A Novel Method for Characterization and Quantification of Flexibility and Mobility in Proteins

by

Elizabeth Eskow

B.S., University of Colorado Boulder, 1979

M.S., University of Colorado Boulder, 1985

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

College of Engineering and Applied Science Department of Computer Science

2014

This thesis entitled:
A Novel Method for Characterization and Quantification of Flexibility and Mobility in Proteins
written by Elizabeth Eskow
has been approved for the College of Engineering and Applied Science Department of Computer Science

Debra Goldberg

Dr. Deanne Sammond

Prof. Richard Byrd

Prof. Larry Hunter

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Eskow, Elizabeth (Ph.D., Computer Science)

A Novel Method for Characterization and Quantification of Flexibility and Mobility in Proteins

Thesis directed by Assistant Professor Debra Goldberg

Proteins in vivo are not completely rigid molecules, and mobilities within their structure play a key role in protein function. We discuss a novel method for measuring two distinct types of protein flexibility by comparing pairs of static protein structure coordinates. The measures focus on the mobility of a subset of atoms in the protein known as the backbone, and they quantify mobility or flexibility at the level of the amino acids (or residues), which are the basic constituents of proteins. We validate our measures against a subset of proteins from the protein-protein docking benchmark, and against a number of individual proteins known to have mobility or flexibility that is significant to their function. We also demonstrate the applicability of our methodology to several important biochemical topics including examples that apply to drug and enzyme design, and evaluation of computational protein structure prediction. We conclude with an analysis of protein structural and energetic terms showing which terms are associated with our flexibility measures, and may therefore be useful within the context of protein modeling algorithms to predict the locality of flexible regions.

Dedication

To Ben, the love and light of my life,

and

To the loving memory of my parents, who would have been so proud.

Acknowledgements

First and foremost, I would like to thank my advisor, Debra Goldberg, for her kind and steady guidance. She allowed me to work on research that was not in her area of expertise, and not only facilitated my collaborations with biochemists, but also advised and supported me every step of the way. I would especially like to thank Deanne Sammond; her questions on protein structure flexibility were the driving force behind all of my research, and her enthusiasm and guidance were a constant source of motivation. I would like to thank Larry Hunter for always challenging me to think beyond my current focus; his questions and expertise have been invaluable. I would like to give special thanks to Bobby Schnabel and Richard Byrd, who have stood by me and worked with me throughout my acadamic career. Clearly, I wouldn't be where I am today without them. It was Bobby Schnabel who planted the seed for my PhD pursuit a long time ago and perhaps someday I can forgive him. I would also like to thank Jessica Feld and Brian Faulker for being instrumental in making my year of teaching at Poudre High School such an incredible experience. I am indebted to Jackie DeBoard for all her assistance, and to the folks in CBP, especially Kathy Thomas and Anis Karimpour-Fard, for all their help and support. I would also like to thank all the other members of the Goldberg research group, and Suzanne Gallagher in particular, for always stepping up to help. I am forever grateful to all my family and friends, for their patience and kindness. And finally, I would like to thank my husband Ben for his undying love, and always encouraging me to "get tough or cry".

I also gratefully acknowledge financial support from NLM Training Grant 63002754 and NSF GK-12 Grant 0841423.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

1.1 Background and Motivation

Proteins are ubiquitous macromolecules present in all living organisms. The structure of proteins allows them to perform, independenntly or collaboratively, a plethora of biological functions including (but not limited to) signaling, regulation, structural support, transportation and catalysis. There are four distinct levels of protein structure and each structural level encompasses the previous ones and is increasingly more complex. The primary level is composed of the sequence of amino acids (often referred to as residues) and there are twenty unique amino acids in nature that comprise the sequences of proteins. Secondary structures include local and non-local formations such as α-helices and β-sheets. Tertiary structure is the full three-dimensional shape (or conformation) and is represented by cartesian or internal (bond and angle) atomic coordinates. The quaternary structure is comprised of more than one protein molecule in a bound complex. We are primarily interested in changes to protein tertiary structure, but all structural levels are involved in this research.

The thermodynamic hypothesis of protein folding postulates that the native (tertiary) structure of a protein is the global minimum of free energy, determined only by its (primary) amino acid sequence. This hypothesis is based on Christian Afinsen's Nobel Prize winning discovery that a denatured protein can spontaneously self-assemble into its native, biologically-active conformation [2]. It follows that under the correct environmental conditions, a protein's native state is a thermostable, correctly-folded tertiary structure. The problem of finding the native state of a protein from its amino acid sequence is the excedingly challenging protein folding problem. This research

does not involve the protein folding problem explicitly, but we address a related area of predicting protein structural changes.

There are known exceptions to the thermodynamic hypothesis. For example, the inhibitor protein serpin has more than one native state. This protein exists as an ensemble of conformers; the native structure is believed to be a metastable, or long-lived intermediate, and its folding pathway to thermostability sometimes results in an inactive but thermostable conformation [107]. This is an example of a protein that adopts multiple conformations, and its lowest energy state is not always the most biologically active. Other exceptions include prions, with a thermostable but misfolded state that has been implicated in a number of diseases. In fact, there is a great deal of evidence that many proteins adopt multiple conformations, especially in induced-fit binding to other proteins [22]. These are not exceptions to the thermodynamic hypothesis because the proteins are influenced by environmental conditions, but nonetheless, the concept of proteins adopting multiple different shapes is extremely important. The study of the underlying differences in conformations of the same protein is the essence of this work. We further motivate this topic by explaining its significance to protein function and the applied fields of protein engineering and design.

Proteins are inherently flexible, and dynamic changes to their conformational shapes are fundamentally important for biological function. According to [103], "proteins are flexible and rapidly fluctuating molecules whose structural mobilities have considerable functional significance". Protein dynamics can occur on a wide range of temporal and spatial time scales. This range encompasses the smallest atomic fluctuations such as bond vibrations, on the order of femtoseconds to 10^{-11} seconds with spatial displacements of 0.01 to 1 Angstroms ($\mathring{A}$), up to large domain or subunit motions that can take seconds with displacements up to 10 $\mathring{A}$ [103]. Thus, understanding the full function of a protein often requires not only a static, experimentally determined (native) structure, but also information about the protein's conformational changes. This crucial information about proteins can be obtained through a number of experimental and computational techniques which are described later in this chapter.

The original and ultimate goal for our research is to provide useful predictions about where

proteins are flexible in order to improve computational design algorithms for protein engineering. Protein engineering is the process of creating (de novo) proteins or modifiying existing proteins to build new and useful ones, and stems from the field of protein design. Protein design encompasses two distinct but interrelated components; the design of biological activity and the design of protein structure [85]. The goal of rational protein design, one of the main strategies used for protein engineering, is to predict the amino acid sequences that will fold to a specific protein structure. This is inversely related to the protein folding problem, but sometimes both problems (protein design and protein folding) are addressed simultaneously. An example is the exciting design of a novel protein from a previously unknown fold, accomplished by optimizing both sequence design and structure prediction [52]. Many protein design and engineering projects have already included structure prediction by allowing for protein backbone flexibililty in their models and algorithms. Our ultimate goal is to improve upon the efficiency and accuracy of these methods by providing more information on the locality and extent of existing or modified protein backbone flexibility.

We review some of the recent breakthroughs in protein engineering, especially the field of rational protein design, that have lead to increasing difficulty and complexity in the types of challenges that have been successfully addressed. Many of these breakthroughs have been achieved by computational algorithms that address the dynamic nature of proteins, using a variety of techniques. These techniques include the incorporatation of simulations of molecular dynamics to generate a variety of starting configurations, or the introduction of variations directly into the structure of the backbone during or prior to the design of a protein sequence. An exciting, improved design of a 300-fold or greater increase in specificity for the protein-protein interface of an immunity protein complex was accomplished using an ensemble of starting conformations with variations in the backbone [44]. An elegant solution to the problem of designing a protein-binding peptide (the GoLoco peptide - G-protein interface) was accomplished using a strategy of sequence design optimization combined with backbone modeling [87]. In a final example, Amy Keating's lab introduced backbone flexiblity into the design of an α-helical ligand to bind the antiapoptotic protein $Bcl - x_L$ using normal mode analysis and successfully designed several new peptides that bind the native

protein [33]. Each of these efforts highlight the benefit of including dynamic information in protein engineering efforts.

The combination of protein design and structure prediction has been a successful strategy in a number of protein engineering acccomplishments, including [87, 52], but requires an enormous cost in computational complexitiy beyond that of just protein design. If we knew where the dynamic regions existed, or which specific residues were flexible in the context of design, we could incorporate backbone flexibility into the design in just those regions where it was needed. This is the original motivation for our study. In later sections, we describe the existing software and explain why it doesn't solve this problem exactly. The main issue is that existing software is useful for detecting a specific *type* of protein flexibility, and apriori, we don't know what type of flexiblity we might need to look for.

Our ultimate goal is to predict flexibility within the execution of a protein design algorithm. Machine learning classification was proposed as a method to predict each residue of the protein as either flexible or rigid. Input attributes to the model would naturally include how tightly the residues were packed, and various computationally available energy terms could be used to determine the protein backbone's ability to change shape or move around, such as atomic attraction or repulsion to other residues in the protein. However, a simple classifier is insufficient because variations in flexibility are more accurately expressed as a continuum of values. Although the ranking of the relative flexibility of the residues can be predicted using machine learning regression, the remaining difficulty is in finding a source of input for the actual values of flexibility that we were trying to predict. In other words, without existing data on residue level flexibility, there is no way to implement a continuous machine learning model of flexibility. Our new method arose in response to this dilemma of lacking accurate experimental data on flexiblity at the residue level.

Why is it difficult to find experimental data that expresses residue level flexibility? Nuclear Magnetic Resonance (NMR) order parameters, which express the variation in NMR coordinate files, are an obvious choice for experimental data. Unfortunately NMR data is limited in terms of the size of the protein and resolution, or defined structural detail. In a comparison of structures

from X-ray crystallography versus NMR structures as templates for computational design, Amy Keating and colleagues found that overall, the X-ray structures were better templates for use with the commonly used Rosetta modelling suite [89, 61]. Another possibility is the use of Root Mean Square Fluctuation (RMSF) outputs from Molecular Dynamics (MD) simulations, but in that case we would be predicting from a computational simulation not validated by experiment. We discovered that there is not an existing method for analyzing flexibility from X-ray data on a per-residue basis, and yet this is precisely what we need. Futhermore, a new method to analyze residue-level flexibility from X-ray crystal coordinates of different protein conformations will be a useful tool for other researchers.

Our algorithm provides an analysis of experimental data that can be used to infer the relative per-residue amplitude of motion without the computational expense of computing energetics. We focus on changes to the backbone resulting in movements within or between secondary structure elements or larger domains while ignoring the very fast timescales corresponding to bond vibrations evident as side chain motions, or hydrogen bond formation. Thus, the method is not intended to replace other extremely valuable computational methodologies that provide either detailed simulation data of movements (molecular dynamics) or an analysis of the ordered motions of an ensemble of configurations such as normal mode analysis. Instead, by utilizing different measurements, we can readily assess different types of movements.

1.2 Problem Description

We investigate the problem of identifying and quantifying the backbone flexiblity of a protein at the residue level, limited to the comparison of pairs of X-ray crystal structures with differing conformations (coordinates) for the same protein sequence. We also explore which protein structure and sequence attributes are associated with backbone flexibility.

1.3 Related Methods

1.3.1 Experimental methods used to detect protein flexibility

Numerous experimental techniques are available to investigate protein dynamics and flexibility. These techniques include:

- Time-resolved x-ray methods attempt to capture protein sructural changes in real time, but this methodology has many practical limitations which prevent widespread usage [106].
- Fluorescence techniques such as fluorescence resonance energy transfer (FRET) [47], and fluorescence correlation spectroscopy (FCS) can be applied small proteins [73].
- Hydrogen-deuterium exchange shows experimental evidence for enhanced flexibility in a mesophilic protein versus its thermophilic homolog [77].
- Nuclear Magnetic Resonance (NMR) is used to compare flexibility in a mesophilic enzyme versus a homologous thermophilic enzyme [38].

1.3.2 Computational methods to simulate or predict protein flexibility

- Molecular Dynamics (MD) simulations have been combined with biophysical data to gain detailed mechanistic information, such as a more complete understanding of molecular recognition dynamics in binding [38, 39, 57].
- MD alone can be used to probe dynamics, although significant computational resources are required and there are limits to the size of macromolecule that can be considered [25].
- MD analysis has been used to predict NMR relaxation data [16] and classify proteins according to mobility patterns [37].
- Normal mode analysis (NMA) computes the low frequency normal modes that are associated with much of the movement of a protein, and the ElNémo Web Server identifies normal modes for movement analysis [95].

- Elastic network models are harmonic models based on a highly simplified energy function, and have been shown to be useful in studying protein conformational changes [67].

- Hinge prediction:

  (1) HingeProt uses elastic network models to predict hinges in a single protein structure [28].

  (2) StoneHinge uses a consensus of two complementary analyses of noncovalent bond networks to predict hinges between domains of a single protein structure [48].

  (3) DynDom uses two conformations to determine domains, hinge axes and hinge bending residues [36].

  (4) FIRST analyzes rigidity and flexibility in network graphs [100].

- TLS motion determination (TLSMD) analyzes a protein crystal structure for evidence of flexibility, such as local or inter-domain motions [78].

