too few hydrophobic	4	0
4	Stabilizing interactions found	23	22

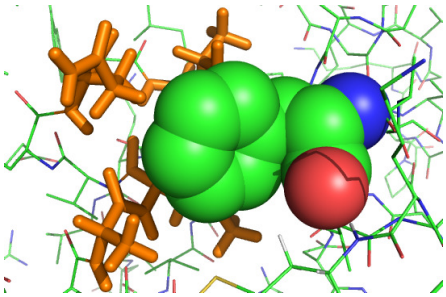
wild-type protein was solvent exposed (more than 50% of the residue was exposed). Because these residues are on the periphery of the protein, their being mutated to glycine would not be expected to have a large effect on the size of the dominant rigid cluster, especially if the side chain of the residue was protruding fully into the solvent (Figure 5.5(a)).

In Group 3, four mutants had wild-type amino acids (Valine, Leucine, Methionine, Phenylalanine) that do form hydrophobic interactions that can be observed by visual inspection. However, because of the packing of these core residues in this structure which were slightly less tight than in many protein cores, the algorithm in KINARI to detect hydrophobic interactions detected far too few of them. Figure 5.5(b) shows a phenylalanine that upon visual inspection should have been stabilized via several hydrophobic interactions, but no atoms in the residue were within 3.5Å of a heavy neighbor atom, so no hydrophobic interactions were detected. Had the atoms of that part of the structure been oriented slightly differently to allow closer packing, KINARI's hydrophobic detection algorithm would have placed more hydrophobic interactions there, which could have caused that residue to be labeled as critical when it was mutated.

Group 4 contains 23 mutants that had more reasonable numbers of hydrophobic interactions which were identified by KINARI, and many of them have hydrogen



(a) In the Streptomyces Subtilisin Protease inhibitor, (PDB file 3ssi), Valine 13 (spheres) is 56% exposed, so only 1 hydrophobic was detected, precluding KINARI analysis.



(b) In Staphylococcal Nuclease, (PDB 1stn) Penylalalanine 61 (green spheres) lies in a hydrophobic pocket (orange), but no hydrophobics were detected.

Figure 5.5. Solvent exposed amino acids not identified as critical. Some amino acids that are highly solvent exposed were not identified as destabilizing, because they did not engage in stabilizing interactions (Figure 5.5(a)). Some residues like those that are completely or largely solvent inaccessible lie more than 3.5Å from the nearest heavy atom, so hydrophobic interactions (orange sticks, Figure 5.5(b)) in the range of 3.6Å to 4.5Å are not found by the hydrophobic detection algorithm in KINARI, preventing quantitative analysis of the impact of the mutation to glycine.

bonds. Of these, 22 were identified as critical, based on the fact that these mutants had dominant rigid clusters that were smaller than the dominant rigid cluster of the wild-type protein.

From the analysis of these 48 mutants, this first implementation of KINARI-Mutagen is able to make qualitative stability predictions. In the cases when residues are highly solvent exposed, KINARI-Mutagen is not as accurate, because such residues do not engage in as many stabilizing interactions as would be expected of them. Similarly, the pre-existing algorithm to detect hydrophobic interactions is not always accurate, when compared to the predicted hydrophobic interactions from a visual inspection. In future work, we plan to address this hydrophobic interaction algorithm.

5.4 Conclusions

In our two case studies, and in the analysis of 48 mutant protein structure files, we have shown that rigidity analysis of even the most simple *in silico* mutations to

glycine provides valuable information about the stability of a mutated protein, that could not have been inferred by other methods, such as SASA measurements, or by ranking of contributions to stabilizing interactions.

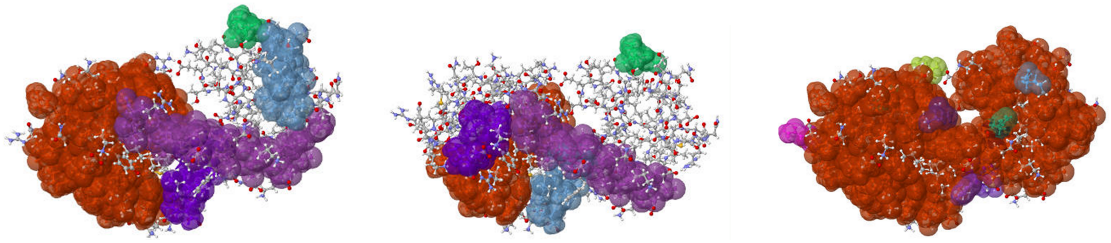
CHAPTER 6

TOWARDS VALIDATION OF MOLECULAR MODELING FOR RIGIDITY ANALYSIS

Rigidity analysis is a relatively new, alternative method, in which a protein structure is analyzed to infer which portions of the molecule are flexible. To perform rigidity analysis, a model is first constructed in which various inter-atomic stabilizing interactions are modeled according to their strength. *No systematic study has been conducted as to what is the most plausible, chemically validated modeling scheme.* All previous implementations have relied on heuristics, which allowed for extracting relevant observations but only for a very limited set of proteins. In this chapter, we describe how we use our KINARI-web server for protein rigidity analysis to systematically vary how stabilizing interactions are modeled. To our knowledge, this work is the first study that attempts to correlate rigidity metrics with experimental data, for a non-trivial sized dataset. We correlate rigidity results of proteins to experimentally derived biological data from the literature, in the form of $\Delta\Delta$G measurements, and we measure the correlation using a non-parametric approach.

6.1 Motivation and Introduction

Rigidity analysis of proteins was initially implemented in MSU-First [34, 33] and the first online tool was *FlexWeb* [79]. These were used to correlate rigidity results with physical properties of several proteins, but they required case-by-case visual inspections of the biomolecules that were involved. Moreover, the choice of modeling of hydrogen bonds and hydrophobic interactions in *FlexWeb* has been determined



(a) Hydrogen bonds modeled as 5 bars; hydrophobic interactions as 2

(b) Hydrogen bonds modeled as 4 bars; hydrophobic interactions as 2

(c) Hydrogen bonds modeled as 5 bars; hydrophobic interactions as 3

Figure 6.1. Choice of modeling affects rigidity results. Changing how hydrogen bonds and hydrophobic interactions are modeled drastically affects the rigidity results.

based on the analysis of a small set of proteins. *Up until now, no systematic study has been performed with the intent to determine a universal modeling for hydrogen bonds and hydrophobic interactions.* In fact, recent work [86] has demonstrated that there is no general agreement about what the correct modeling should be, so that rigidity results for a large protein dataset correlate with biological phenomena.

As is shown in Figure 6.1, even small changes to how hydrogen bonds and hydrophobic interactions are modeled can drastically alter the rigidity results. *Our goal is to correlate rigidity results of proteins to experimentally derived biological data from the literature, in the form of* $\Delta\Delta G$. The experimental data designates whether a variant of a protein is stable, and we use that as the ground-truth. We investigate which set of modeling options for hydrogen bonds and hydrophobic interactions produces rigidity results that correlate to this biological data. To do this, we have retrieved from the PDB the structure files of the wild type and 158 variants of Lysozyme from bacteriophage T4. For each variant, we have searched the ProTherm [42] database for its $\Delta\Delta$G value. We systematically varied how hydrogen bonds and hydrophobic interactions are modeled during rigidity analysis of the 158 variants. In this chapter we correlate rigidity results of our protein dataset to experimentally derived biological data from the literature, in the form of $\Delta\Delta$G.

6.2 Constructing a Dataset of Protein Data Files and Systematically Varying Modeling of Stabilizing Interactions

In order to have as rich as possible a dataset of protein structure files along with their experimentally measured data, we searched the ProTherm database [42], which catalogues numerical data of thermodynamic parameters of proteins. Of the more than 25,000 entries in the ProTherm database, 1,719 of them are for variants of lysozyme from bacteriophage T4, more than any other protein. Of those, we retained the 158 entries which had a $\Delta\Delta$G measurement and had corresponding structure files in the PDB. For all of these 158 variants, the reference protein (the wild-type, non-mutated form) was the protein structure in PDB file 2lzm [85]. The $\Delta\Delta$G values of these protein mutants ranged from very negative, meaning that a mutant was much less stable than the wild-type, to positive, indicating that the mutant was more stable than the wild-type.

In preparation for rigidity analysis, single and double covalent bonds were modeled in the associated graph as 5 bars and 6 bars, respectively. This modeling represents that single covalent bonds impose one degree of freedom between the corresponding atoms in the mechanical model of the protein, which is equivalent to allowing rotation along the bond, and that the double covalent bonds retain zero degrees of freedom and do not permit rotation.

Because there is no agreed-upon way of performing the mechanical modeling of hydrogen bonds and hydrophobic interactions, we used KINARI's customizable modeling feature to systematically vary the modeling of these constraints.

Both hydrophobic interactions and hydrogen bonds were modeled as 1, 2, 3, 4, 5, 6 bars, or as hinges, when building the mechanical framework. Because hydrogen bonds have associated with them energies that determine their strength, they were also modeled in seven additional ways: strong hydrogen bonds were modeled with more bars, than weaker bonds (Table 6.1). In previous implementations of software

Table 6.1. Modeling of hydrogen bonds according to their energies. Hydrogen bonds were binned into six energy categories (**ByEnergy1**), three energy categories (**ByEnergy2**), or into two energy categories (**ByEnergy3** through **ByEnergy7**). Bonds in a specific bin were modeled using the number of bars (or as hinges) designated in the **How modeled** column.

Scheme	Energy of hydrogen bond (kcal/mol)	How modeled
ByEnergy1	strength < -5	6 Bars
	-5 < strength < -4	Hinges
	-4 < strength < -3	4 Bars
	-3 < strength < -2	3 Bars
	-2 < strength < -1	2 Bars
	-1 < strength	1 Bar
ByEnergy2	strength < -4	Hinges
	-4 < strength < -1	4 Bars
	-1 < strength	3 Bars
ByEnergy3	strength < -2	4 Bars
	-2 < strength	Hinges
ByEnergy4	strength < -3	4 Bars
	-3 < strength	Hinges
ByEnergy5	strength < -4	4 Bars
	-4 < strength	Hinges
ByEnergy6	strength < -5	4 Bars
	-5 < strength	Hinges
ByEnergy7	strength < -6	4 Bars
	-6 < strength	Hinges

for rigidity analysis of proteins, all hydrogen bonds were modeled the same, irrespective of strength. Thus we had 13 different ways that hydrogen bonds were modeled, and 7 ways that hydrophobic interaction were modeled, for a total of $13 \times 7 = 98$ unique modeling settings. For each protein, 98 different mechanical frameworks were generated, and their rigidity analyzed using the KINARI software.