1.4 Thesis Overview

We define some of the basic biochemical terms commonly used in the thesis at the end of this chapter in §1.5. Chapter 2 describes a framework for characterizing protein flexiblity at the residue level, and a method for scoring flexibility within the context of the new framework. Background for each of the individual measurements is also provided. Chapter 3 validates the method by examining the flexiblity scores of a collection of protein-protein docking benchmark data. We also validate a number of proteins against the flexiblity described in the structure literature. Chapter 4 gives several applications of the new method, and Chapter 5 examines correlations of our flexibility scores with structural, sequence and energetic attributes. We conclude with a summary of contributations and future work in Chapter 6.

1.5 Definitions

- **Amino Acid** The basic constituent (monomer) of a protein (polypeptide), containing a carboxyl and amino group. The R group, or side chain, differs for each of the 20 naturally occurring amino acids.

- **Backbone** All atoms in the amino acids of the protein except for the R group atoms, or side chains. This includes the amino group, the carboxyl group, and the central carbon, or C_α atom.

- **Dihedral Angle** There are three different dihedral angles in the protein backbone, but in this thesis only two are considered: the phi (ϕ) angle of rotation between the nitrogen and C_α atoms in the backbone, and the psi (ψ) angle of rotation between the C_α and the carbonyl carbon atoms in the backbone.

- **Nuclear Magnetic Resonance (NMR) Spectroscopy** An experimental technique that reveals the three-dimensional structure of a protein in solution.

- **Residue** Another name for Amino Acid, used interchangably throughout the thesis.

- **Secondary structure** Regions of repetitive coiling (helices) or folding of the protein backbone due to hydrogen bonding between constituents of the backbone.

- **X-ray crystallography** The first method developed to determine the three dimensional structure of a protein (or protein complex) in atomic detail. This experimental technique still provides the clearest visualization of protein structures currently available [8].

Chapter 2

Movers and Shapers: Characterizing and Quantifying Backbone Flexibilty at the Residue Level

2.1 Introduction

The primary goal of our new method is to provide a framework to characterize and quantify the backbone flexibility found by comparing different conformations (or shapes) of experimental protein structures. The enormous variety of protein conformational changes represented by comparing proteins crystalized under different conditions makes this a challenging endeavor. Specifying and quantifying *how* protein backbones change in different circumstances and *where* they are more or less flexible may lead to insights into protein function [38]. In the protein structure literature, we've found that many of the observations concerning conformational flexibility are gleaned by using analytical techniques which are similar to the measures we combine to calculate our flexibility scoring functions. Our method is both novel in the way we characterize and quantify residue level flexibility in a single framework, and general in its applicability to many different types of conformational changes.

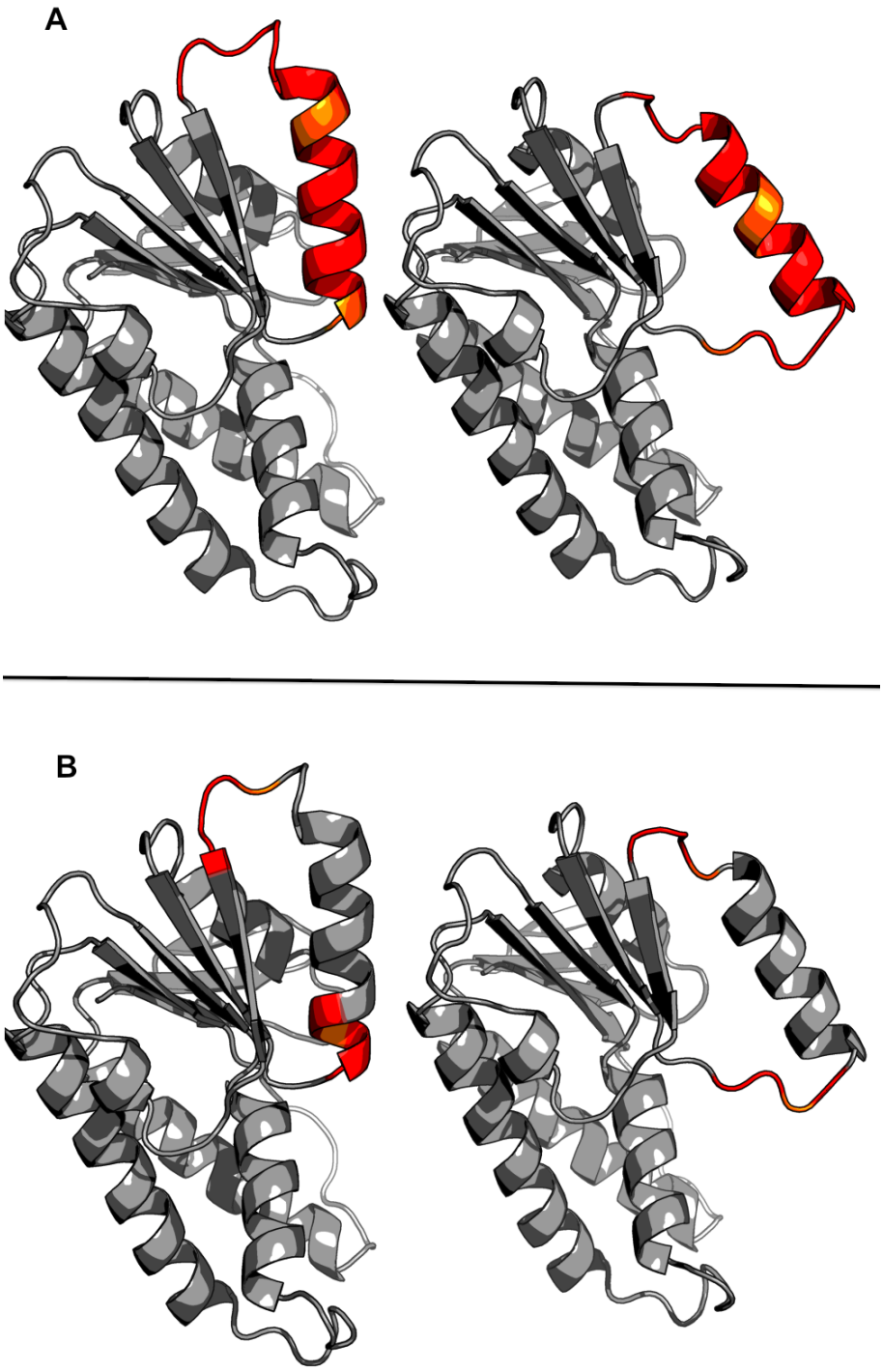
We quantify residue level flexibility by analyzing both local and global differences in protein structures. More specifically, residues within regions that appear mobile with respect to the remainder of the protein are characterized as *movers* exhibiting mobility, whereas those residues within locally deformed areas of the protein are characterized as *shapers* exhibiting shape pliability. Some degree of both characterizations can apply to the same residues. For example, when a domain moves with respect to the rest of the protein but also exhibits smaller scale deformations

within the domain then some residues may have relatively high levels of both mobility and shape pliability. This can be seen clearly in the example of the characterizations of mobility and shape pliability shown in Figure 2.1 A and B respectively. In part A, the high level of mobility of the top right helix is demonstrated by the way its position changes with respect to the remainder of the protein. The shape pliability in this case are hinges, or deformations to the shape of the backbone enabling the mobility highlighted in A to occur, and the two regions of high shape pliability are highlighted in Figure 2.1 B for the same conformations as shown in A. In part B the helix is not highlighted because it doesn't change shape when it moves away from the rest of the protein. All figures containing protein structures, here and elsewhere in the thesis, were created using PyMol [90], unless otherwise noted.

According to [40], the movement of a protein backbone segment in an ordered structure can be classified as either an internal or external motion. These motions are fundamentally different; an internal motion refers to a deformation of the segment itself and an external motion refers to translational and rotational motion of the rigid segment. Because we are comparing static structures and analyzing changes to the backbone without regard to underlying motions, we use our own terminology to describe these changes based on differences in observable measurements from experimental structures. We note that while shape pliability and mobility describe different aspects of the measured changes, some regions may exhibit high levels of both. Therefore, the measures for mobility and shape pliability are different but not exclusively so.

Using X-ray diffraction experimental data, the dynamic nature of a protein can be revealed by comparing conformational differences between independent crystal structures of the same protein [56]. We currently apply our method exclusively to this type of data. X-ray crystal structures of different conformations show less variability than NMR structures, but provide more detailed structural information [111, 101] . Figure 2.2 shows independent crystal structures for the protein calmodulin; a classic example of a protein changing conformational shape upon binding to a peptide [17]. Calmodulin plays a central role in intracellular calcium signaling such that after calcium is sensed by the structure's lobes, the extreme flexibility of the central helix enhances the likelihood

Figure 2.1: Movers and Shapers: A. Movers: Mobility of the top right helix with respect to the remainder of the protein in 2 different conformations of the protein CobU. Highly mobile residues are highlighted in red and orange, with the remainder of the protein colored gray for contrast. B. Shapers: Residues with extremely high shape pliability scores are highlighted red and orange in the same 2 CobU conformations. The shape changes in B enable the large-scale movement in A.



that a target peptide bound to one lobe will bind to the second lobe [94].

B factors from X-ray diffraction data or order parameters from NMR structures are also used to obtain protein structure flexibility information. B factors (also known as temperature or Debye-Waller factors) are a measure of atomic displacement due to both static disorder (where substates present in solution are trapped in the crystal and may induce modeling errors) as well as the actual atomic motion, or dynamic disorder [38]. Therefore, using B factors to represent only the amplitude of atomic fluctuations can be misleading. Our goal is to define a continuous measure of backbone flexibility similar to the B factor (or NMR order parameter), but representing only the amplitude of the atomic fluctuations found by comparing x-ray crystal structures of different conformations of the same protein.

2.2 Methods

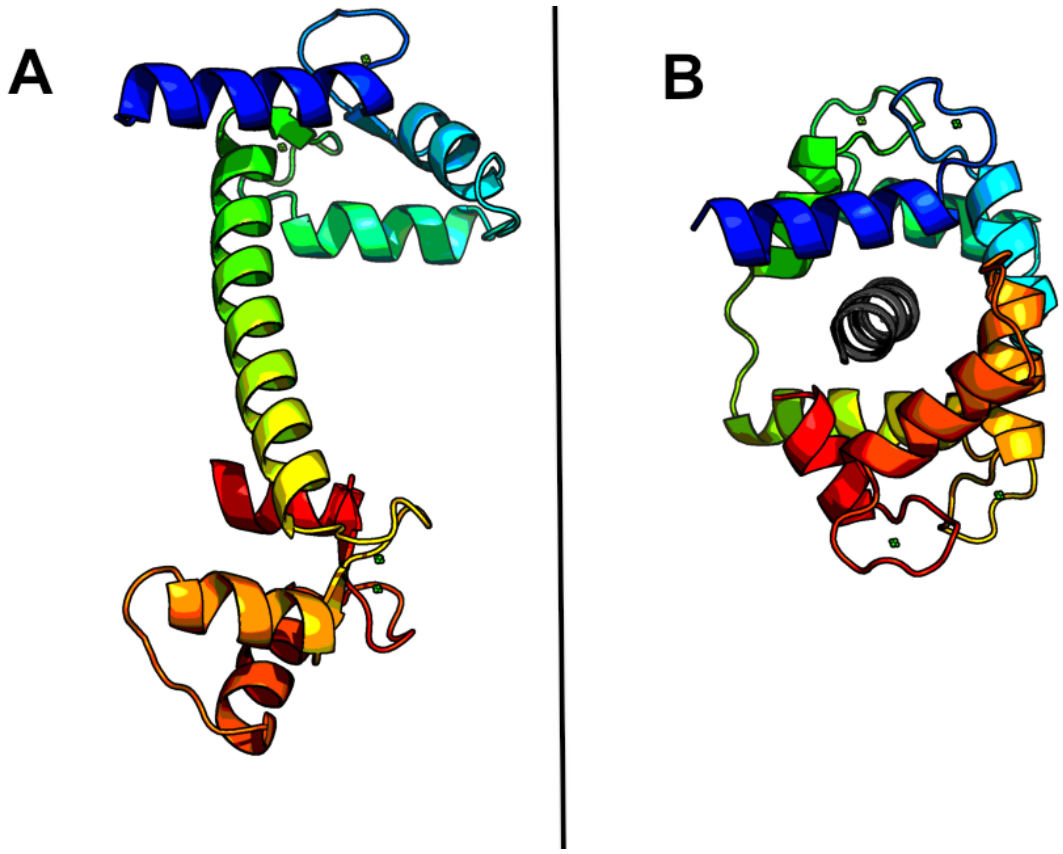
2.2.1 Overview of Measurement Calculations

Three types of measurements are used in computing per residue mobility and shape pliability scores. Below is a brief overview of these measurements, and the subsequent sections we provide more details.

(1) **C_α Distances from our Superpositioning by Secondary Structure elements (S^3) algorithm.** The minimum distance obtained between C_α atoms for each corresponding pair of residues is computed by superpositioning the pair of the conformations as a whole, and recursively superpositioning down to the level of individual secondary structure elements, as necessary, to improve the structural alignment. This calculation does not necessarily find a continuous alignment; instead it optimizes the alignment of individual elements of secondary structure, such as α-helices and β-strands. This measure is used in computing mobility or shape pliability scores, depending on the relative magnitude of the per residue distance measured.

(2) **Intramolecular distance calculations.** Intramolecular distances are computed over

Figure 2.2: X-ray crystal structures of the protein calmodulin bound to calcium: A. Calmodulin without a peptide is dumb-bell shaped, with similar lobes (or hands) connected by a central α-helix and B. Calmodulin when bound to a peptide (central unattached gray helix) differs significantly from the unbound structure in the central α-helix, while the lobes differ only slightly (1CDL) [17, 68]. Each lobe contains 3 α-helices and 2 loops binding calcium (shown as small green dots).The two structures are colored using a rainbow spectrum with the same colors for corresponding helices and loops to highlight the similarities.



the C_α atoms for each conformation, and the corresponding per residue differences of the distances are calculated. The percentage of differences pertaining to each residue and above a certain threshold is called the Difference Distance Matrix Percentage (DDMP). DDMP is used in computing the per residue mobility scores.

(3) **Dihedral angle differences.** Comparisons of dihedral angles measuring differences in the phi (ϕ) and psi (ψ) backbone dihedral angles for corresponding residues are computed, and the measurement for the dihedral angle difference is the maximum of the per residue ϕ or ψ angle difference. The computation of backbone dihedral angle changes can highlight important shape pliabilities that may not be visible using superpositioning alone, hence this measure is extremely important in the computation of shape pliability.

2.2.2 Superpositioning

2.2.2.1 Background

Many protein design and structure prediction articles use superpositioning to reveal the functional movements of proteins in different environments or with different binding partners. An example of this is the recent study of estrogen receptors, which are the main targets of estrogens and biomarkers for certain types of breast cancer [63]. The structures of the ligand binding domain of an estrogen receptor (ER) in complex with 3 different ligands were superimposed to display the differences in the ER movements responding to the different ligands. The information gleaned by this superpositioning may aid the design of novel ligands useful in treating or diagnosing diseases associated with estrogen receptors.

Superpositioning two (or more) structural conformations to minimize rotational and translational differences, and computing the resulting root-mean-square deviation (RMSD) measured by comparing the locations of the corresponding C_α (or all) atoms is a commonly used approach to evaluating structural differences. Two different types of superpositioning methods include rigid superpositioning algorithms such as [98, 24, 62, 66, 75] that do not allow the structures to change

shape, and flexible superpositioning algorithms [110, 93] that change the shape of one of the structures to find a superposition with lower RMSD. Structural similarity results from superpositioning are used for purposes of structure prediction, fold classification or database searching for structural homology; information about the proteins as a whole. In contrast, the purpose of our analysis is to compare the individual corresponding residues of conformational pairs of proteins to determine protein flexibility at the residue level. We have implemented our own superpositioning algorithm called "Superpositioning by Secondary Structure elements" (S^3) which gives a more accurate per-residue level comparison than existing superpositioning algorithms.

We experimented with a number of different superpositioning algorithms before implementing our own. The Theseus superposition algorithm [97, 98] uses maximum likelihood instead of least squares to find the optimal translation and rotation. For proteins that are somewhat rigid, the Theseus algorithm tends to find more accurate solutions by differentially weighting structural regions and correcting for proximal atomic correlations. For molecules that undergo large conformational changes, however, a superposition algorithm such as FATCAT [110] that allows modifications to one of the protein structures finds lower RMSD solutions. FATCAT superpositioning takes into account the flexibility of the molecule by adding twists between aligned fragment pairs (AFPs), with the simultaneous goals of optimizing the alignment and minimizing the number of twists [110].

For many cases, the FATCAT superpositioning results demonstrate that unaligned regions of the protein match what is known to be flexible. However, in some cases involving larger motions, the unaligned portions don't always match precisely what is known to be flexible. An example of a FATCAT alignment of the receptor protein ($uPAR$) from a protein-protein complex is shown to illustrate the problem we found when attempting to use this (incredibly valuable) tool to find flexibility. We show the alignment between bound and unbound comformations for one domain (D^I) of the 3-domain receptor, to focus on the distances to the C_α atoms obtained by the FATCAT alignment. In Figure 2.3 the FATCAT alignment is shown with the 2 aligned conformations colored in green and magenta. In the reference for the $uPAR$ structure, two hairpin turns in domain D^I are described as extremely flexible, with the labelled residues inside the hairpins, Glu34 and Leu61,

having RMSD's of 6.4$\mathring{A}$ and 5.1$\mathring{A}$ respectively [4]. Figure 2.3 demonstrates that the FATCAT alignment results for the flexible hairpin turns are far apart in the superposition, as expected. But most of the other turns in the FATCAT alignment are also far apart. Herein lies the problem of trying to discriminate highly flexible from less flexible residues without falsely specifying that all turns in the protein are highly flexible.

The flexible hairpin turns described in [4] are located at residues 34-36 and 59-62, and these residues have green boxes around them in the measurements reported in Figures 2.4 and 2.5. The column under the heading "Fatcat" reports the C^{α} distances measured from the FATCAT alignment displayed in 2.3. There are a number of regions with reported distances $> 2\mathring{A}$, including residues 1-24 (excluding residue 8), 32-36, 43-50, 59-62 and 72-77. In fact, the residues in region 1-24 have higher C^{α} distances than those in 59-62, although 59-62 is the region described as the flexible hairpin in [4].

Occasionally, a display of superimposed structures is divided up into domains to clearly demonstrate which parts of the protein have changed, because it is not feasible to display a single superimposed image showing disparate changes. For example, the receptor protein $uPAR$ was superimposed by its three domains separately in the discussion of structural changes from unbound to bound conformations [4]. As an additional example, the crystal structure conformations for the human S100A6 calcium sensor protein (known to be overexpressed in certain tumor cells) are shown in [76] and the first 2 helices are superimposed in a separate figure from the remaining 2 helices. The figures from [76] motivated the development of the Superpostion by Secondary Structure elements (S^3) algorithm, because existing superpositioning algorithms didn't explicitly provide the same information about flexibility as the manually created figures and measurements of the distinct protein regions described in the structure paper. The analysis of mobility and shape pliability for the residues of the calcium sensor protein are discussed later in §3.3.2.

After an alignment between C_{α} atoms in the respective structures has been determined, a superpositioning algorithm determines the translation and rotation that best relates the two sets of C_{α} atoms [32]. Within the context of this thesis, the sequence alignment is trivial because we

Figure 2.3: FATCAT [110] alignment of the bound versus unbound conformations of a single domain, isolated from the alignment of the full three-domain receptor protein (**uPAR**) of a protein-protein complex. The gold colored hairpin turns are highlighted in the figure because they are highly flexible (and contain the 2 identified residues), with relatively large RMSD's for the C^{α} atoms of residues Leu61 and Glu34 specifically noted in [4].

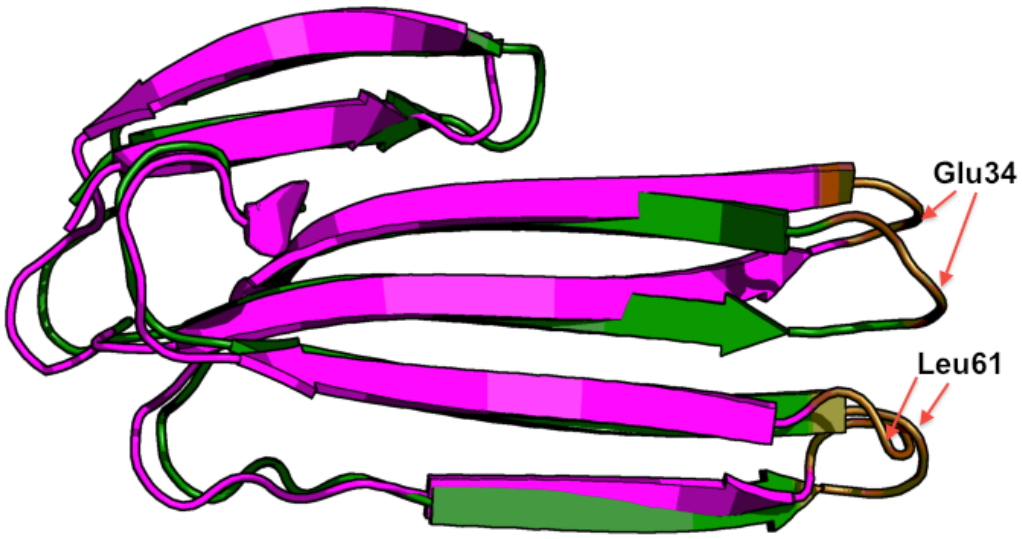


Figure 2.4: Measurements for comparing bound and unbound protein $uPAR$ domain D^I. Residue comparisons within the green rectangle are highly flexible, and located within a hairpin turn [4]. See Figure 2.5 for column heading definitions.

PDB1	RESID	PDB2	RESID	D_phi	D_psi	DDMP	S^3	Fatcat	Shape	Mobil	S^2
2i9b E	LEU1	1ywh A	LEU1	0	-0.7	0.53	0.42	4.21	13.9	52.71	- -
2i9b E	ARG2	1ywh A	ARG2	2.2	14.7	0.45	0.41	4.09	14.72	45.35	E E
2i9b E	CYS3	1ywh A	CYS3	1.9	17.6	0.46	0.4	3.41	17.57	46.12	E E
2i9b E	MET4	1ywh A	MET4	-33.8	-15.5	0.44	0.32	2.89	33.84	43.8	E E
2i9b E	GLN5	1ywh A	GLN5	9.5	-7.6	0.45	0.33	2.95	10.9	44.96	E E
2i9b E	CYS6	1ywh A	CYS6	9.6	-6.8	0.43	0.08	2.25	9.62	43.02	E E
2i9b E	LYS7	1ywh A	LYS7	-3.5	-6.9	0.45	0.38	2.3	12.77	45.35	- -
2i9b E	THR8	1ywh A	THR8	18.6	-8.9	0.42	0.56	1.93	18.7	42.25	T T
2i9b E	ASN9	1ywh A	ASN9	-33.2	16.6	0.45	0.47	2.63	33.23	44.96	T T
2i9b E	GLY10	1ywh A	GLY10	32.5	-14.1	0.42	0.32	2.04	32.49	41.86	S S
2i9b E	ASP11	1ywh A	ASP11	-7.4	-15.6	0.43	0.47	2.3	15.62	42.64	- -
2i9b E	CYS12	1ywh A	CYS12	35.4	42.8	0.43	0.24	2.43	42.85	42.64	E E
2i9b E	ARG13	1ywh A	ARG13	-43	-8.7	0.44	0.28	2.91	43.03	44.19	E E
2i9b E	VAL14	1ywh A	VAL14	1.9	-5.4	0.45	0.2	3.45	6.51	45.35	E E
2i9b E	GLU15	1ywh A	GLU15	6.5	-6.9	0.47	0.46	3.89	15.24	46.51	E E
2i9b E	GLU16	1ywh A	GLU16	-1.3	17.1	0.46	0.3	4.17	17.09	46.12	E E
2i9b E	CYS17	1ywh A	CYS17	-19.7	-3.9	0.48	0.17	4.06	19.7	48.45	- -
2i9b E	ALA18	1ywh A	ALA18	7.8	2.5	0.52	0.27	4.4	8.85	51.55	- -
2i9b E	LEU19	1ywh A	LEU19	-7.1	15.4	0.5	0.38	4.38	15.43	50.39	T T
2i9b E	GLY20	1ywh A	GLY20	-10.9	32	0.52	0.27	4.18	32.01	51.94	T T
2i9b E	GLN21	1ywh A	GLN21	-29.5	-1.4	0.46	0.35	3.35	29.48	45.74	- -
2i9b E	ASP22	1ywh A	ASP22	20	-23.1	0.47	0.11	3.33	23.15	47.29	- -
2i9b E	LEU23	1ywh A	LEU23	-7.3	-14.8	0.45	0.32	3.11	14.78	45.35	E E
2i9b E	CYS24	1ywh A	CYS24	2.8	17.3	0.47	0.37	3.12	17.27	47.29	E E
2i9b E	ARG25	1ywh A	ARG25	-31.9	6.7	0.44	0.42	1.89	31.9	44.19	E E
2i9b E	THR26	1ywh A	THR26	-1	-11.5	0.45	0.12	1.66	11.53	44.57	E E
2i9b E	THR27	1ywh A	THR27	5.6	14.7	0.43	0.42	1.56	14.74	43.41	E E
2i9b E	ILE28	1ywh A	ILE28	-20.2	6.8	0.44	0.87	1.44	29.03	44.19	E E
2i9b E	VAL29	1ywh A	VAL29	-7.6	-2.4	0.41	0.88	1.13	29.42	41.47	E E
2i9b E	ARG30	1ywh A	ARG30	-3.4	-24.1	0.42	0.83	0.81	27.74	42.25	E E
2i9b E	LEU31	1ywh A	LEU31	59.9	-63.5	0.34	0.98	0.81	62.34	33.72	E E
2i9b E	TRP32	1ywh A	TRP32	8.2	9.7	0.52	1.11	4.05	36.86	52.33	E E
2i9b E	GLU33	1ywh A	GLU33	-21.1	-4.8	0.69	1.34	7	44.52	69.38	E E
2i9b E	GLU34	1ywh A	GLU34	-26.5	-146.7	0.72	2	9.21	100	71.71	S T
2i9b E	GLY35	1ywh A	GLY35	172.7	-112.3	0.61	2.24	8.26	100	74.83	S T
2i9b E	GLU36	1ywh A	GLU36	98.5	2.2	0.42	0.13	4.44	85.7	42.25	_ E
2i9b E	GLU37	1ywh A	GLU37	-60	10.9	0.36	0.59	1.72	59.99	36.05	E E
2i9b E	LEU38	1ywh A	LEU38	7.7	7.6	0.34	0.9	0.93	29.98	33.72	E E
2i9b E	GLU39	1ywh A	GLU39	-8.9	-5.6	0.38	0.33	0.98	11	37.98	E E
2i9b E	LEU40	1ywh A	LEU40	12.2	8.7	0.42	0.57	1.21	19.15	41.86	E E
2i9b E	VAL41	1ywh A	VAL41	-10.8	-5.8	0.41	0.16	1.31	10.82	41.09	E E
2i9b E	GLU42	1ywh A	GLU42	4.5	1.5	0.42	0.32	1.84	10.52	42.25	E E
2i9b E	LYS43	1ywh A	LYS43	-2.8	1.7	0.44	0.38	2.31	12.78	43.8	E E
2i9b E	SER44	1ywh A	SER44	0.2	-8.4	0.47	0.44	2.99	14.55	46.51	E E
2i9b E	CYS45	1ywh A	CYS45	4.6	50.6	0.47	0.2	2.92	50.59	46.51	E E
2i9b E	THR46	1ywh A	THR46	-52.2	-3.2	0.48	0.46	2.91	52.22	48.45	E E
2i9b E	HIS47	1ywh A	HIS47	9.9	6.8	0.46	0.49	2.75	16.25	46.12	- -
2i9b E	SER48	1ywh A	SER48	2.7	-11.1	0.52	0.52	2.81	17.26	52.33	T T
2i9b E	GLU49	1ywh A	GLU49	12.4	1.8	0.6	1.31	2.96	43.57	59.69	T T
2i9b E	LYS50	1ywh A	LYS50	8.6	-34.1	0.5	1.01	2.46	34.06	50.39	- -
2i9b E	THR51	1ywh A	THR51	32.8	11.4	0.45	0.7	1.74	32.81	44.96	- -
2i9b E	ASN52	1ywh A	ASN52	-23.6	1.9	0.42	0.48	0.99	23.63	41.86	- -

Figure 2.5: Continuation of Measurements for $uPAR$ domain D^I, the green rectangle highlights the second extrememly flexible hairpin turn described in [4]. Column headings: The 2 PDB files (PDB1 and PDB2) with residues (RESID), differences in the ϕ (D_phi) and ψ dihedral angles (D_psi), the Distance Distance Matrix Percentage (DDMP), S^3 C_α distance (S^3) , the FATCAT C_α distance (Fatcat), shape pliability (Shape) and mobility (Mobil) scores, and the secondary structure for each residue (S^2) , see §5.2 for secondary structure definitions.