6.3 Correlating Rigidity Parameters to Experimental Data

In an attempt to gain insight into the possible correlation between rigidity results and experimental data, for each protein structure, for each of the 98 different combinations of modeling hydrogen bonds and hydrophobic interactions, we calculated three rigidity metrics: Cluster Configuration Entry (**CCE**, Section 2.4), Dominant

Rigid Cluster size (**DRC**), and Average Cluster Size (**ACS**). In this section we provide scatter plots for the rigidity metrics versus experimental results, and we calculate the Spearman's rank correlation coefficient to measure the extent of their statistical dependence. Finally, assuming a linear relationship, we propose a general method for correlating rigidity metrics with $\Delta\Delta$G measurements.

6.3.1 Scatter Plots of Rigidity Measurements and Experimental Data

To get a sense of the relationship between the calculated rigidity properties and experimental data for the 158 variants in our dataset, we generated scatter plots of the three rigidity metrics versus $\Delta\Delta$G measurements for the 98 combinations of modeling hydrogen bonds and hydrophobic interactions. Figure 6.2 shows the scatter plots for the Dominant Rigid Cluster versus $\Delta\Delta$G measurements. In Figure 6.3, we show the scatter plots for the $\Delta\Delta$G measurement versus Cluster Configuration Entropy (Section 2.4), and in Figure 6.4, we show the scatter plots for $\Delta\Delta$G versus Average Cluster Size. For the scatter plots, there is no discernible relationship between any of the three rigidity metrics and $\Delta\Delta$G.

6.3.2 Calculating Correlation Using Spearman's Rank Coefficient Testing

To see if there is some sort of monotonic, linear, or non-linear relationship between the calculated rigidity parameters and experimental $\Delta\Delta$G data that the scatter plots did not reveal, we performed statistical significance testing on our rigidity results. We calculated Spearman's rank correlation coefficient, or ρ [73]. It is a non-parametric measure of statistical dependence between two variables. We used it to assess how well the relationship between the rigidity metrics and experimental data can be described using a monotonic function. Specifically, we employed the *cor.test* function in R [67], which outputs a measure of association in the range [-1,1]. A Spearman correlation $\rho = 0$ indicates no association between the two variables, while a value of $\rho = 1$ results when the two variables being compared are monotonically related, even if their

relationship is not linear. A positive Spearman correlation coefficient corresponds to an increasing monotonic trend between the two variables, while a negative value of ρ corresponds to a decreasing monotonic trend. Note also that R did not output confidence intervals for the p-values because there were ties when the $\Delta\Delta$G and rigidity metrics were ranked.

In Table 6.2, we list the rigidity metrics and the modeling options for hydrogen bonds and hydrophobic interactions for which $|\rho|$ was the greatest. From that table, we see that the largest association between rigidity measurements and experimental data occurred for the Cluster Configuration Entropy (CCE) metric. When hydrogen bonds were modeled as 1 Bar, and hydrophobic interactions as 5 Bars, the correlation between CCE and $\Delta\Delta$G was -0.234. Although there are four ways of modeling hydrogen bonds and hydrophobic interactions so that the absolute value of the correlation between the rigidity metric and experimental data is greater than 0.2, note that values of -0.234, -0.227, -0.218, and -0.212 are each quite small and far removed from -1. This indicates that the correlation between experimental data and the CCE metric is small. Also, the negative correlation values for Average Cluster Size (ACS) and Dominant Rigid Cluster (DRC) are not what was expected. For those two metrics, an increase in the $\Delta\Delta$G measurement for a protein should correspond to an increase in the DRC and ACS. To the contrary, the values for the two ACS and one DRC value were -0.179, -0.179, and -0.171, respectively.

On the other hand, Cluster Configuration Entropy (CCE) is a function of the probability that a vertex in the mechanical model is part of a cluster of size s (explained in Section 2.4). Protein variants with high CCE values are more disordered than proteins with low CCE values. For the purpose of our experiments, proteins that have a high CCE value correlate well with experimental data whose $\Delta\Delta$G values are low. Thus, unlike for the DRC and ACS metrics, a negative slope for the model-

Table 6.2. Non-Parametric Spearman's Correlation Testing; Lowest p-values

Metric	HBonds Modeling	HPhobes Modeling	p-value	correlation
CCE	1Bar	5Bars	0.002	-0.234
CCE	3Bars	3Bars	0.003	-0.227
CCE	2Bars	5Bars	0.005	-0.218
CCE	1Bar	Hinge	0.007	-0.212
CCE	2Bars	Hinge	0.013	-0.195
ACS	E6	4Bars	0.023	-0.179
ACS	E1	4Bars	0.023	-0.179
CCE	3Bars	2Bars	0.026	-0.176
DRC	E6	4Bars	0.030	-0.171
CCE	2Bars	4Bars	0.031	-0.170

ing versus CCE metric value indicates the expected correlation between the rigidity metric and experimental data.

Finally, even with these low correlation measurements, notice that for the top five rows in Table 6.2, the modeling of hydrogen bonds and hydrophobic interactions was confined to a small region of the set of possible combinations of modeling these stabilizing interactions. Namely, both hydrogen bonds and hydrophobic interactions should not be both modeled as very few bars, nor should they both be modeled as too many bars. In any case, the low correlation values indicate that there is no convincing evidence that rigidity models correlate well with experimentally measured stability data given any of our rigidity metrics.

6.3.3 A General Method for Correlating $\Delta\Delta G$ With Rigidity Metrics, Assuming a Linear Relationship

Under the assumption that there is a linear correlation between a rigidity metric and experimentally derived $\Delta\Delta G$ measurements, we want to assess the predictive ability of our model with respect to $\Delta\Delta G$ for the 158 proteins. To do this, we tabulated how many of the proteins, for each of the 98 combinations of modeling options, would have been correctly identified as stable (relative to the wild-type)

using the rigidity metric (Algorithm 1). Using our approach, it is possible to rank the different modeling choices based on their ability to provide rigidity results that correlate with experimental data.

To illustrate on a small sample dataset, Table 6.3 lists the $\Delta\Delta$G, and rigidity DRC values, for three variants of lysozyme, for two separate modeling schemes that differ only in how the hydrogen bonds were modeled. Our quantitative correlation analysis of the two schemes reveals that the DRC metric would have correctly identified the stability of the three proteins if modeling scheme 1 were used. However, if modeling scheme 2 were used, then the quantitative correlation analysis indicates that the relative stability of only one of the three proteins would have been correctly identified using the DRC metric. In Figure 6.5, we plot $\Delta\Delta$G vs. the DRC metrics for the three proteins, for both modeling schemes. The red lines indicate slopes where the DRC metric did NOT positively correlate with the $\Delta\Delta$G values; green lines indicate proteins who have metric-$\Delta\Delta$G slopes for the variant and wild-type protein where the rigidity metric positively correlates with the experimental data.

Figures 6.6 through 6.8 show the quantitative correlation scores for the correlation of the three metrics versus the experimentally derived $\Delta\Delta$G values, for the 98 different ways that stabilizing interactions were systematically modeled using the dataset of 158 protein mutant structures and one wild-type.

In each of the Figures 6.6 - 6.8, the number at position x,y in the grid indicates for how many of the 158 proteins did the rigidity results quantitatively correlate with experimentally derived data from the literature. Boxes whose color tends towards yellow designate the hydrogen bond (y-axis) and hydrophobic modeling choice (x-axis) whose rigidity results correlated best with $\Delta\Delta$G data.

Table 6.3. Sample dataset for correlating experimental with rigidity data. Sample experimental ($\Delta\Delta$G) values and rigidity metrics (Dominant Rigid Cluster), for three variants of bacteriophage T4 lysozyme proteins 255l, 1l73, and 1l67, for two separate modeling schemes, were used. The wild-type protein, 2lzm, is the reference protein against which $\Delta\Delta$G values are reported for the three proteins in the literature. Our quantitative correlations indicates that modeling scheme 1 versus 2 generates rigidity metrics that correlate better with experimental data.

	Modeling Scheme 1	**Modeling Scheme 2**
Hydrogen Bonds	4 Bars	2 Bars
Hydrophobic Interactions	3 Bars	3 Bars
DRC_{2lzm}	510	641
DRC_{255l}	743	810
DRC_{1l73}	603	270
DRC_{1l67}	440	820
$\Delta\Delta G_{2lzm}$	0.00	
$\Delta\Delta G_{255l}$	+1.30	
$\Delta\Delta G_{1l73}$	+0.85	
$\Delta\Delta G_{1l67}$	-2.10	
Correct Count	3/3	1/3

Algorithm 1: Measuring Quantitative Correlation of Rigidity DRC with Experimental $\Delta\Delta$G. HP=Hydrophobic Interaction, HB=Hydrogen Bond

Input: Rigidity Results of proteins
Input: $\Delta\Delta$G values for all protein-mutant pairs
Initialize *correlationCorrectCounts*[][] matrix to be all 0s
foreach i^{th} *protein analyzed* **do**
  foreach j *modeling of HP* **do**
    foreach k *modeling HB* **do**
      DRC_{var} = size of DRC for i when modeling j, k
      DRC_{wt} = size of DRC for **wt** when modeling j, k
      ddg = $\Delta\Delta$G for **wt**, i pair
      if $DRC_{var} < DRC_{wt}$ **then**
        if $ddg < 0$ **then**
          $correlationCorrectCounts[j][k]++$
        end
      else
        if $ddg > 0$ **then**
          $correlationCorrectCounts[j][k]++$
        end
      end
    end
  end
end
Output: *correlationCorrectCounts* matrix

6.3.3.1 Evaluation of the Dominant Rigid Cluster Metric

The Dominant Rigid Cluster metric is the count of atoms that are in the largest rigid cluster. Figure 6.3.3.1 illustrates that there are no ways of modeling hydrogen bonds and hydrophobic interactions so that the DRC metric correlates positively with at least 100 of 158 analyzed proteins. When hydrogen bonds were modeled as 3 bars, and hydrophobic interactions as 6 Bars, then the DRC metric correctly identified the stability of 95 of the 158 proteins, respectively.