PDB1	RESID	PDB2	RESID	D_phi	D_psi	DDMP	S^3	Fatcat	Shape	Mobil	S^2
2i9b	E ARG53	1ywh	A ARG53	-17.1	5.9	0.4	0.42	0.36	17.07	39.92	E E
2i9b	E THR54	1ywh	A THR54	0.9	1.8	0.37	0.35	0.29	11.83	37.21	E E
2i9b	E LEU55	1ywh	A LEU55	-3.5	-10.7	0.37	0.49	0.89	16.47	36.82	E E
2i9b	E SER56	1ywh	A SER56	16.1	-7.8	0.32	0.17	0.65	16.07	31.78	E E
2i9b	E TYR57	1ywh	A TYR57	10	-9.3	0.36	0.6	0.77	20.02	36.05	E E
2i9b	E ARG58	1ywh	A ARG58	-36.5	-18.3	0.36	0.56	1.02	36.49	36.43	E E
2i9b	E THR59	1ywh	A THR59	7.7	101.8	0.56	0.47	2.82	87.88	56.2	
2i9b	E GLY60	1ywh	A GLY60	-150	77.4	0.63	1.33	3.6	100	63.1	S S
2i9b	E LEU61	1ywh	A LEU61	-23.3	19.9	0.59	2.22	2.84	23.27	74.15	S S
2i9b	E LYS62	1ywh	A LYS62	-65.4	-1.4	0.58	1.18	3.24	63.57	57.75	E S
2i9b	E ILE63	1ywh	A ILE63	-4.6	-20.3	0.41	0.35	0.9	20.28	41.47	E E
2i9b	E THR64	1ywh	A THR64	-24.8	-4.9	0.43	1.03	0.79	34.33	42.64	E E
2i9b	E SER65	1ywh	A SER65	14.1	-34.4	0.39	0.81	0.53	34.37	38.76	E E
2i9b	E LEU66	1ywh	A LEU66	38.7	24.3	0.47	0.35	1.29	38.69	46.51	E E
2i9b	E THR67	1ywh	A THR67	-30	1.1	0.46	0.19	1.26	29.97	46.12	E E
2i9b	E GLU68	1ywh	A GLU68	-16.2	10.5	0.42	0.32	1.03	16.21	41.86	E E
2i9b	E VAL69	1ywh	A VAL69	-6.9	-15.8	0.43	0.31	1.55	15.78	42.64	E E
2i9b	E VAL70	1ywh	A VAL70	8.1	-7.3	0.43	0.34	2.14	11.43	43.41	E E
2i9b	E CYS71	1ywh	A CYS71	13.5	7.6	0.45	0.45	2.6	15.1	44.96	E E
2i9b	E GLY72	1ywh	A GLY72	-17.2	23.2	0.45	0.48	3.49	23.23	44.96	- -
2i9b	E LEU73	1ywh	A LEU73	-21.6	7.1	0.47	0.17	3.79	21.62	46.51	S S
2i9b	E ASP74	1ywh	A ASP74	3.1	-10.9	0.46	0.41	3.47	13.78	45.74	T T
2i9b	E LEU75	1ywh	A LEU75	7.9	-4.5	0.45	0.32	3	10.54	44.96	T T
2i9b	E CYS76	1ywh	A CYS76	5.8	-10.7	0.46	0.35	2.79	11.62	46.12	T T
2i9b	E ASN77	1ywh	A ASN77	0.5	21.9	0.47	0.42	2.41	21.9	47.29	T T

focus only on conformations with the same sequence. A simple objective for superpositioning is to minimize the squared distances between the corresponding C_α atoms of the two structures. The optimal translation is the vector that relates the repective controid (*i.e.* average over all points) of each structure. The classic solution to finding the optimal rotation vector is the Kabsch algorithm [45, 32] using singular value decomposition (SVD). Given protein structures X and Y which have been translated by their respective centroids, a covariance matrix $C = X^T Y$ is computed. The SVD of $C = U\Sigma V^T$ is calculated (with computational complexity $O(N^3)$ [23]) such that the columns of U are the eigenvectors of CC^T, the columns of V are the eigenvectors of $C^T C$ and Σ is a diagonal matrix containing the square roots of the non-zero eigenvalues (*i.e.* singular values) of both CC^T or $C^T C$. The optimal rotation matrix $R = VDU^T$, such that $D = \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{smallmatrix}\right)$ and $d = sign(determinant(C)) \times 1.0$.

2.2.2.2 S^3 Method

When analyzing a pair of protein conformations by using simple structural alignment tools from environments such as PyMOL [90], if the intact rigid structures don't align well, it is natural to align just one or two secondary structure elements first. Using this concept, we superposition secondary structure elements when bigger fragments of the protein don't align well. We use the Define Secondary Structure of Proteins algorithm (DSSP) [46] to define the individual secondary structure elements that correspond between the two structures. Instead of introducing twists into one of the structures, as is done in FATCAT [110], we calculate the minimum C_α distances that result from partial superpositioning. We don't create a continuous superimposed structure because we are only interested in the minimum C_α distances used in describing the relative flexiblity of the individual residues of the protein.

Algorithm 1 superpositions the two structures using the standard Singular Value Decomposition (SVD) algorithm within the **DistancesFromSuperposition** function, and if the RMSD is within the cutoff ($1\mathring{A}$) then the algorithm is complete. Otherwise, it finds a single midpoint of the reference structure that's not within a secondary structure element (SSE), or 2 "midoints" if there

is a SSE at the midpoint, and computes the translation and rotation of the optimal superposition and checks the resulting RMSD recursively for each side. If the midpoint was within a long enough SSE, it superpositions the residues of that SSE as well. At each superpositioning, we update the C_α distances to be the minimum distance seen so far. Figure 2.6 shows specifically the S^3 iterations of applying Algorithm 1 (the S^3 algorithm) to the S100 calcium sensor protein.

There are faster and more robust methods for solving the superposition problem, but our S^3 implementation uses the classic approach because we are mainly superpositioning small subsets of the coordinates where the efficiency of these small superpositions is not an issue. The details of finding the miniminum C_α distance using the SVD algorithm are given in **Function Distances-FromSuperposition**.

The difference between the C_α distances found by S^3 versus those found by FATCAT for the example of the receptor protein $uPAR$ discussed previously can be seen under the column headings S^3 and Fatcat in Figures 2.4 and 2.5. Notably, the flexible hairpin residues in the green boxes have the highest S^3 C_α distances, and are the only residues with S^3 C_α distances $> 1.5\mathring{A}$. This example is typical of the types of differences we see in the two different alignment algorithms.

2.2.3 Difference Distance Matrix Percentages

In [82] the intramolecular distances between 2 structures are compared without requiring superpositioning, resulting in an $O(N^2)$ triangular difference matrix (DM), where N is the number of residues in each protein. Differences obtained from subtracting the DM's of 2 structures results in a Difference Distance Matrix (DDM). In [82] an RMSDD is computed to measure the overall similarity of the 2 structures based on their respective difference distance entries and furthermore, elements of DDM's for comparing independently determined crystal structures of the same protein were characterized as "large" outside the range of -1 to 1Å.

Because we are interested in per residue flexibility measures, we summarize the percentage of values outside of the range of -1 to 1Å from the DDM for each residue, and assign each a Difference Distance Matrix Percentage (DDMP) value. In a protein with two extremely different

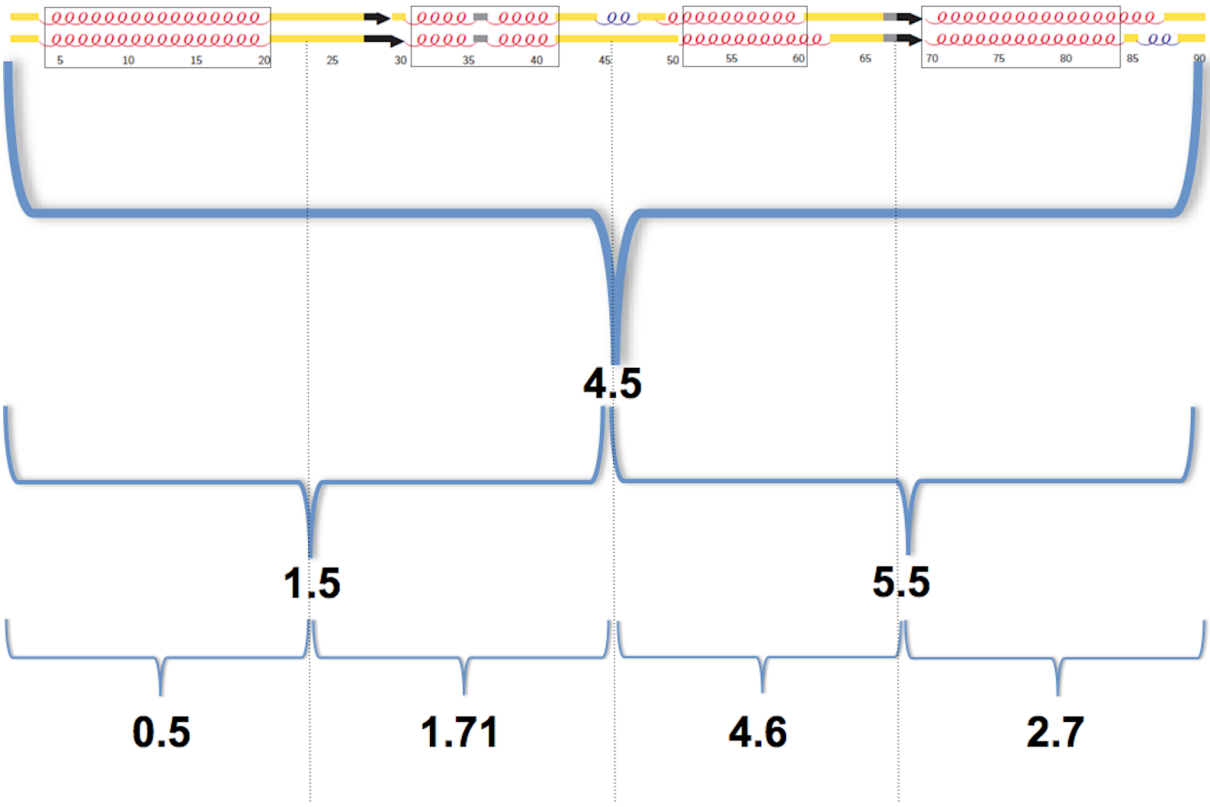
Algorithm 1: Superposition by Secondary Structure Elements (S^3)

Input : X,Y arrays of 3 dimensional coordinates of the C_α atoms for 2 sequence-aligned structures

Output: $MinC_\alpha Dist$ vector of the min C_α distance for each sequence-aligned residue from X and Y

```
1  MinSubDomainLength ⟵ 8
2  RMSDCutoff ⟵ 1.0
3  First ⟵ 1
4  Last ⟵ N; (N=Number of aligned residues)
5  RMSD,MinCαDist ⟵ DistancesFromSuperposition (First,Last,X,Y)
6  if RMSD > RMSDCutoff and (Last - First) > (2 × MinSubDomainLength) then
7      ToExploreList ⟵ ∅; (List of SubDomains specified by (First, Last) residues)
8      AppendToExploreList (First,Last)
9      SSEList ⟵ List of (First, Last) residues for aligned Secondary Structure Elements
       defined using DSSP [?] (e.g. α-helices and β-strands) with length ≥ 4 residues
10     while ToExploreList ≠ ∅ do
11         First,Last ⟵ GetSubdomainFromList (ToExploreList)
12         MidPt1,Midpt2 ⟵ FindMidPoints (First,Last,SSEList); if the midpoint is within
           a SSE then set MidPt1 and Midpt2 to flank the SSE otherwise MidPt1 =
           Midpt2 = the midpoint between First and Last
13         if MidPt1 - First > MinSubDomainLength then
14             RMSD,MinCαDist ⟵ DistancesFromSuperposition (First,MidPt1,X,Y)
15             if RMSD > RMSDCutoff and (MidPt1 - First) > (2 × MinSubDomainLength)
               then
16                 AppendToExploreList (First,MidPt1)
17         if Last - Midpt2 > MinSubDomainLength then
18             RMSD,MinCαDist ⟵ DistancesFromSuperposition (Midpt2,Last,X,Y)
19             if RMSD > RMSDCutoff and (Last - Midpt2) > (2 × MinSubDomainLength)
               then
20                 AppendToExploreList (Midpt2,Last)
21         if (Midpt2 - MidPt1) > MinSubDomainLength then
22             RMSD,MinCαDist ⟵ DistancesFromSuperposition (MidPt1,Midpt2,X,Y)
23 return MinCαDist
```

Figure 2.6: S^3 partial superpositioning of the S100 calcium sensor with boxes around matching secondary structure elements between the 2 conformations. Below, iterations of S^3 partial superpositioning are shown with resulting **RMSD**.