6.3.3.2 Evaluation of the Cluster Configuration Entropy Metric

The Cluster Configuration Entropy (CCE) is a function of the probability that a vertex in the mechanical model is part of a cluster of size s (explained in Section 2.4). Protein variants with high CCE values are more disordered than proteins with low CCE values. For the purpose of our experiments, proteins that have a high CCE value correlate well with experimental data whose $\Delta\Delta$G values are low.

Figure 6.3.3.2 indicates that there are multiple ways of modeling hydrogen bonds and hydrophobic interactions, so that at least 100 of the 158 analyzed proteins had rigidity CCE values that quantitatively correlated with the experimentally derived stability $\Delta\Delta$G data. From among these 8, when hydrogen bonds were modeled as 3 Bars, and hydrophobic interactions as 3 Bars, then the stability of 111 of the 158 variants would have been correctly identified using the CCE metric.

6.3.3.3 Evaluation of the Average Cluster Size Metric

The Average Cluster Size metric computes the average number of atoms that are contained among all of the rigid bodies identified in a protein. Figure 6.3.3.3 illustrates that there is only a single modeling combinations where the ACS metric would have correctly identified the stability at least 100 of the 158 proteins (Hydrogen Bonds modeled as 1 Bar, and Hydrophobic Interactions modeled as 6 Bars).

6.4 Conclusions

Based on the analysis of our method of systematically varying how hydrogen bonds and hydrophobic interactions are modeled, we observed several behaviors.

There is no one single choice of modeling of hydrogen bonds and hydrophobic interactions that would have given rigidity results that correlated positively for all 158 variants with $\Delta\Delta$G data. This is corroborated by the results of Wells, *et al* [86], in which he showed that the choice of modeling of hydrogen bonds and selecting of the hydrogen bond cutoff needs to be chosen on a protein case-by-case basis so that rigidity results can be verified against biological data. One possible explanation for this is that the size of the dominant rigid body is not the best metric as an indicator of a protein's stability. Similarly, correlating rigidity results to $\Delta\Delta$G data might require a multi-dimensional analysis that uses several rigidity metrics.

Moreover, we found that there is no choice of modeling of hydrogen bonds and hydrophobic interactions which would have generated DRC, CCE, and ACS rigidity metrics that would have quantitatively positively correlated with $\Delta\Delta$G data in at least 150 of the 158 variants in our dataset. There are certain reasons why this may be the case.

It may be that some critical stabilizing interactions are not identified in the protein's structure in building the mechanical model of the molecule. This would affect the rigidity results. Some hydrophobic interactions not being found could be caused by slight non-optimal packing of the atoms in the protein, which would preclude KINARI's hydrophobic detection algorithm from identifying critical interactions. Secondly, there may be yet another scheme of modeling hydrogen bonds according to their energies, which would yield rigidity results that 100% of the time correlate quantitatively with experimental data. Lastly, hydrophobic interactions may also need to be modeled according to energy, and not all the same way, as was done in

these experiments. At the current time, modeling of stabilizing interactions according to energy in the KINARI software is available only for hydrogen bonds.

Although we did not identify a single choice of modeling of hydrogen bonds and hydrophobic interactions which generated rigidity results that positively correlated with $\Delta\Delta$G data in all of our proteins in our dataset, we have demonstrated our method (Algorithm 1) in correlating rigidity metrics to experimental data. Moreover, our method is not dependent on a case-by-case analysis of the proteins that were studied, but instead requires only experimental data (here $\Delta\Delta$G), and rigidity results. As such, we believe that our method can be used with other rigidity metrics, possible other experimental data, as well as on other, potentially larger protein datasets.

Figure 6.2. Scatter plots for Change of the Dominant Rigid Cluster versus ΔΔG. The y-axis of each plot designates the Change in the Dominant Rigid Cluster (DRC) metric, while the x-axis designates the ΔΔG values, of the proteins in the dataset. The left-most label for each row indicates how hydrogen bonds were modeled, and the bottom-most label for each column designates how hydrophobic interactions were modeled in that column.

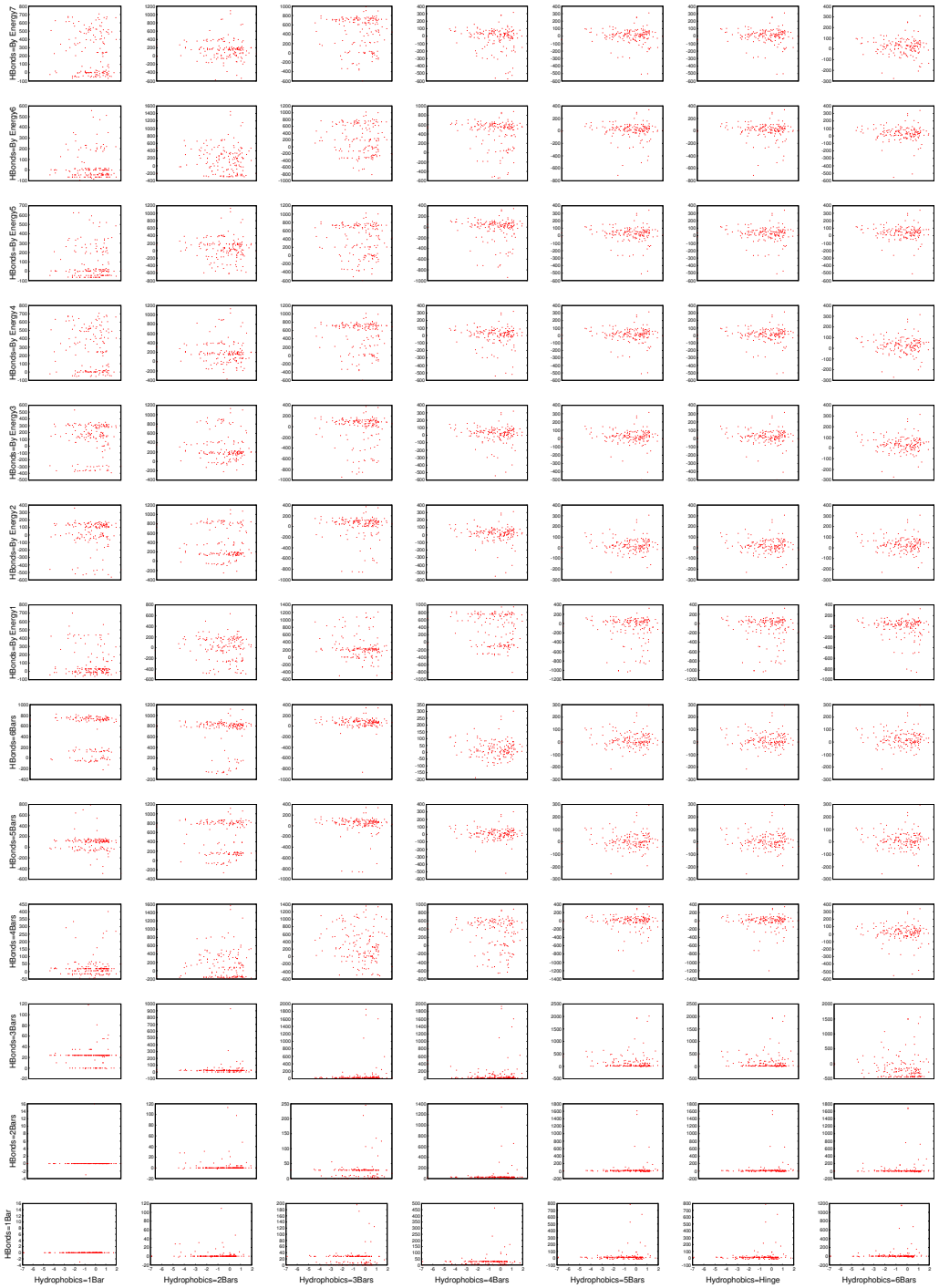


Figure 6.3. Scatter plots for Change in Cluster Configuration Entropy versus ΔΔG. The y-axis of each plot designates the Change in the Cluster Configuration Entropy (CCE) metric, while the x-axis designates the ΔΔG values, of the proteins in the dataset. The left-most label for each row indicates how hydrogen bonds were modeled, and the bottom-most label for each column designates how hydrophobic interactions were modeled in that column.

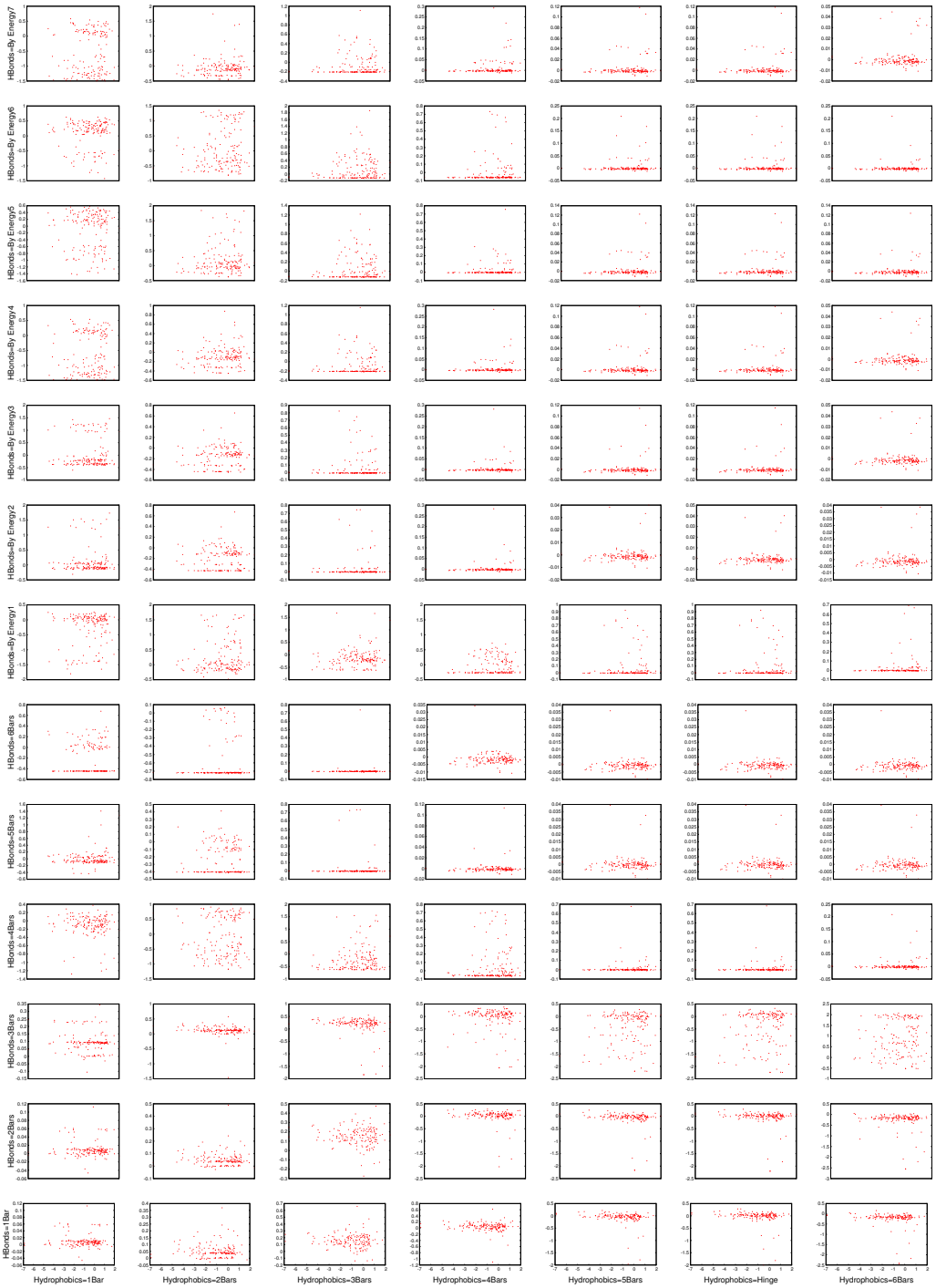
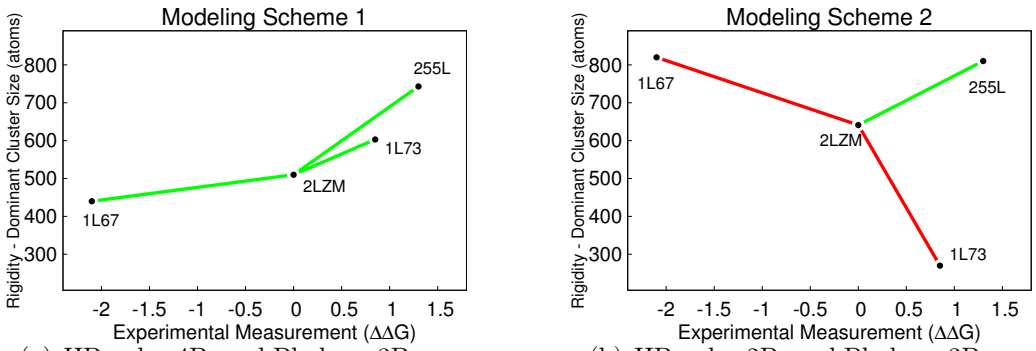


Figure 6.4. Scatter plots for Change of the Average Cluster Size versus ΔΔG. The y-axis of each plot designates the Change in the Average Cluster Size metric, while the x-axis designates the ΔΔG values, of the proteins in the dataset. The left-most label for each row indicates how hydrogen bonds were modeled, and the bottom-most label for each column designates how hydrophobic interactions were modeled in that column.



(a) HBonds=4Bars; hPhobes=3Bars (b) HBonds=2Bars; hPhobes=3Bar

Figure 6.5. Correlating rigidity metrics with experimental data. When Modeling Scheme 1 (a) is used, the DRC rigidity metric correctly identifies the stability of the three proteins relative to the wild-type, 2lzm. When Modeling Scheme 2 (b) is used, the DRC metric incorrectly identifies protein 1L67 as more stable than 2lzm (the negative $\Delta\Delta$G value for 1l67 indicates that the protein is less stable than 2lzm), and protein 1l73 is incorrectly identified as less stable than 2lzm, (the positive $\Delta\Delta$G for 1l73 designates that it is more stable than 2lzm).

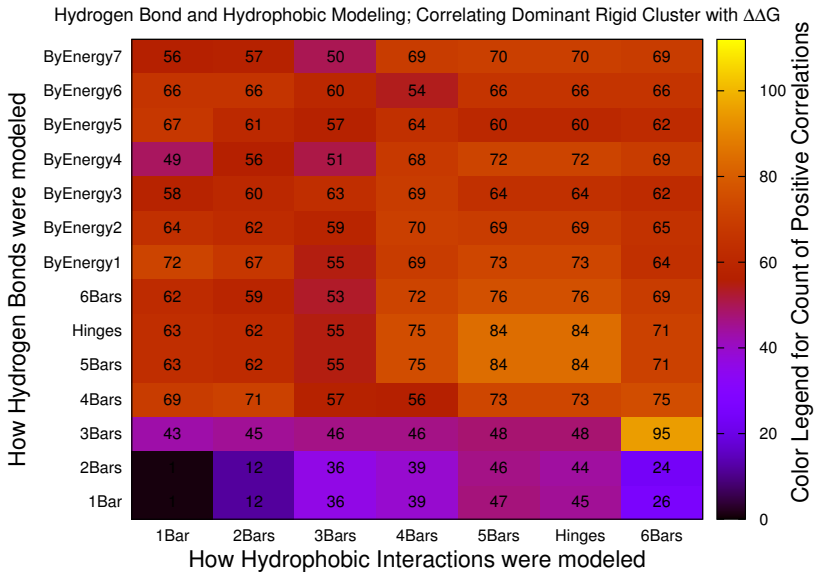


Figure 6.6. Quantitative correlations for Dominant Rigid Cluster and $\Delta\Delta$G. The number at position x,y in the grid indicates for how many of the 158 proteins did the rigidity results quantitatively correlate with experimentally derived data from the literature. Boxes whose color tends towards yellow designate the modeling choices whose rigidity results correlated best with $\Delta\Delta$G data.

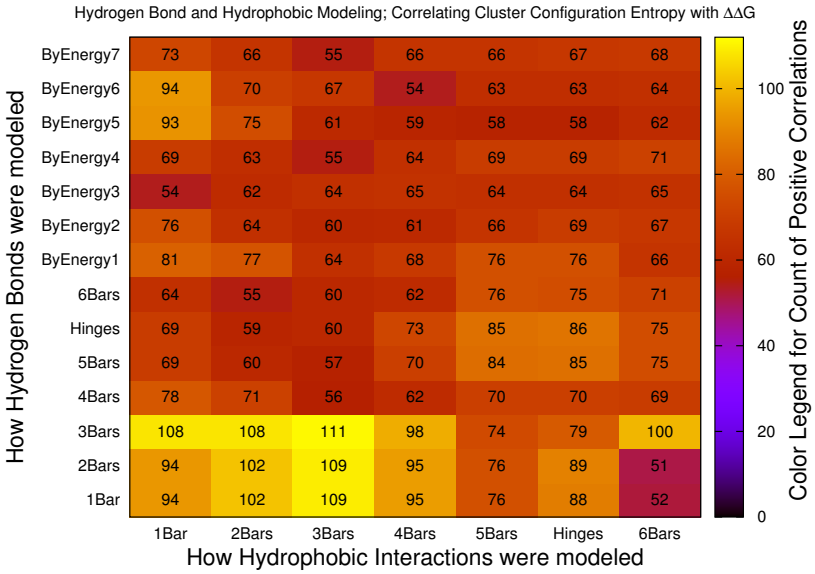


Figure 6.7. Quantitative correlations for Cluster Configuration Entropy and ΔΔG. The number at position x,y in the grid indicates for how many of the 158 proteins did the rigidity results quantitatively correlate with experimentally derived data from the literature. Boxes whose color tends towards yellow designate the modeling choices whose rigidity metric values correlated best with ΔΔG data.

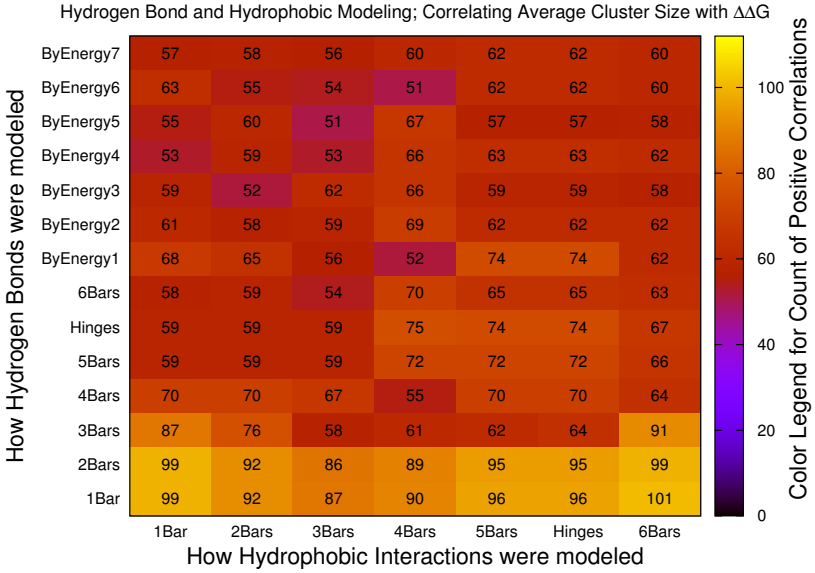


Figure 6.8. Quantitative correlations for Average Cluster Size and ΔΔG. The number at position x,y in the grid indicates for how many of the 158 proteins did the rigidity results quantitatively correlate with experimentally derived data from the literature. Boxes whose color tends towards yellow designate the modeling choices whose rigidity metric values correlated best with ΔΔG data.