Function DistancesFromSuperposition($Begin, End, X, Y$)

Input : Begin and End are the positions of the subdomain to be superpositioned, and X and Y are the entire C_α coordinate arrays (with size $N \times 3$) used in the distance computation

Output: RMSD is the Root Mean Squared Deviation value of the subdomain superpositioning, and MinC_αDist vector is updated to contain the minimum distance between sequence-aligned residues from all superpositioning operations.

$X_{DOM} \longleftarrow X[\mathsf{Begin} : \mathsf{End}]$

$Y_{DOM} \longleftarrow Y[\mathsf{Begin} : \mathsf{End}]$

$Cen(X_{DOM} or Y_{DOM})_{i=1}^{3} \longleftarrow (\sum_{i=1}^{3} \sum_{j=\mathsf{Begin}}^{\mathsf{End}} XorY[i,j])/(\mathsf{End} - \mathsf{Begin})$; Compute centroids

$X_{DOM} \longleftarrow X_{DOM} - Cen(X_{DOM})$; $Y_{DOM} \longleftarrow Y_{DOM} - Cen(Y_{DOM})$; Translate both subdomains so its centroid corresponds with the origin of the coordinate system

$C \longleftarrow (X_{DOM})^T Y_{DOM}$; Compute a covariance matrix

$U\Sigma V^T \longleftarrow \mathtt{SVD}(C)$; Calculate the Singular Value Decomposition (SVD) of the covariance matrix C

$d \longleftarrow sign(det(C))$; $D \longleftarrow \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{smallmatrix}\right)$

$\mathsf{Rot} \longleftarrow VDU^T$; The optimal least squares rotation matrix.
$\mathsf{Tran} \longleftarrow Cen(X_{DOM}) - (Cen(Y_{DOM} \cdot \mathsf{Rot})$

$(Y_{DOM})' \longleftarrow (Y_{DOM} \cdot \mathsf{Rot}) + \mathsf{Tran}$

$\mathsf{RMSD} \longleftarrow \sqrt{(\sum_{i=1}^{N} (X_{DOM})_i - (Y_{DOM})')_i)^2/N}$

$Y' = (Y \cdot \mathsf{Rot}) + \mathsf{Tran}$

for $j \longleftarrow 1$**to**N **do**

  MinC_αDist$[j] \longleftarrow Minimum(\mathsf{Min}C_\alpha\mathsf{Dist}[j], \mathtt{Distance}(Y'[j], X[j])$

return $RMSD, MinC_\alpha Dist$

conformations, the DDMP values tend to range from .5 to 1.0, with the more mobile subregions exhibiting relatively higher values.

In Figures 2.4 and 2.5 residues 33-35 and 59-62 exhibit the highest relative DDMP measures compared to the remainder of the domain, and these are close to the exact regions of the flexible hairpin turns described earlier [4]. Because there are somewhat larger, but local, changes in the other two domains (not reported here), we see that the DDMP measures throughout the domain reflect the movements in the other domains. In this case, there isn't a large-scale domain movement as seen in other proteins, where everything changes relative to everything else, as seen in the calmodulin measurements in Figure 2.7.

2.2.4 Dihedral angle differences

Dihedral angles describe the angle of rotation along the line of intersection of two planes, and can be computed from the coordinates of 4 contiguous (backbone) atoms such that the first three atoms define one plane and the last 3 atoms define the second. Considering the 3 bond vectors formed by the 4 contiguous atoms, the dihedral angle can also be described as the angle of rotation about the vector in the middle. Looking down on the middle vector, the dihedral angle describes the angular distance observed between the first and third vectors. The backbone dihedral angles of proteins are known as phi (ϕ), involving the backbone atoms $C' - N - C_\alpha - C''$, psi (ψ), involving the backbone atoms $N - C_\alpha - C' - N$ and omega (ω), involving the backbone atoms $C_\alpha - C' - N - C_\alpha$). The planarity of the peptide bond usually restricts ω to be either 180°; the typical *trans* case, or 0°; the *cis* case, with the *cis* case occuring much less frequently. We include $\Delta\phi$ and $\Delta\psi$ (the differences of corresponding ϕ and ψ dihedral angles from 2 conformations) in our analysis but ignore $\Delta\omega$, since the likelihood of a nonzero $\Delta\omega$ value is small.

Comparisons of dihedral angles measure differences in the ϕ and ψ backbone dihedral angles of NMR structures are computed in [113] and others measure differences between psuedo-dihedral angles of every 4 C_α atoms [39, 31]. The computation of backbone dihedral angle changes can highlight important regions with high shape pliabilities that may not be visible using superposi-

tioning alone. Because each psuedo-dihedral angle involves the inclusion of neighboring residues, we focus only on the individual ϕ and ψ dihedral angle differences involving the C_α atom for each (computationally feasible) residue to obtain a per residue value used in computing shape pliability scores.

In [74], a method called dihedral angle transition (DTA) characterizes the effects of large dihedral angle changes, including those changes described as $\Delta\phi$ and $\Delta\psi$ above. Large-scale transitions to the backbone of protein fragments are defined as Δ's $\geq 120^\circ$, or the sum of ϕ and ψ Δ's per residue $\geq 120^\circ$ during comparisons of structural pairs of proteins. Our scoring functions, described below, do not classify the dihedral Δ's as large or small, but the range of changes in the highest bin of shape pliability scores include those with values similar to the large-scale transitional changes in DTA.

In Figures 2.4 and 2.5 residues 34-36 and 59-60 have relatively high $\Delta\phi$ or $\Delta\psi$ measures, and these residues are all found within the hairpin turns described in [4] as flexible. In this running example, we see that the three measurements from alignment, intermolecular distances and dihedral angles all indicate increased flexibility, and thus influence both shape pliability and mobility, described in more detail in the next section. Often, high levels involving both characterizations of backbone changes are found in the most extremely flexible regions. On the other hand, there are a number of examples discussed in Chapters three and four demonstrating that one of the characterizations plays a much greater role in quantifying the flexibility of the backbone in a particular region or domain.

We considered taking the average of the $\Delta\phi$ and $\Delta\psi$ per residue for our shape pliability score instead of the maximum over the pair of dihedral angle changes per residue, but this resulted in too much "smoothing" over the high dihedral angle differences. Especially in cases (which are not uncommon) where there is a pair of dihedral angles differences such that a high $\Delta\psi$ of one residue is immediately followed by a high $\Delta\phi$ in the next residue, averaging over the $\Delta\phi$ and $\Delta\psi$ of each residue tends to diminish the relative spikes of the observed high Δ's. For example, in Figure 2.4, the the high shape pliability score for residue 36 would be greatly diminimished if the $\Delta\phi$ and $\Delta\psi$

were averaged, instead of using just the maximum high value of the $\Delta\phi$. And in Figure 2.5, all of the high shape pliability scores for the flexible loop of residues 59-62 would be greatly diminished. The advantage of using the maximum Δ over the average was especially evident in studying the hinges of adenlyate kinase compared to what was found in the literature; as presented in Table 4.4.

2.2.5 Scoring Functions for Mobility and Shape Pliability

Our scoring functions apply to individual residues, and are scaled so that the values for shape pliability and mobility scores range between 0 and 100, with higher scores indicative of the respective flexiblity. The scores are based on the premise that shape pliability or local changes to the structure involve both changes to dihedral angles and distances between the C_α atoms of residues that don't superimpose exactly, but are within a specified tolerance or cutoff parameter (using our S^3 algorithm described above). When the distances from superpositioning are greater than our cutoff distance (typically set to 1.5 Angstroms), then we assume the residues are involved in mobility. Mobility can also be high when the percentage of difference distances from the DDM (*i*.e. the DDMP value) is high. There are many possible calculations for combining individual measurements into a scoring function, and the following simple function is designed to emphasize any evidence of flexibility:

(1) **foreach** residue compute:

- ${S^3}_{SCALED} \longleftarrow Scaling\_Factor_1 \times C_\alpha$ distance computed from Algorithm 1.
- $\Delta\phi\psi_{SCALED} \longleftarrow Scaling\_Factor_2 \times (\mathbf{max}(\Delta\phi, \Delta\psi))$
- DDMP (Distance Distance matrix percentage).

(2) **if** ${S^3}_{SCALED} < Cutoff$ **then**

Shape Pliability $\longleftarrow \mathbf{max}(S^3_{SCALED}, \Delta\phi\psi_{SCALED})$

else Shape Pliability $\longleftarrow \Delta\phi\psi_{SCALED}$

(3) **if** ${S^3}_{SCALED} \geq Cutoff$ **then**

Mobility $\longleftarrow \mathbf{max}(S^3_{SCALED}, DDMP)$

else Mobility $\longleftarrow DDMP$

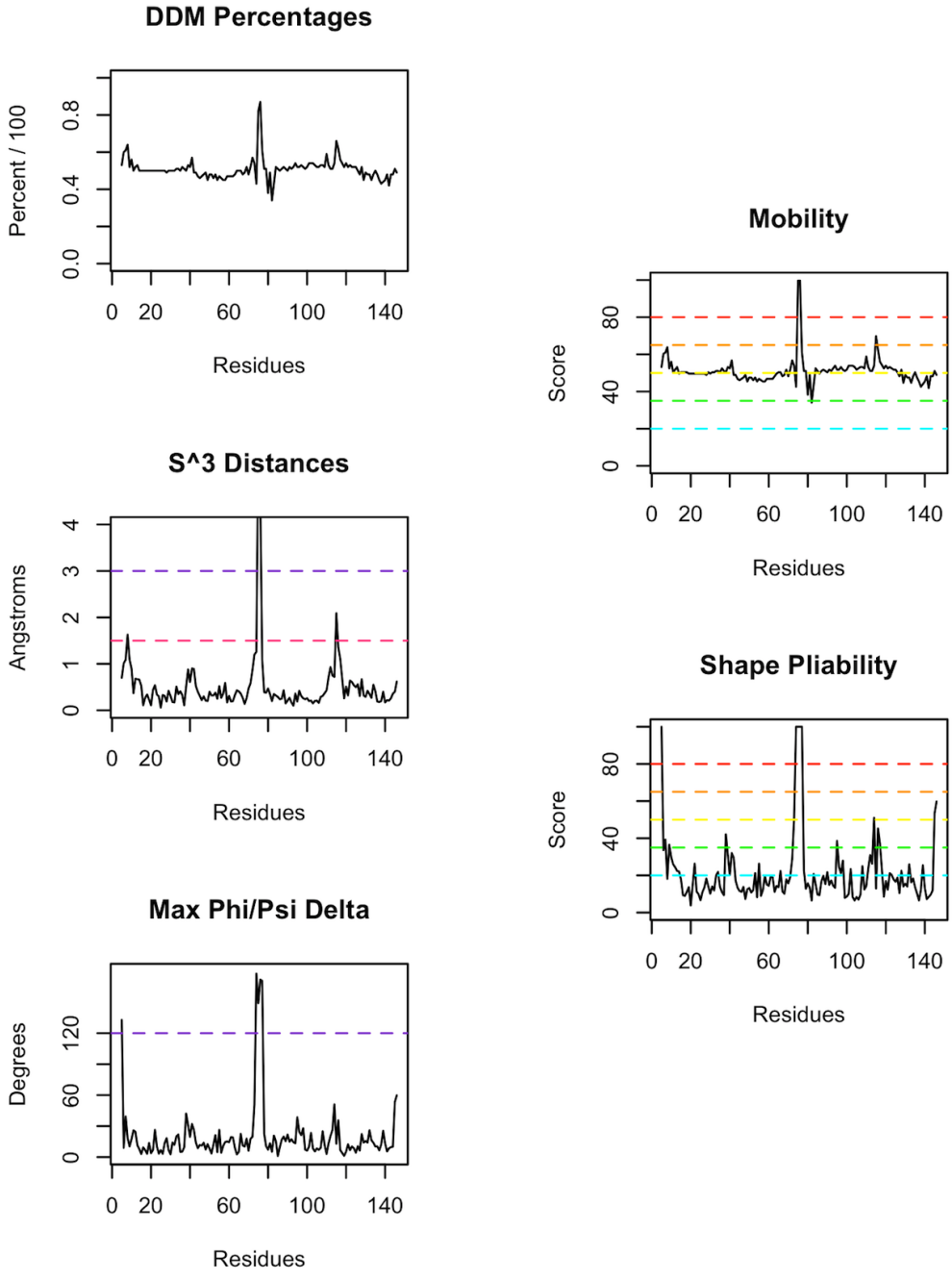
We demonstrate the combination of individual measurements into shape pliability and mobility scores in Figure 2.7. In observing the Calmodulin measurements (left column of Figure 2.7) and scores (right column of Figure 2.7) the mobility scores reflect mainly the DDMP measures, which have similar values of near 50% throughout, because all residues have moved with respect to each other. The peak near residue 75 is increased in value (over the DDMP measurements) due to the S^3 distances in that region. The shape pliability scores reflect mainly the ϕ/ψ maximum deltas in this particular protein, and the shape of the shape pliability score graph is almost identical to that of the Max Phi/Psi Delta measurements.