CHAPTER 7

CONCLUSIONS

7.1 Summary of Contributions

Proteins bend and flex, and interact with other molecules, in order to perform their functions. Scientists would like to understand, on an atomic level, how proteins move. Having knowledge of where and how proteins bend and flex can guide the design of drugs aimed to regulate proteans associated with diseases. Unfortunately there are not existing experimental methods that permit observing on the atomic level, in real-time, how proteins bend and flex. To gain insight into these motions, simulation based methods have been developed, but unfortunately they are computationally intensive.

Rigidity analysis is an alternative, complimentary approach to simulation methods. Its goal is not to predict or simulate motion, but instead to infer which parts of a protein are rigid, and which are flexible. In rigidity analysis, a protein's atoms and chemical interactions are used to build a mechanical model, which is associated to a graph composed of nodes that represent atoms, and edges that correspond to chemical constraints.

Rigidity analysis of proteins was first implemented in *MSU-First* and the first online tool was *FlexWeb*. Beginning in the late 1990s, the usefulness of rigidity analysis was demonstrated in inferring various structural and functional properties of proteins. Many such studies relied on heuristics to determine which choice of modeling settings of important stabilizing interactions allowed for extracting relevant biological observations from rigidity analysis results of a small set of proteins. This is one reason why large-scale validate of protein flexibility has not been performed.

Also, experimental methods such as X-ray crystallography produce the asymmetric unit, which is the smallest portion of a crystallized protein on which symmetry operations can be applied to reproduce the crystal lattice. The asymmetric unit most often does not represent the biological functional form of a protein. The generate the biological form of a protein, its asymmetric unit has be translated, rotated, copied, etc. If done by hand, it is a time consuming process. *MSU-First* and *FlexWeb* do not provide tools to generate the biological assembly of a protein, so performing rigidity analysis on large datasets of biological forms of proteins cannot be done easily using those tools.

Also, because *MSU-First* and *FlexWeb* do not provide the user with easily accessible options to designate how important stabilizing interactions should be modeled in the mechanical model of a protein, these tools cannot be used to perform large-scale studies to infer how changing the modeling of these interactions affects the rigidity results. A consequence of this is that there is no agreed-upon choice of how chemical interactions should be modeled in the mechanical framework of a protein.

In this thesis, we have made progress in addressing some of these obstacles, which prevent high-throughput, large-scale validation of using rigidity analysis to infer protein flexibility. To achieve that, we have developed the KINARI software. This has allowed us to generate and study the rigidity of a large set of biological assemblies. Also, because KINARI is highly customizable, we've performed the first systematic study to investigate the modeling of hydrogen bonds and hydrophobic interactions so that rigidity results correlate with experimental data. The specifics of each contribution are described below.

7.1.1 KINARI: Infrastructure for Rigidity Analysis of Proteins

The first tools that implemented rigidity analysis of proteins offered few options for curating PDB data files, and the choices of modeling of important stabilizing

interactions were fixed. To provide an infrastructure to easily test if and how rigidity analysis can function as a predictive tool for inferring biophysical properties of proteins, we have developed KINARI-Web. It is a general, well-tested, versatile web server for rigidity analysis of molecular structures. It relies on a mechanical model of a protein that is customizable by the user, it performs rigidity analysis of the mechanical framework, and it includes an interactive visualizer for exploring the rigidity results. Moreover, the release of the C++ libraries for rigidity analysis allows a researcher to easily integrate these tools into custom-made scripts meant for high-throughput experiments of protein rigidity. The benchmarking experiments of more than 25,000 proteins that were performed as part of the this dissertation are an example of the use of these freely-available tools.

7.1.2 Inferring Structural and Functional Information of Protein Biological Assemblies and Crystals

PDB files contain only the asymmetric unit, which is the smallest part of a crystal on which symmetry operations are applied to generate a crystal lattice and biological form of a protein. The majority of previous rigidity-theoretic studies of protein flexibility analyzed these asymmetric units. We extended KINARI and developed the KINARICrystal and BioAssembly tools for generating crystal lattices and biological assemblies from PDB structure files. With the features of KINARI that were not available prior to the work presented in this thesis, it is now possible to perform larger scale studies of the rigidity properties of the biological assemblies of proteins. Generating the biological forms of proteins is now easily done using the KINARICrystal and BioAssembly feature that are integrated into the Curation feature of KINARI-Web.

As a demonstration, we have performed rigidity analysis of over 900 crystal lattices and biological assemblies that we generated using these new tools. We've shown that

when rigidity analysis is performed on only the asymmetric unit or just isolated portions of a protein, then structural and functional information is missed.

7.1.3 KINARI-Mutagen: Inferring Critical Residues

To further expand the use of rigidity analysis in inferring structural and biological properties of proteins, we developed KINARI-Mutagen. It infers which residues are critical in maintaining a protein's stability. The interpretation of the rigidity results is not dependent on any in-depth, case-by-case knowledge of the biophysical properties of studied protein. KINARI-Mutagen permits fast evaluation of *in silico* mutations that may not be easy to perform *in vitro*. For two cases studies and a dataset of 48 proteins, we have shown that KINARI-Mutagen identifies critical residues that would not have been easily identified using existing methods, or by ranking of residues by their involvement in hydrogen bonds or hydrophobic interactions.

7.1.4 Correlating Rigidity Parameters to Experimental Data

A large-scale study correlating rigidity metrics to experimental data has not been performed up until now. In Chapter 6, we have explain our method in which we systematically varied how hydrogen bonds and hydrophobic interactions were modeled for a dataset of 158 variants of lysozyme from bacteriophage T4. In correlating three rigidity metrics for each of the proteins against $\Delta\Delta$G data, we have found that there is no one single "best" choice of modeling hydrogen bonds and hydrophobic interactions. However, for our dataset, there were a few modeling schemes so that the rigidity metrics for more than 100 of the 158 variants correlated against $\Delta\Delta$G data.

Although we did not identify a single choice of modeling of hydrogen bonds and hydrophobics which generated rigidity results that positively correlated with $\Delta\Delta$G data in all of the protein structures that we studied, we have demonstrated the use of our method in correlating rigidity metrics to experimental data. In addition, we have shown that there are several combinations of modeling hydrogen bonds and

hydrophobic interactions so that the Cluster Configuration Entropy metric correlates better with experimental data than the Dominant Rigid Cluster metric. Moreover, our method is not dependent on a case-by-case analysis of the studied proteins, but instead requires only experimental data (here $\Delta\Delta G$), and rigidity metrics. As such, with this method, we have provided a general, unbiased approach to correlate rigidity metrics with experimental data. This now permits ranking rigidity metrics based on how well they correlate with experimental data.

7.2 Future Directions

In the course of the work leading up to this dissertation, several future research directions were identified. We describe a few of them here.

Rigidity of Protein Biological Assemblies and Crystal Structures

The crystal lattices that were generated using KINARICrystal were relatively small, at most 2×2×2 unit cells. However, even these small crystals contained many atoms (the largest lattice contained 54,107 atoms (PDB file 3hon, Table 4.2)). The reason why larger crystals were not generated and analyzed was because curation, modeling, and parts of the rigidity analysis required upwards of 10 minutes of run-time when analyzing structure files with more than 10,000 atoms.

Several advancements to the software might be made. Firstly, stabilizing interactions do not need to be computed for every unit cell. Instead the symmetry among unit cells might be taken advantage of, which would require calculating interactions for one unit cell only, under the assumption that the same interactions would exist in other unit cells. If such a scheme were used, interactions would need to be also identified in the boundary areas where unit cells abut. Secondly, a systematic study could be performed on a dataset of biological assemblies. The KINARI curation tools permit a user to easily generate components of a biological assembly. That feature

could be used to classify proteins based on the degree to which each subunit contributes (if at all) to the stability of the entire biological assembly.

Using Rigidity Analysis to Infer Which Residues are Critical In Stabilizing a Protein's Structure

KINARI-Mutagen performs *in silico* mutations to glycine, only, and calculates their effects on the protein's rigidity. The mutation engine can be expanded to permit generating amino substitutions to other residues. Doing so would permit validating KINARI-Mutagen against an even larger dataset of proteins, for which mutations to a host of different residues have been performed.

Using other rigidity metrics, such as Cluster Configuration Entropy and Average Cluster size, as predictors of which residues are critical, might permit identifying important residues that the current version of the software missed. Moreover, in order for KINARI-Mutagen to quantitatively predict the role of residues in stabilizing a protein, a multi-dimensional analysis that incorporates several rigidity metrics, might be required.

Correlating Rigidity Metrics with Experimental Data, and Evaluating How Stabilizing Interactions Should Be Modeled

In this thesis, the hydrogen bonds and hydrophobic interactions were systematically varied, and the resulting rigidity metrics were correlated with experimental data. In our studies, no one single universal choice of modeling of these stabilizing interactions was identified, that enabled any of the three rigidity metrics to always predict the stability of a variant protein structure. One possible extension to our method would entail modeling hydrophobic interactions according to their energies, just as we did for hydrogen bonds.

APPENDIX A

RIGIDITY RESULTS OF PROTEIN BIOLOGICAL ASSEMBLIES AND CRYSTAL LATTICES

Table A.1. Rigidity results for putative protein from the gram-negative bacterium *Thermus thermophilus*. The count of the sizes of the rigid clusters for PDB file 2yzt are shown for the asymmetric unit (AU, column 2), the unit cell (column 3), the 2×1×1 crystal (column 4), and 2×2×1 crystal (column 5)

Size (atoms) of rigid cluster	**AU**	**111.2yzt**	**211.2yzt**	**221.2yzt**
3	4	26	49	91
4	21	106	201	308
5	122	632	1165	2126
6	22	88	133	1880
7	1	4	6	8
11	5	26	49	93
12	2	8	12	16
26	1	4	6	8
463	1	2	3	4
2504	0	1	0	0
6084	0	0	1	0
14328	0	0	0	1

Table A.2. Rigidity results of the scaffolding protein of Vaccinia Virus. The number of each type of rigid cluster in PDB file 3saq is listed for the asymmetric (AU) and biological units (BU). Column 2 lists the count of the different sizes of the rigid bodies in the asymmetric unit, which contains one copy of chain A and B. Columns 3 and 5 list the counts of the different sizes of the rigid bodies in one-third of the two biological units, respectively. Columns 4 (three copies of chain A) and 6 (three copies of chain B) list the counts of the different sizes of the rigid bodies for the two complete biological assemblies.