2.3 Discussion and Summary

Our shape pliability involves changes in dihedral angles and sometime subtle differences in C_α structural alignments. This allows us to locate different flexible regions than other computational methods find, and is perhaps advantageous. Having both mobility and shape pliability scores within the same framework allows us measure flexibility for a wide variety of proteins, as will be shown in Chapter 3.

Future improvements (which are not included in the thesis) to S^3 might replace the least squares optimization with maximum likelihood or another more efficient superpositioning, such as quaternions [21]. In the future we might also consider a generalization to find C_α distances for homologous proteins. This would require an alignment step first, followed by the superpositioning, and might also require modifications to S^3.

Figure 2.7: Calmodulin measurements (left column) and flexibility scores (right column).



Chapter 3

Method Validation

3.1 Introduction

The new method described in detail in Chapter 2 is a novel way of characterizing protein backbone flexibility that was developed to fill the void for accessing flexibility information at the residue level. This information is required for protein structural flexibility prediction and also for understanding the variations in x-ray crystal structures of different conformations of the same protein. However, because this method is novel, there are no comparable methods to validate the residue level scores we obtain. We therefore approach validation in two ways: 1) by comparing our results to a sufficiently large database of proteins with documented flexibilities pertaining to the problem of protein-protein docking and 2) by examining flexibilities of 10 indiviual protein pairs and comparing our findings to the descriptions of conformational differences in the structure literature. Both approaches have some limitations but this strategy avoids the pitfalls of validating a computational method with other computational methods that don't provide the same type of flexibility information.

The protein-protein docking benchmark [27, 42] provides a database of protein complexes and their constituent monomers. The benchmark classifies the complexes according to the perceived difficulty for docking, such that the more flexible the interface, or specifically, the more the monomers have changed at the interface upon being bound in the complex, the more difficult the docking problem is. Thus, the dateset contains a set of proteins such that the apo (unbound) monomer conformations can be compared to the same proteins bound in complex, and furthermore,

each protein complex is classified by the the residue flexibility at the interface. We selected a subset of proteins from the benchmark to see how our measures of flexibility compare to the categories of flexiblity in the benchmark. We present the comparisons in §3.2.

In §3.3 we present an analysis of pairs of selected protein conformations using coordinates obtained from the Protein Data Bank (PDB) [9]. In contrast to the docking benchmark section which focuses on classes of flexiblity and average flexibility measurements pertaining to each protein or classification of docking difficulty, this second analysis validates the method at the residue level. The proteins were chosen from the [27, 42] databases to cover a range of different motions and protein sizes. The specific examples provided here are cases where the literature describes the flexibility found in different conformations of the protein, so that we can compare these published descriptions with our findings. We also selected PDB files (and chains) that were specifically mentioned in the literature, when possible, to more precisely match the residues in the descriptions and to our corresponding flexibility scores.

3.2 Validation for Categories from the Docking Benchmark

The protein-protein docking benchmark 4.0 consists of 176 complexes of high-resolution, non-redundant structures with their unbound constituents [42]. The complexes are classified by rigid, medium or difficult and the benchmark is expressly useful for development and assessment of computational protein-protein docking methods. The classifications are given based on the Root Mean Squared Distances (RMSD) in the positions of all interface residue C_α atoms after superpositioning the bound and unbound monomers; called the $I\text{-}RMSD$. Interface residues have C_α atoms within a cutoff distance of $5\mathring{A}$ between bound proteins in the complex. These residues are used to calculate $f_{non\text{-}nat}$ which is the fraction of non-native contacts of the superposed unbound structures divided by the total number of contacts in the complex interface [69]. The three classifications are:

(1) **Rigid:** $I\text{-}RMSD \leq 1.5\mathring{A}$ *and* $f_{non\text{-}nat} \leq 0.4$

(2) **Medium:** $[I\text{-}RMSD \leq 2.2\mathring{A}]$ *or* $[I\text{-}RMSD \leq 1.5\mathring{A}$ *and* $f_{non\text{-}nat} > 0.4]$

(3) **Difficult:** $I\text{-}RMSD > 2.2\mathring{A}$

3.2.1 Categories of Docking

We selected a subset of protein complexes from the benchmark consisting of complexes made up of only two constituent proteins where at least one of the bound proteins could be compared to its unbound monomer. The two requirments for selecting the unbound monomers were: 1.) structure was obtained by X-ray diffraction and 2.) the PDB files for the bound and unbound conformations were amenable to analysis by our method. The subset of bound/unbound protein conformation pairs were categorized by their classification in the docking benchmark and contain roughly equivalent numbers of the three categories: 22 Flexible, 31 Medium and 28 Rigid. "Flexible" is synonomous with the "Difficult" category defined above, because the difficult complexes have more flexible interfaces. Interface residues for the complexes were calculated using Rosetta [61] using a cutoff distance of $5.5\mathring{A}$ instead of the $5.0\mathring{A}$ cutoff used by the benchmark which may result in our comparisons having a higher number of residues considered to be on the interface of the complex.

We examine the trends in shape pliability and mobility for each of the docking benchmark classifications. In Figures 3.1 and 3.2, we look at the whole population of residues or interface residues distinguished by classification of rigid, medium or flexible without regard to individual proteins. The scores on the x-axes reflect shape pliability or mobility from less flexible to more, whereas the y-axes give the percentage of that score seen in the subset of residues or interface residues of the three classifications. The trends for the Mobility graphs (right side) show a higher percentage of less mobile residues with the Rigid Interface than with the Flexible Interface, and likewise, a higher percentage of very mobile residues (80-100 bins) in the Flexible versus Rigid Interfaces. The Medium Interface residues seem to fall somewhere in the middle, as expected. On the other hand, the Shape Pliability graphs (left side) do not show these expected trends. The mobility scores are computed using distance measures, whereas shape pliability scores also involve dihedral angle calculations. The docking benchmark uses only distance measurements in

the calculations of *I-RMSD* and $f_{non\text{-}nat}$ contacts. Therefore, we could expect a discrepancy between shape pliability measurements and classifications of interface residues, and in fact, that is what we see in Figures 3.1 and 3.2.

In Table 3.1, we calculate the means and medians for mobility and shape pliability over the protein monomer scores for each protein conformation pair analyzed in the dataset. We then summarize over the set of means and medians to get the distribution statistics displayed in the table. Again, we see the expected trends for mobility scores across the three categories; that is, increasingly higher means and medians as the docking difficulty increases. The exception is in the minimum values (row 1) for each set of means or medians, but because this is a single value it is probably not statistically significant. We also see this trend in the shape pliability means, but not for the medians, and the minimum values in the first row again do not follow the expected trend. The discrepancy in the shape pliability trends compared to what would be expected for the medians are also likely to be a factor of the dihedral angle calculations, as discussed above for Figures 3.1 and 3.2 .

3.2.2 Docking categories for Individual Protein Monomers

Ideally, the average flexibility over all the interface residues per protein would correlate with the docking classification of the associated complex such that rigid complexes would have lower interface flexibility scores, difficult complexes would have higher flexibility scores and medium would, of course, fall somewhere in the middle. In Figure 3.3 we see a large spread for all classifications, irrespecive of the number of residues compared. The reasons that the docking classifications and our average flexibility measurements may be in contrast are the following: 1) the *I-RMSD* and $f_{non\text{-}nat}$ contacts that underly docking benchmark categories are based on the interface of *complexes* and our measurements are based on each separate bond monomer of the complex compared to its unbound equivalen. If both sides of the complex interface display an equal amount of flexibility, then this problem is avoided, but otherwise the flexibility measurements versus the docking benchmark calculations do not necessarily correlate. 2.) the docking calculations are based solely on distance

Figure 3.1: Each graph shows either the `mobility` (A) or `shape pliability` (B) score percentages within 5 score divisions for *all* the residues in each classified protein . The three interface classification: rigid, medium, and flexible (difficult), are obtained from the protein classifications in the Docking Benchmark [42]

.

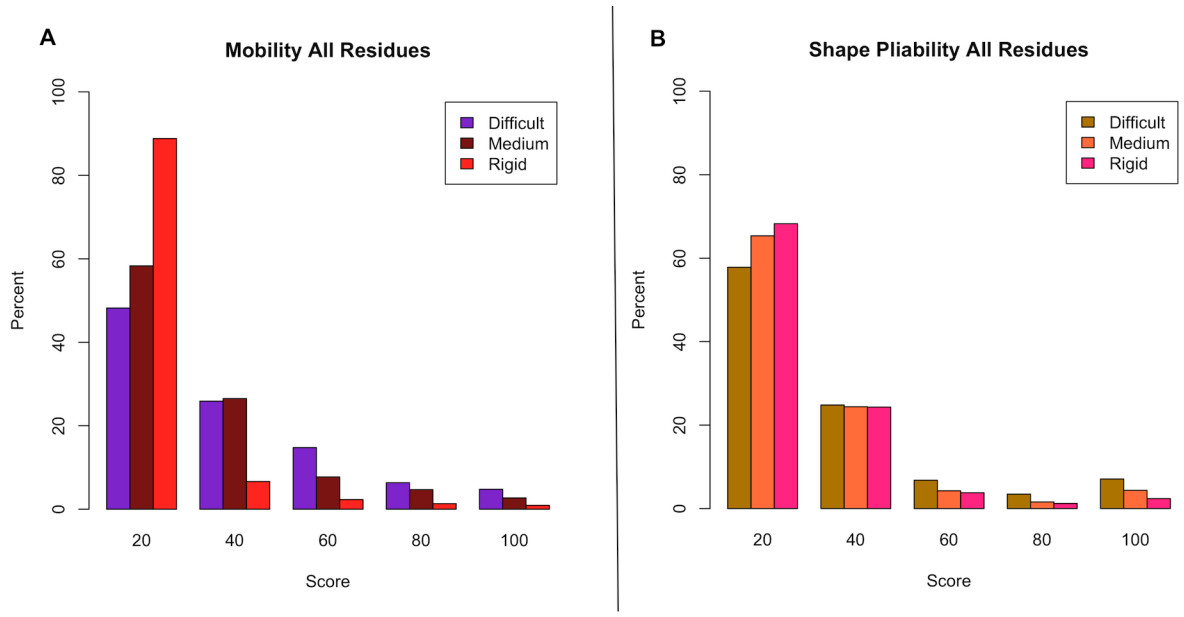
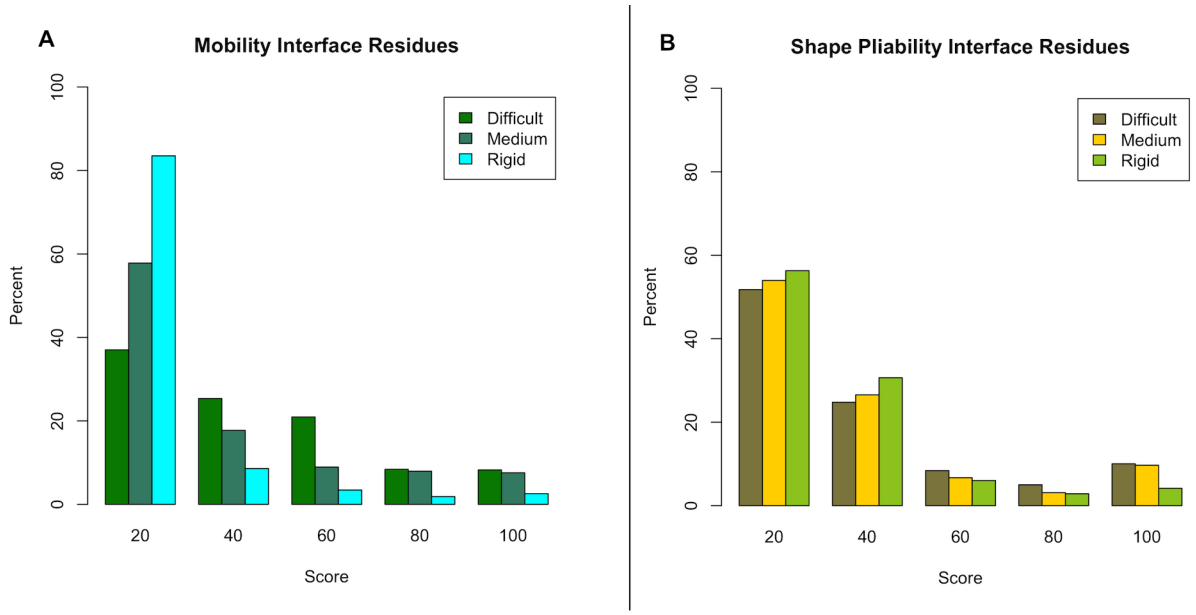


Table 3.1: Summary of the means and medians of mobility scores and shape pliability scores over all interface residues for each protein monomer with the specified docking classification.

Mobility Means				**Mobility Medians**		
	Rigid	Medium	Flexible	Rigid	Medium	Flexible
Min	0.05	2.71	1.03	0.00	1.06	0.42
1st Q.	2.97	12.74	17.82	1.10	6.75	6.00
Mean	8.08	25.62	36.07	3.02	11.49	29.54
Med.	11.40	25.96	33.11	6.08	18.09	26.78
3rd Q.	17.05	36.56	48.45	9.23	27.01	39.59
Max	31.91	62.09	66.10	29.50	66.12	69.38
Shape Pliability Means				**Shape Pliability Medians**		
	Rigid	Medium	Flexible	Rigid	Medium	Flexible
Min	12.44	10.51	11.01	10.46	6.77	9.99
1st Q.	19.71	21.06	22.24	15.00	14.90	13.64
Mean	21.91	26.92	29.43	18.44	19.31	18.96
Med	24.72	28.60	31.38	18.86	20.40	25.45
3rd Q.	27.99	35.43	40.16	21.29	24.43	30.04
Max	45.97	52.15	71.37	31.98	43.06	83.15

Figure 3.2: Each graph shows either the `mobility` (A) or `shape pliability` (B) score percentages within 5 score divisions for all the *interface* residues. The three interface classification: rigid, medium, and flexible (difficult), are obtained from the protein classifications in the Docking Benchmark [42]

.



calculations, and our shape pliability measurements also include torsion angle differences which may not coincide with the distance differences. 3.) each complex and its interface is given a single categorical discription (rigid, medium or difficult), whereas our calculations measure two distinct types of flexibility. Because the spread of average flexibilities in Figure 3.3 does not correlate well with the docking categories, we further investigate a few of the outliers below.

First, we analyze the two highest average scores from the rigid category. In Table 3.2 we see that the flexibility scores are very high for the monomer 2CGA_B compared to 1CGI_E. We also notice that the *I-RMSD* for the complex is 2.02. This complex seems to be miscategorized, as the cutoff for rigid complexes is $I\text{-}RMSD \leq 1.5\mathring{A}$. The second rigid monomer outlier 1BR9_A compared to complex 1GXD_A:C, with $I\text{-}RMSD = 1.39$, is within specification but has the second highest *I-RMSD* value in the set of rigid category complexes. After further visual examination (using pymol [90]) of the 21 interface residues and their superpositioning from unbound monomer to bound protein in the complex, the average score of 29.9 is representative of the visible differences in the relative positions. Perhaps the average score of somewhere near 30 is the upper bound for a rigid protein monomer, and this second "outlier" is within the range of normal rigid scores.

The 4 difficult category monomers with the lowest mobility averages were easier to validate. The outlier average flexibilities for the four are 1) 3DNI=1.14, 2)1ZM8=4.54, 3) 1ILR=10.1 and 4) 1KWM=12.78. In 1) 3DNI compared to complex 1ATN_D, the other monomer from the interface is an NMR structure so it wasn't compared, but 3DNI superposes onto 1ATN_D with almost no differences. For outlier 2) 1ZM8 compared to bound monomer of complex 2O3B_A, again there are no visible differences, and the comparison for the other half of the complex, namely 2O3B:B compared to unbound monomer 1J57_A exhibits all of the variation but was not analyzed using our measurements due to PDB issues. In 3) 1ILR_1, the comparison to complex 1IRA_YX shows a large domain movement for the other binding partner, but very little variation in the 1ILR:1 comparision to 1IRA_X. And for the final outlier examined, 4) 1KWM, the situation is that the other binding partner is again an NMR structure so the comparison has not been performed. For the monomer 1KWM_A compared to the bound conformation in the complex of 1ZLI_A, the overall comparison

involves many residues without coordinates in 1KWM_A, but the interface residue coordinates are present and do not appear to vary significantly from those in the complex. In conclusion, it appears that for the classification of these four "flexible" protein monomers with low average shape pliability and mobility scores, the partners of the complexes exhibit most of the flexibility, and these are all cases where only one half of the docking interface is flexible.

Figure 3.3: The mobility and shape pliability scores per residue are averaged, and the mean flexibility of all the interface residues per monomer is plotted against the total number of residues compared for the monomer. The three interface classifications: rigid, medium, and flexible (difficult), are obtained from the protein classifications in the Docking Benchmark [42].

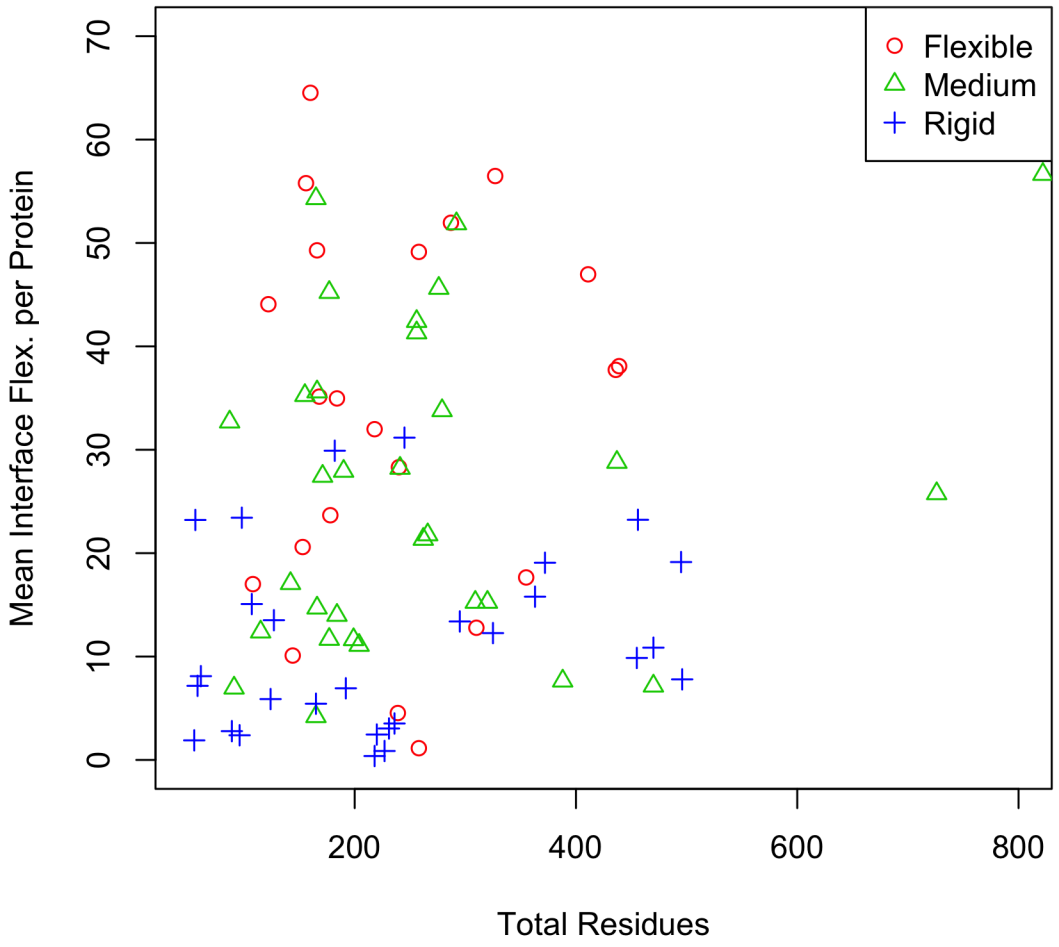


Table 3.2: Table of the *I-RMSD* per rigid complex ([42]) and the means for mobility and shape pliability scores calculated over all interface residues for the monomer-bound protein conformation comparisons.

Complex RIGID (28)	I-RMSD	PDBID1	I-Mob. Mean	I-SP Mean	PDBID2	I-Mob. Mean	I-SP Mean
1KXQ_H:A	0.72	1KXQ_H			1PPI_	8.0	21.3
1AVX_A:B	0.47	1QQU_A	1.8	19.7	1BA7_B		
1AY7_A:B	0.54	1RGH_B	2.0	17.5	1A19_B	2.6	23.8
1BVN_P:T	0.87	1PIG_	20.9	33.7	1HOE_		
1CGI_E:I	2.02	2CGA_B	31.9	43.9	1HPT_	22.4	33.2
1D6R_A:I	1.14	2TGT_	3.4	19.3	1K9B_A	8.1	30.6
1DFJ_E:I	1.02	9RSA_B	6.0	25.9	2BNH_	26.0	19.7
1E6E_A:B	1.33	1E1N_A	10.1	15.0	1CJE_D	13.9	22.5
1EAW_A:B	0.54	1EAX_A	0.8	15.2	9PTI_	1.9	20.9
1EWY_A:C	0.80	1GJR_A	13.5	27.1	1CZP_A	23.6	46.0
1F34_A:B	0.93	4PEP_	12.0	21.7	1F32_A	13.9	26.4
1FLE_E:I	1.02	9EST_A	3.0	20.2	2REL_A(4)		
1GL1_A:I	1.21	1K2I_1	2.8	13.8	1PMC_A(6)		
1GXD_A:C	1.39	1CK7_A			1BR9_A	30.6	39.5
1JTG_B:A	0.49	3GMU_B	6.4	22.2	1ZG4_A		
1MAH_A:F	0.61	1J06_B			1FSC_	7.8	32.5
1OC0_A:B	1.00	1B3K_A	15.8	21.4	2JQ8_A(4)		
1OPH_A:B	1.21	1QLP_A	22.9	25.0	1UTQ_A	0.1	12.4
1TMQ_A:B	0.86	1JAE_	6.9	20.8	1B1U_A		

Table 3.3: Table of the *I-RMSD* per medium complex ([42]) and the means for mobility and shape pliability scores calculated over all interface residues for the monomer-bound protein conformation comparisons.

Complex MEDIUM (31)	I-RMSD	PDBID1	I-Mob. Mean	I-SP Mean	PDBID2	I-Mob. Mean	I-SP Mean
1ACB_E:I	2.26	2CGA_B	27.0	35.7	1EGL_		
1JIW_P:I	2.07	1AKL_A	7.5	15.1	2RN4_A(1)		
1M10_A:B	2.10	1AUQ_	6.6	30.4	1M0Z_B	24.0	29.0
1NW9_B:A	1.97	1JXQ_A			2OPY_A	7.8	22.4
1GRN_A:B	1.22	1A4R_A	25.6	26.8	1RGP_		
1HE8_B:A	0.92	821P_	14.1	29.9	1E8Z_A	26.8	32.4
1I2M_A:B	2.12	1QG4_A	55.9	36.4	1A12_A	5.3	10.5
1LFD_B:A	1.79	5P21_A			1LXD_A	30.3	40.8
1MQ8_A:B	1.76	1IAM_A	14.0	26.9	1MQ9_A	29.3	35.1
1R6Q_A:C	1.67	1R6C_X	17.8	20.8	2W9R_A		
1WQ1_R:G *	1.16	6Q21_D	35.1	50.1	1WER_	15.0	29.3
1XQS_A:C	1.77	1XQR_A	38.7	21.3	1S3X_A	2.7	18.8
1ZM4_A:B	2.11	1N0V_C	62.1	23.6	1XK9_A	8.1	25.4
2H7V_A:C	1.63	1MH1_			2H7O_A	21.1	23.5
2HRK_A:B	2.03	2HRA_A	10.6	17.8	2HQT_A	12.7	16.9
2J7P_A:D	1.93	1NG1_A	55.7	41.5	2IYL_D	44.3	31.2
2NZ8_A:B	2.13	1MH1_	47.9	41.7	1NTY_A	35.6	20.5
2OZA_B:A	1.89	3HEC_A			3FYK_X	43.6	52.1
2Z0E_A:B	2.15	2D1I_A	12.8	18.2	1V49_A(1)		
3CPH_G:A	2.12	3CPI_G	29.0	23.5	1G16_A	37.1	38.8

Table 3.4: Table of the *I-RMSD* per difficult complex ([42]) and the means for mobility and shape pliability scores calculated over all interface residues for the monomer-bound protein conformation comparisons.

Complex DIFFICULT (22)	I-RMSD	PDBID1	I-Mob. Mean	I-SP Mean	PDBID2	I-Mob. Mean	I-SP Mean
1F6M_A:C	4.90	1CL0_A			2TIR_A	15.0	29.6
1FQ1_A:B	3.41	1B39_A	47.1	71.4	1FPZ_F	21.6	29.4
1ZLI_A:B	2.53	1KWM_A	10.6	17.2	2JTO_A(6)		
2O3B_A:B	3.13	1ZM8_A	3.4	12.8	1J57_A		
1ATN_A:D	3.28	1IJJ_B			3DNI_	1.0	15.0
1BKD_R:S	2.86	1CTQ_A	49.7	71.4	2II0_A	38.5	11.0
1H1V_A:G	6.62	1IJJ_B			1D0N_B	58.2	22.3
1IBR_A:B	2.54	1QG4_A	36.3	31.3	1F59_A	38.8	22.2
1IRA_Y:X	8.38	1G0Y_R			1ILR_1	8.7	23.4
1JK9_B:A	2.51	1QUP_A	31.8	22.8	2JCW_A	19.6	23.8
1R8S_A:E	3.73	1HUR_A	66.1	52.9	1R8M_E	35.9	29.5
1Y64_A:B	4.69	2FXU_A	17.1	49.5	1UX5_A	48.9	44.7
2I9B_E:A	3.79	1YWH_A	49.3	38.7	2I9A_A	46.0	37.6
2OT3_B:A	2.79	1YZU_A	54.8	45.1	1TXU_A	29.4	19.3

3.3 Validation of flexibiliy scores for selected protein conformations

We apply our method to the comparison of pairs of X-ray crystal structures with the same sequence (or with minor mutations) for validation purposes. We have prepared the PDB coordinates by eliminating alternate location data for our results such that we always using the coordinates for Location "A". We left the residue numbering intact, and screen out cases where the sequence numbering for the 2 conformations in question does not agree. In the future, we can implement the renumbering, but because PDB files contain so much variability, this requires a very careful analysis and thorough testing.