Size (atoms) of rigid cluster	**AU**	**BU1a**	**BU1**	**BU2a**	**BU2**
3	321	303	911	140	434
4	77	45	132	40	117
5	1462	905	2715	696	2117
6	179	187	561	66	217
7	1	4	12	1	4
11	16	12	36	8	25
12	45	31	93	19	56
13	2	1	3	1	3
15	2	3	9	0	0
16	2	1	3	1	3
19	7	8	24	2	7
22	1	2	6	0	0
25	1	2	6	0	0
33	1	1	3	0	1
38	1	1	0	0	0
42	0	0	3	0	0
48	1	1	3	0	0
71	1	1	3	0	0
98	2	1	3	1	3
104	1	2	6	0	1
2277	0	1	3	0	0
3912	0	0	0	0	1
4562	0	0	0	1	0
7883	1	0	0	0	0
9475	0	0	0	0	1

Table A.3. Rigidity results for Nucleoprotein from Rift Valley Fever Virus. The first biological of PDB file 3ouo assembly is a hexamer, where each of the 6 units is a dimer (chains A and B). The second biological assembly is made up of six copies of Chain C. The number of each size (column 1, number of atoms) of rigid cluster is listed for the asymmetric (AU) and biological (BU) units. Column 2 designates the number of rigid clusters of the asymmetric unit, which contains one copy of chains A, B, and C. Columns 3 and 4 list the count of the rigid clusters for chains A and B of the first biological assembly. Columns 5 and 8 list the counts of the rigid bodies of the dimer (chains A and B) and the monomer (chain C) in the asymmetric unit, respectively. Column 7 lists the counts of the rigid clusters of two copies of the monomer (chain C). Columns 6 and 9 list the counts of the rigid bodies for the first and second complete biological assemblies, respectively.

Size (atoms) of rigid cluster	AU	BU1a	BU1b	BU1ab	BU1	BU2cc	BU2c	BU2
3	215	67	66	142	454	72	148	458
4	91	33	29	62	186	30	60	182
5	1289	429	431	854	2556	435	854	2520
6	159	52	57	107	318	53	102	300
7	10	3	3	8	27	2	4	12
11	32	10	10	19	57	13	25	72
12	31	9	12	20	57	11	21	60
13	1	1	0	1	3	1	2	6
15	5	3	2	4	12	1	1	0
16	11	4	3	7	18	4	9	30
17	1	0	0	0	0	1	1	0
19	5	2	1	3	9	2	4	12
22	6	2	2	4	12	2	4	12
23	1	0	0	0	0	1	2	6
30	2	0	2	2	6	0	0	0
38	1	1	0	1	3	0	0	0
39	0	0	0	0	0	1	2	6
55	1	1	0	1	3	0	0	0
56	2	0	0	0	0	2	4	10
57	3	1	1	2	6	1	2	6
58	3	0	1	1	3	1	2	6
60	0	0	0	0	0	0	0	2
64	3	1	1	2	6	1	2	6
66	1	1	0	1	3	0	0	0
73	1	0	1	1	3	0	0	0
86	1	0	1	1	0	0	0	0
89	0	1	0	0	0	0	0	0
90	0	0	1	0	0	0	0	0
91	2	1	0	1	3	1	2	6
92	1	0	0	0	0	2	3	6
93	2	1	0	1	3	0	0	0
97	1	0	0	0	0	1	1	0
100	1	0	1	1	3	0	0	0
105	1	0	0	0	0	1	2	6
111	1	0	1	1	6	0	0	0
113	1	0	1	1	3	0	0	0
115	1	1	0	1	0	0	0	0
118	1	1	0	1	0	0	1	6
122	1	0	0	0	0	1	1	0
152	0	0	1	0	0	0	0	0
174	1	2	0	1	3	0	0	0
175	1	0	0	1	3	0	0	0
187	1	0	1	1	3	0	0	0
197	1	0	0	0	0	1	1	0
221	1	0	0	0	0	1	2	6
237	0	0	0	0	3	0	0	0
277	1	0	0	0	0	1	1	0
381	1	0	0	1	3	0	0	0
536	1	0	1	1	3	0	0	0
585	1	1	0	1	3	0	0	0
737	0	0	0	0	0	0	1	6

Table A.4. Rigidity results for Type III Antifreeze Protein RD1. The number of each type of cluster for PDB file 1ucs is shown for the asymmetric unit (AU, column 2), the unit cell (column 3), the 2×1×1 crystal (column 4), and the 2×2×1 crystal (column 5).

Size (atoms) of rigid cluster	AU	111.1ucs	211.1ucs	221.1ucs
2	1	5	11	22
3	63	255	511	1029
4	8	32	64	128
5	181	718	1434	2868
6	46	189	375	747
7	3	14	30	63
8	4	16	32	64
11	1	4	8	16
12	1	4	8	16
19	3	12	24	48
23	1	3	5	10
27	1	4	8	16
36	0	1	3	6
45	1	4	8	16
67	1	4	8	16

Table A.5. Rigidity results for Ribonuclease A. The count of different sized rigid clusters of PDB file 5rsa is show for the asymmetric unit (AU, column 2), the unit cell (column 3), the 2×1×1 crystal (column 4), and the 2×2×1 crystal (column 5).

Size (atoms) of rigid cluster	AU	111.5rsa	211.5rsa	221.5rsa
3	62	124	248	496
4	20	40	80	160
5	386	772	1544	3088
6	31	62	124	248
10	3	6	12	24
11	3	6	12	24
12	6	12	24	48
19	1	2	4	8
21	1	2	4	8
22	1	2	4	8
24	1	2	4	8
25	2	4	8	16
29	1	2	4	8
35	1	2	4	8
65	1	2	4	8

APPENDIX B

EXPERIMENTAL AND RIGIDITY DATA FOR 48 MUTANT PROTEINS ANALYZED BY KINARI-MUTAGEN

Table B.1. Protein structures with no stabilizing interactions at substitution. For these, the wild-type residue did not engage in stabilizing interactions, so *in silico* mutating the residue was not expected to change the rigidity results. DRC=Dominant Rigid Cluster; HPhobe=Hydrophobic Interaction; HBond=Hydrogen Bond.

PDB file	Protein	Wild-type Residue	Wild-type Residue Hydrophobicity	Wild-type Residue SASA (%)	$\Delta\Delta$G from ProTherm	Mutant Loss of HBonds	Mutant Loss of HPhobes	Change to DRC (atoms) upon mutation
1stn	Staphylococcal Nucl.	18,I	very	59	-2.6	0	0	0
1stn	Staphylococcal Nucl.	37,L	very	9	-3.9	0	0	0
1stn	Staphylococcal Nucl.	60,A	slightly	77	-1.5	0	0	0
1stn	Staphylococcal Nucl.	62,T	-	0	-3.4	0	0	0
1rtb	Thymidylic Acid	63,V	very	41	-3.5	0	0	0
1rtb	Thymidylic Acid	64,A	slightly	77	-0.44	0	0	0
1lz1	Human Lysozyme	2,V	very	68	-2.3	0	0	0
3mbp	Maltodextrin-Binding	276,A	slightly	0	-1.5	0	0	0
2rn2	Ribonuclease H	52,A	slightly	1	-2.7	0	0	0
3ssi	Streptomyces Subtilisin Protease Inh.	73,M	slightly	98	-0.49	0	0	0
1bvc	Biliverdin apomyoglobin	8,Q	-	59	-0.5	0	0	0
1ftg	Apo Flavodoxin	84,A	slightly	0	-1.25	0	0	0
1cto	Granulocyte	45,V	very	62	-1.9	0	0	0

Table B.2. Protein structures with too fee stabilizing interactions at substitution. For these structures, the wild-type residue engaged in fewer than exepcted hydrogen bonds or hydrophobic interactions, so *in silico* mutating the residue was not expected to affect the rigidity results as much as if all expected hydrogen bonds and/or hydrophobic interactions were detected. DRC=Dominant Rigid Cluster; HPhobe=Hydrophobic Interaction; HBond=Hydrogen Bond.

PDB file	Protein	Wild-type Residue	Wild-type Residue Hydrophobicity	Wild-type Residue SASA (%)	$\Delta\Delta$G from ProTherm	Mutant Loss of HBonds	Mutant Loss of HPhobes	Change to DRC (atoms) upon mutation
1stn	Staphylococcal Nucl.	61,F	medium	34	-4.7	0	0	0
3ssi	Streptomyces Subtilisin Protease Inh.	103,M	slight	0	-6.8	0	0	0
1stn	Staphylococcal Nucl.	36,L	very	0	-5.4	0	2	0
1rtb	Thymidylic Acid	54,V	very	0	-4.87	0	1	0

Table B.3. Protein structures with sufficient stabilizing interactions at substitution. For these structures, the wild-type residue engaged in as many hydrogen bonds or hydrophobic interactions as was expected via a visual inspection, so *in silico* mutating the residue was expected to affect the rigidity results. In all but one case (protein 1bpi, residue 35), the change to the DRC positively correlated with the experimental $\Delta\Delta G$ value. DRC=Dominant Rigid Cluster; HPhobe=Hydrophobic Interaction; HBond=Hydrogen Bond

PDB file	Protein	Wild-type Residue	Wild-type Residue Hydrophobicity	Wild-type Residue SASA (%)	$\Delta\Delta G$ from ProTherm	Mutant Loss of HBonds	Mutant Loss of HPhobes	Change to DRC (atoms) upon mutation
1rtb	Thymidylic Acid	47,V	very	22	-7.35	1	2	6
1rtb	Thymidylic Acid	108,V	very	4	-7.29	0	3	9
1bpi	Trypsin	43,N	-	15	-5.7	3	0	7
1rtb	Thymidylic Acid	57,V	very	14	-5.52	0	3	12
1bpi	Trypsin	35,Y	-	7	-5.0	0	3	0
1rtb	Thymidylic Acid	81,I	very	34	-4.81	0	2	6
1bpi	Trypsin	44,N	-	20	-4.7	1	2	18
1stn	Staphylococcal Nucl.	76,F	medium	25	-4.7	0	1	13
1lz1	Human Lysozyme	59,I	very	2	-3.83	0	1	6
1stn	Staphylococcal Nucl.	95,D	-	67	-3.1	1	0	5
1stn	Staphylococcal Nucl.	83,D	-	63	-2.8	4	0	49
1rtb	Thymidylic Acid	118,V	very	20	-2.7	0	1	6
1iob	Interleukin1	9,T	very	4	-2.6	1	1	7
2rn2	Ribonuclease H	68,S	-	2	-2.4	2	0	12
1lz1	Human Lysozyme	38,Y	-	12	-2.32	0	4	16
1stn	Staphylococcal Nucl.	77,D	-	0	-2.2	4	0	9
2abd	Acyl-coenzyme A	9,A	slight	12	-1.8	0	3	3
2abd	Acyl-coenzyme A	34,A	slight	0	-1.57	0	4	3
1pga	Streptococcal G	53,T	-	38	-1.2	1	1	8
1rtb	Thymidylic Acid	16,V	very	18	-1.18	0	2	9
3mbp	Maltodextrin-Binding	8,V	very	31	-1.0	0	1	6
1rtb	Thymidylic Acid	109,A	slight	8	-0.43	0	2	3
1ank	Adenylate Kinase	88,R	-	36	-0.2	2	1	19

Table B.4. Structure files with solvent exposed mutation points. Mutants for which the wild-type residue at the mutation point was more than 50% solvent exposed were not expected to be identified as critical using KINARI-Mutagen. DRC=Dominant Rigid Cluster; HPhobe=Hydrophobic Interaction; HBond=Hydrogen Bond.

PDB file	Protein	Wild-type Residue	Wild-type Residue Hydrophobicity	Wild-type Residue SASA (%)	$\Delta\Delta G$ from ProTherm	Mutant Loss of HBonds	Mutant Loss of HPhobes	Change to DRC (atoms) upon mutation
1rtb	Thymidylic Acid	93,P	-	69	-2.6	0	0	0
1lz1	Human Lysozyme	103,P	-	94	-0.1	0	0	0
1lz1	Human Lysozyme	78,H	-	86	-0.12	0	0	0
1rtb	Thymidylic Acid	114,P	-	81	-3.6	0	0	0
1lz1	Human Lysozyme	74,V	very	62	-0.22	0	0	0
1lz1	Human Lysozyme	71,P	-	57	-1.6	0	0	0
1iob	Interleukin1	97,P	-	51	-1.2	0	0	0
3ssi	Streptomyces Subtilisin Protease Inh.	13,V	very	56	-9.3	0	0	0

BIBLIOGRAPHY

[1] Abyzov, A., Bjornson, R., Felipe, M., and Gerstein, M. RigidFinder: a fast and sensitive method to detect rigid blocks in large macromolecular complexes. *Proteins 78*, 2 (February 2010), 309–324.

[2] Alber, T., Dao-pin, S., Nye, J.A., Muchmore, D.C., and Matthews, B.W. Temperature-sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. *Biochemistry 26*, 13 (1987), 3754–3758.

[3] Alber, T., Dao-pin, S., Wozniak, J.A., Cook, S.P., and Matthews, B.W. Contributions of hydrogen bonds of Thr 157 to the thermodynamic stability of phage T4 lysozyme. *Nature 330* (November 1987), 41–46.

[4] Alder, B.J., and Wainwright, T.E. Studies in molecular dynamics. I. General method. *Journal of Chemical Physics 31*, 2 (1959), 459–466.

[5] Bahar, M., Graham, S., Stuart, D., and Grimes, J. Insights into the evolution of a complex virus from the crystal structure of vaccinia virus D13. *Structure 19* (July 2011), 1011–1020.

[6] Bell, J.A., Becktel, W.J., Sauer, U., Baase, W.A., and Matthews, B.W. Dissection of helix capping in T4 lysozyme by structural and thermodynamic analysis of six amino acid substitutions at Thr 59. *Biochemistry 31*, 14 (1992), 3590–3596.

[7] Bello, J., and Nowoswiat, E.F. The activity of crystalline ribonuclease A. *Biochimica et Biophysica Acta (BBA) - Enzymology and Biological Oxidation 105*, 2 (1965), 325–332.

[8] Berova, N., Nakanishi, K., and Woody, R. *Circular Dichroism: Principles and Applications*. Wiley-VCH, 2000.

[9] Brown, S., Fawzi, N.J., and Head-Gordon, T. Coarse-grained sequences for protein folding and design. *Proceedings of the National Academy of Sciences 100*, 19 (2003), 10712–10717.

[10] Capriotti, E., Fariselli, P., and Casadio, R. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics 20, Supplemental* (2004), i63–i68.

[11] Center for Biological Physics, Arizon State University. FIRST 6.2.1 User Guide. Available at: http://flexweb.asu.edu/ (10 June 2011, date last accessed).

[12] Cheng, J., Randall, A., and Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *PROTEINS: Structure, Function, and Bioinformatics 62* (2006), 1125–1132.

[13] Clark, P., Grant, J., Monastra, S., Jagodzinski, F., and Streinu, I. Periodic rigidity of protein crystal structures. In *2nd IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS'12)* (February 2012).

[14] Clegg, Robert M. Forster resonance energy transfer–FRET what is it, why do it, and how it's done. In *Fret and Flim Techniques*, T.W.J. Gadella, Ed., vol. 33 of *Laboratory Techniques in Biochemistry and Molecular Biology*. Elsevier, 2009, pp. 1–57.

[15] Diez, M, Zimmermann, B., Borsch, M, Konig, M., Schweinberger, E., Steigmiller, S., Reuter, R., Felekyan, S., Kudryavtsev, V., Seidel, C.A., and Graber, P. Proton-powered subunit rotation in single membrane-bound F0F1-ATP synthase. *Nature Structural and Molecular Biology 11*, 5 (2004), 135–41.

[16] Ebihara, A., Manzoku, M., Iino, H., Kanagawa, M., Shinkai, A., Yokoyama, S., and Kuramitsu, S. Crystal structure of uncharacterized protein ttha1756 from thermus thermophilus hb8: Structural variety in upf0150 family proteins. *Proteins: Structure, Function, and Bioinformatics 71* (2008), 2097–2101.

[17] Echols, N., milburn, D., and Gerstein, M. MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Research 31* (2003), 478–482.

[18] Eriksson, A.E., Baase, W.A., Zhang, X.J., Heinz, D.W., Baldwin, E.P., and Matthews, B.W. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science 255* (1992), 178–183.

[19] Ferron, F., Li, Z., Danek, E.I., Luo, D., Wong, Y., Coutard, B., Lantez, V., Charrel, R., Canard, B., Waiz, T., and Lescar, J. The hexamer structure of the rift valley fever virus nucleoprotein suggests a mechanism for its assembly into ribonucleoproteiin complexes. *PLoS Pathogens 7*, 5 (May 2011).

[20] Fox, N., Jagodzinski, F., Li, Y., and Streinu, I. KINARI-Web: A server for protein rigidity analysis. *Nucleic Acids Research 39 (Web Server Issue)* (2011), W177–W183.

[21] Fox, N., Jagodzinski, F., and Streinu, I. Kinari-lib: a C++ library for pebble game rigidity analysis of mechanical models. In *Minisymposium on Publicly Available Geometric/Topological Software, Chapel Hill, NC, USA* (June 2012).

[22] Fox, N., and Streinu, I. Towards accurate modeling for protein rigidity analysis. In *2nd IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS'12). Feb. 23-25* (February 2012).

[23] Garman, S.C., and Garboczi, D.N. Structural basis of Fabry disease. *Molecular Genetics and Metabolism 77*, 1-2 (2002), 3 – 11.

[24] Gilis, D., and Rooman, M. Predicting protein stability changes upon mutation using database-dervied potentials: Solvent accessibility determines the importance of local versus non-local interactions along the sequence. *Journal of Molecular Biology 272*, 2 (1997), 276–290.

[25] Gohlke, H., and Radestock, S. Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. *Engineering Life Science 8* (2008), 507–522.

[26] Granzin, J., Puras-Lutzke, R., Landt, O., Grundert, H-P, Heinemann, U., Saenger, W., and Hahn, U. RNase T1 mutant Glu46Gln binds the inhibitors 2'GMP and 2'AMP at the 3' subsite. *Journal of Molecular Biology 225*, 2 (1992), 533–542.

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

[28] Holder, T. Supercell. `http://www.pymolwiki.org/index.php/Supercell`, August 2011.

[29] Hutschison, C.A., Philipps, S., Edgell, M.H., Gillham, S., Jagnke, P., and Smith, M. Mutagenesis at a specific position in a DNA sequence. *Journal of Biological Chemistry*, 18 (1978), 6551–6560.

[30] Hyun, J.K., Accurso, C., Hijnen, M., Schult, P., Pettikiriarachchi, A., Mitra, A.K., and Coulibaly, F. Membrane remodeling by the double-barrel scaffolding protein of poxvirus. *PLoS Pathogens 7*, 9 (September 2011).

[31] Ishima, R., Freedberg, D.I., Wang, Y.-X., Louis, J.M., and Torchia, D.A. Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. *Structure 7*, 9 (1999), 1047–1055.

[32] Jacobs, D.J., and Hendrickson, B. An algorithms for two-dimensional rigidity percolation: the pebble game. *Journal of Computational Physics 137* (1997), 346–365.

[33] Jacobs, D.J., Rader, A.J., Thorpe, M.F., and Kuhn, L.A. Protein flexibility predictions using graph theory. *Proteins 44* (2001), 150–165.

[34] Jacobs, D.J., and Thorpe, M.F. Generic rigidity percolation: the pebble game. *Physics Review Letters 75* (1995), 4051–4054.

[35] Jagodzinski, F., Clark, P., Liu, T., Grant, J., Monastra, S., and Streinu, I. Rigidity analysis of periodic crystal structures and protein biological assemblies. *Submitted, BMC BioInformatics* (2012).

[36] Jagodzinski, F., Hardy, J., and Streinu, I. Using rigidity analysis to probe mutation-induced structural changes in proteins. In *Workshop on Computational Structural Bioinformatics* (November 2011), IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM'11), pp. 432–437.

[37] Jagodzinski, F., Hardy, J., and Streinu, I. Using rigidity analysis to probe mutation-induced structural changes in proteins. *Journal of Bioinformatics and Computational Biology 10*, 3 (2012).

[38] Jagodzinski, F., and Streinu, I. Towards biophysical validation of constraint modeling for rigidity analysis of proteins. *BMC Bioinformatics*, Accepted.

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

[40] Ko, T.-P, Robinson, H., Gao, Y.-G, Cheng, C.-H.C., DeVries, A.L, and Wang, A.H.-J. The refined crystal structure of an eel pout type III antifreeze protein RD1 at 0.62-A resolution reveals structural microheterogeneity of protein and solvation. *Biophysical Journal 84* (2003), 1228–1237.

[41] Krebs, W.G., and Gerstein, M. The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Research 28*, 8 (2000), 1665–1675.

[42] Kumar, M.D., Bava, K.A., Gromiha, M.M., Prabakaran, P., Kitajima, K., Uedaira, H., and Sarai, A. Protherm and pronit: thermodynamic databases for proteins and proteinâĂŞnucleic acid interactions. *Nucleic Acids Research 34*, suppl 1 (2005), D204–D206.

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

[44] Lakowicz, J.R. *Principles of Fluorescence Spectroscopy*. Springer, 3rd edition, 2006.

[45] Laman, G. On graphs and rigidity of plane skeletal structures. *Journal of Engineering Mathematics 4* (1970), 331–340.

[46] Lee, A., and Streinu, I. Pebble game algorithms and sparse graphs. *Discrete Mathematics 308*, 8 (2008), 1425—1437.

[47] Lee, B., and Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology 55*, 3 (1971), 379–400.

[48] Lee, C., and Levitt, M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature 352* (1991), 448–451.

[49] Lou, X., Tu, X., Pan, G., Xu, C., Fan, R., Lu, W., Deng, W., Rao, P., Teng, M., and Niu, L. Purification, N-terminal sequencing, crystallization and preliminary structural determination of atratoxin-b, a short-chain alpha-neurotoxin from Naja atra venom. *Acta Crystallography 59*, 6 (2003), 1038–1042.

[50] Lovell, S.C., Word, J.M., Richardson, J.S., and Richardson, D.C. The penultimate rotamer library. *Proteins Structure Function and Genetics 40* (2000), 389–408.

[51] Matsumura, M., Becktel, W.J., and Matthews, B.W. Hydrophobic stabilization in t4 lysozyme determined directly by multiple substitutions of ile 3. *Nature 334* (1988), 406–410.

[52] Matsumura, Masazumi, Becktel, Wayne J., Levitt, Michael, and Matthews, Brian W. Stabilization of phage t4 lysozyme by engineered disulfide bonds. *Proceedings of the National Academy of Sciences 86*, 17 (1989), 6562–6566.

[53] Maxwell, J.C. On the calculation of the equilibrium and stiffness of frames. *Philosophical Magazine Series 4 27* (1864), 294–299.

[54] Mayo, S.L., Dahiyat, B.I., and Gordon, D.B. Automated design of the surface positions of protein helices. *Protein Science 6*, 6 (1997), 1333–1337.

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

[56] McCoy, R.H., Meyer, C.E., and Rose, W.C. Feeding experiments with mixtures of highly purified amino acids. viii. isolation and identification of a new essential amino acid. *Journal of Biological Chemistry 112* (1935), 283–302.

[57] Mcnaught, A. D., and Wilkinson, A. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*, xml on-line corrected version: http://goldbook.iupac.org (2006-) created by m. nic, j. jirat, b. kosata; updates compiled by a. jenkins ed. Blackwell Scientific Publications, Oxford, 1997.

[58] Michalet, X., Weiss, S., and Jager, M. Single-molecule fluorescence studies of protein folding and conformational dynamics. *Chemical Reviews 106*, 5 (2006), 1785–1813.

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

[60] Nachman, J., Miller, M., Gilliland, G.L., Carty, R., Pincus, M., and Wlodawer, A. Crystal structure of two covalent nucleoside derivatives of ribonuclease a. *Biochemistry 29*, 4 (1990), 928–937.

[61] Nicholson, H., Soderlind, E., Tronrud, D.E., and Matthews, B.W. Contributions of left-handed helical residues to the structure and stability of bacteriophage T4 lysozyme. *Journal of Molecular Biology 210*, 1 (1989), 181–193.

[62] Nicholson, L. K., Yamazaki, T., Torchia, D. A., Grzesiek, S., Bax, A., Stahl, S. J., Kaufman, J. D., Wingfield, P. T., Lam, P. Y. S., Jadhav, P. K., and Others. Flexibility and function in HIV-1 protease. *Nature Structural Biology 2*, 4 (1995), 274–280.

[63] Palmer, A.G, Grey, M.J., and Wang, C. Solution nmr spin relaxation methods for characterizing chemical exchange in high-molecular-weight systems. In *Nuclear Magnetic Resonance of Biological Macromolecules*, Thomas L. James, Ed., vol. 394 of *Methods in Enzymology*. Academic Press, 2005, pp. 430 – 465.

[64] Pande, Vijay S., Baker, Ian, Chapman, Jarrod, Elmer, Sidney P., Khaliq, Siraj, Larson, Stefan M., Rhee, Young Min, Shirts, Michael R., Snow, Christopher D., Sorin, Eric J., and Zagrovic, Bojan. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers 68*, 1 (2003), 91–109.

[65] Perryman, A.L., Lin, J.-H., and McCammon, J.A. HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: Possible contributions to drug resistance and a potential new target site for drugs. *Protein Science 13*, 4 (2004), 1108–1123.

[66] Prevost, M., Wodak, S.J., Tidor, B., and Karplus, M. Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96-Ala mutation in barnase. *Proceedings of the National Academy of Sciences, U.S.A. 88*, 23 (1991), 10880–10884.

[67] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[68] Rader, A. J., and Bahar, I. Folding core predictions from network models of proteins. *Polymer 45*, 2 (2004), 659–668.

[69] Rader, A.J., Anderson, G., Basak, I., Bahar, I., and Klein-Seetharaman, J. Identification of core amino acids stabilizing rhodopsin. *Proceedings of the National Academy 0f Science of the United States of America 101*, 19 (May 2004), 7246–7251.

[70] Rader, A.J., Hespenheide, B.M, Kuhn, L.A., and Thorpe, M.F. Protein unfolding: Rigidity lost. *Proceedings of National Academy of Sciences, U.S.A.* (2002), 3540–3545.

[71] Rupp, B. *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*, 1st ed. Garland Science, New York, 2009.

[72] Senechal, Marjorie. *Crystalline Symmetries, An Informal Mathematical Introduction.* Adam Hilger Publishing, 1990.

[73] Smith, M.J-de. STATSREF: Statistical analysis handbook. `http://www.statsref.com/`, 2010.

[74] Smith, S.O., Eilers, M., Song, D., Crocker, E., Ying, W., Groesbeek, M., and Aimoto, G. Metz M. Ziliox S. Implications of threonine hydrogen bonding in the glycophorin a transmembrane helix dimer. *Biophysics Journal 82* (2002), 2476–2486.

[75] Taketomi, H., Ueda, Y., and Gō, N. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *International Journal of Peptide and Protein Research 7*, 6 (1975), 445–459.

[76] Tay, T.-S. Rigidity of multigraphs I: linking rigid bodies in n-space. *Journal of Combinatorial Theory, Series B 36* (1984), 95–112.

[77] Teeter, M.M., and Hendrickson, W.A. Highly ordered crystals of the plant seed protein crambin. *Journal of Molecular Biology 127*, 2 (1979), 219–223.

[78] Teeter, M.M., Mazer, J.A., and L'Italien, J.J. Primary structure of the hydrophobic plant protein crambin. *Biochemistry 20*, 19 (1981), 5437–5443.

[79] Thorpe, M.F., Lei, M., Rader, A.J., and Kuhn, D.J. Jacobs L.A. Protein flexibility and dynamics using constraint theory. *Journal of Molecular Graphics and Modeling 19*, 1 (2001), 60–9.

[80] Thorpe, Michael F., Chubynsky, M. V., Hespenheide, B. M, Menor, Scott, Jacobs, Donald J., Kuhn, Leslie A., Zavodszky, Maria I., Lei, Ming, Rader, A. J., and Whiteley, Walter. *Flexibility in Biomolecules.* Current Topics in Physics. Imperial College Press, London, 2005, ch. 6, pp. 97–112. R.A. Barrio and K.K. Kaski, eds.

[81] Topham, C.M., Srinivasan, N., and Blundell, T. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitutions and propensity tables. *Protein Engineering 10*, 1 (1997), 7–21.

[82] Tsang, I.R., and Tsang, I.J. Cluster size diversity, percolation, and complex systems. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics 60* (1999), 2684–2698.

[83] Tsodikov, O.V., Record, M.T., and Sergeev, Y.V. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *Journal of Computational Chemistry 23*, 6 (2002), 600–609.

[84] Ueda, Y., Taketomi, H., and Gō, N. Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A. three-dimensional lattice model of lysozyme. *Biopolymers 17*, 6 (1978), 1531–1548.

[85] Weaver, L.H., and Mathews, B.W. Structure of bacteriophage T4 lysozyme refined at 1.7 A resolution. *Journal of Molecular Biology 193* (1987), 189–199.

[86] Wells, S.A., Jimenez-Roldan, J.E., and Romer, R.A. Comparative analysis of rigidity across protein families. *Physical Biology 6*, 4 (2009).

[87] Westbrook, John, Berman, Helen M., Feng, Zukang, Gilliland, Gary, Bhat, T. N., Weissig, Helge, Shindyalov, Ilya N., and Bourne, Philip E. The protein data bank. *Nucleic Acids Research 28* (2000), 235–242.

[88] Worth, C.L., Preissner, R., and Blundell, L. SDM-a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Research 39*, Web Server Issue (2011), W215–W222.

[89] Xu, J., Baase, W.A., Baldwin, E., and Matthews, B.W. The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Science 7*, 1 (1998), 158–177.

[90] Zhang, X.-J., Wozniak, J.A., and Matthews, B.W. Protein flexibility and adaptability seen in 25 crystal forms of t4 lysozyme. *Journal of Molecular Biology 250* (1995), 527–552.