Table 3.5 contains the list of selected proteins, their PDB codes and the motion described by [27] or [42]. The majority of the set of selected proteins have been studied extensively in the literature, either because they are functionally or commercially important or representative of a significant conformational change. During the process of verifying the analysis of conformational differences found between 2 distinct PDB structure files, some PDB files were specifically selected for analysis due to the discovery of references that made the same or similar comparisons for the protein in question.

For all cases presented, we analyze the mobility and shape pliability and present scores based on spectrum colors of the binned scores shown in Figure:3.4 A. See the previous chapter for a complete description of the measurements used to compute mobility and shape pliability scores. The heatmap colors display more flexbility and higher scores toward the red end of the spectrum, and conversely more rigidity at the blue end. In our discussions of individual proteins, we compare the flexibility found in our measurements with references for the structures. The references frequently describe the importance of the high mobility measurements relating to function, whereas the residues with high shape pliability scores found by our analysis are mentioned much less often. The shape pliability measurements, while perhaps more difficult to verify than the mobility measurements, are equally important in characterizing backbone flexibility. Furthermore, because shape pliability is more local in nature, the residues demonstrating high shape pliability may ulti-

Table 3.5: Proteins analyzed and compared to literature (articles cited within the text descriptions to follow.) The true positive rate (TP) measures the sensitivity of the residues we score with high mobility or shape pliablity compared to residues described as flexible in the literature, whereas the true negative rate (TN) measures the specificity of the residues that scored lower in mobility and shape pliability compared to the remainder of residues not mentioned as flexible in the literature. All shape pliable and mobile residues with scores ≥ 50.0 are classified as positive for flexibility in calculating the TP and TN rates. For proteins with large domain movements, TP and TN are recalculated with a cutoff ≥ 65.0 and printed in blue below the original ≥ 50.0 rates.

ID	Protein Name	*Motion**	PDB1	PDB2	$N^{\dagger}$	TP	TN
1	CobU	Sheer (Frag.)	1CBU:A	1C9K:B	180	24/29 (83%)	149/151 (99%)
2	S100 Calcium sensor	Sheer (Frag.)	1K9P:A	1K9K:A	87	47/47 (100%)	35/40 (87%)
3	HIV-1 Protease	Hinge (Frag.)	1RPI:A	3PHV:A	99	8/20 (40%)	71/79 (90%)
4	β-Lacto-globulin	Hinge (Frag.)	1BEB:A	1B0O:A	156	20/22 (91%)	132/134 (98%)
5	Che Y	Unclass. (Frag.)	3CHY:A	1CYN:A	126	9/15 (60%)	108/111 (97%)
6	Cytochrome P450BM-3	Sheer (Dom.)	1BU7:B	1JPZ:A	455	53/61 (87%)	373/394 (95%)
7	Adenylate Kinase	Hinge (Dom.)	1AKE:A	4AKE:A	214	68/68 (100%) 66/68 (97%)	80/146 (55%) 127/146 (87%)
8	Calmodulin	Hinge (Dom.)	1CLL:A	1CDL:A	141	4/4 (100%) 4/4 (100%)	44/137 (32%) 135/137 (99%)
9	G-Protein α_{i1} subunit	Refold (Part.)	1BOF:A	1KJY:A	299	20/33 (60%)	257/266 (97%)
10	Gelsolin (Actin)	Flex. Int.	1H1V:G	1D0N:B	327	130/130 (100%) 126/130 (97%)	39/197(20%) 167/197 (85%)

* Motion for Proteins 1-9 described in [27], Protein 10 classified in [42]. † N is the number of residues compared and excludes missing residues from either PDB.

mately be amenable to prediction, whereas mobility may be much more difficult to predict. The analysis in Chapter 5 of both types of flexibility and how they correlate to structural and energetic terms lends credence to these ideas.

We also present sequence diagrams (generated with TEXshade [6]) shaded by shape pliability and mobility using the same spectrum colors shown in Figure 3.4 A. Secondary stucture elements as defined by DSSP [46] (described in §5.2) of the two conformations compared are given above the shaded sequences, and the symbols used are defined in Figure 3.4 B. Major differences in secondary structure elements between the two conformations are usually mentioned within the literature descriptions, and also tend to have higher shape pliability scores.

3.3.1 Protein 1: CobU

The flexibility of CobU was presented in the previous chapter in Figure 2.1 to illustrate the concepts of mobility versus shape pliability. The CobU protein is an enzyme with dual functions (adenosylcobinamide kinase/adenosylcobinamide phosphate guanlyltransferase) and it occurs as a homotrimer assuming a pinwheel shape [99]. In the discussion of the structure of CobU complexed with GMP (PDB coordinate file 1C9K) compared to apo (uncomplexed) CobU (PDB 1CBU), the largest movement is at α-helix 2 (residues 33-48), with unwinding and rewinding at the the helix ends [99] . This is in agreement with the high mobility scores at helix 2, and the shape pliability at the helix 2 ends, as evident in the sequence Figure 3.5 C. [99] also describes weak electron density and conformation flexibility in the loops at residues 58-60 and 94-97. Similarly, we see higher shape pliability scores at residues 58, 96 and in the short 3-10 helix at residue Glu 100. There is also high shape pliability at residue 71, which is not confirmed as a flexible region in [99]. The movement of helix 2 is readily apparent in Figure 3.5 A. and the flexiblity at the helix ends in Figure 3.5 B.

3.3.2 Protein 2: S100A6 Calcium Sensor

The S100A6 calcium sensor protein is a member of the S100 family of Ca^{2+} binding proteins. Modified levels of expression of S100 family members may be complicit in Altheimer's disease,

Figure 3.4: Colors and Symbols used in the structure and sequence figures. A. The spectrum of heatmap colors are used for coloring sequence and structure figures, and gray represents residues with no coordinates present for at least one of the PDB coordinate files. The same scoring range is used for mobility and shape pliability scores. B. Secondary structure elements are defined for each PDB file using DSSP [46], and given symbolic representations in sequence figures, located above the sequences which are shaded by their shape pliability and mobility scores.

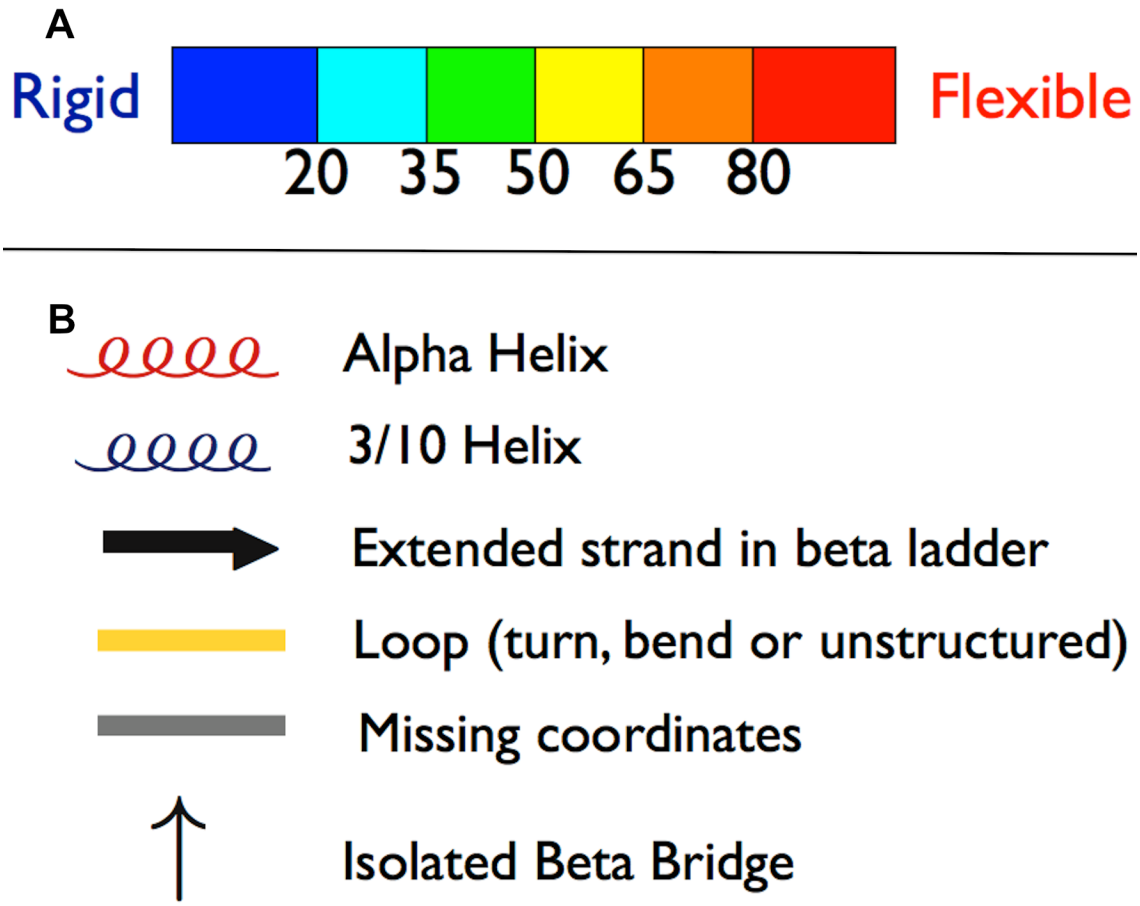
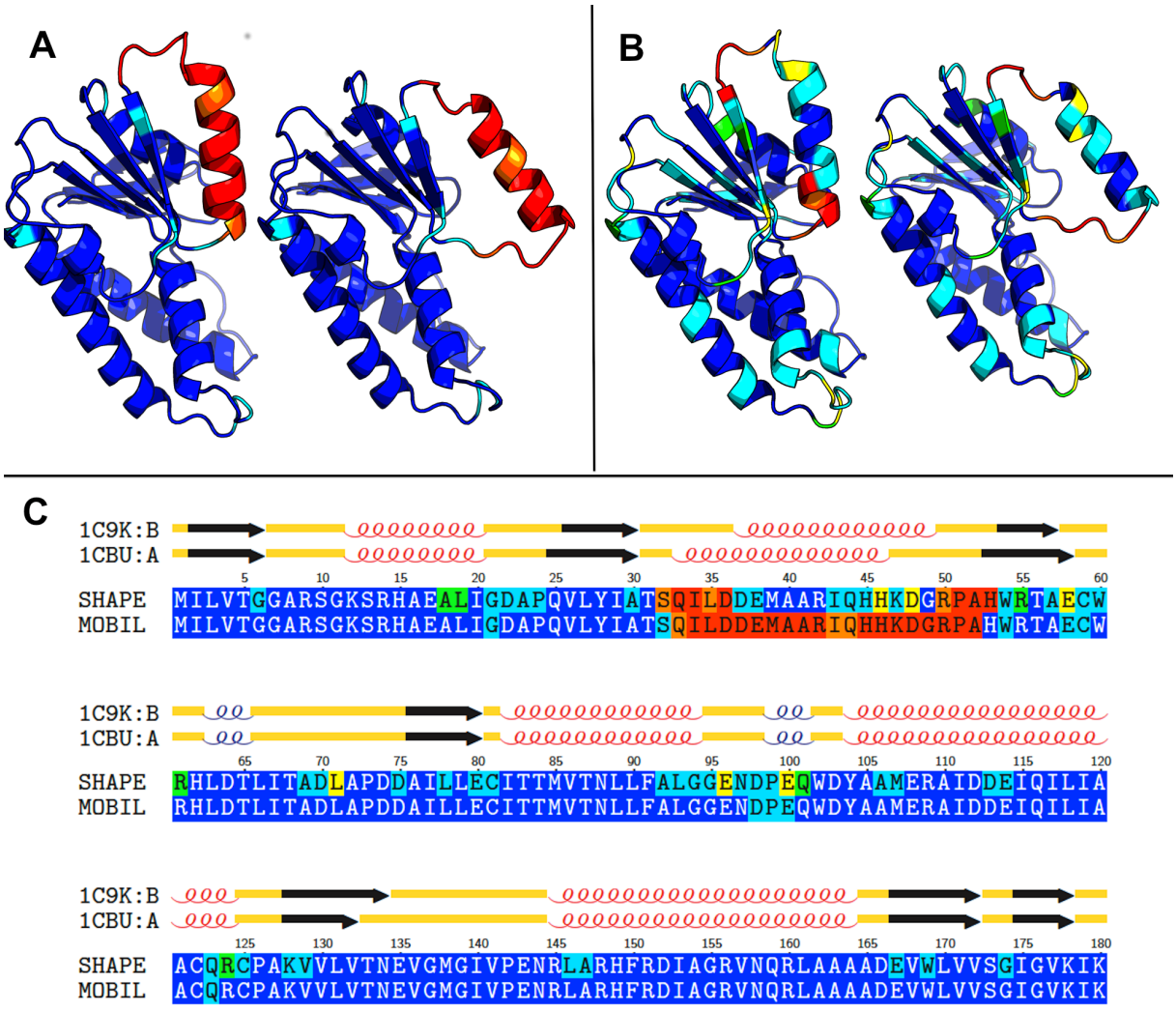


Figure 3.5: **Protein 1 CobU** Mobility and Shape Pliability: A. Structure comparison for the protein CobU from 1CBU(A) and 1CDK(B) colored by mobility, B. Structure comparison for CobU from 1CBU(A) (right) and 1CDK(B) (left) colored by shape pliability. C. Sequence comparison for Cobu from 1CBU(A) and 1CDK(A) colored by shape pliability (top) and mobility (bottom). Secondary structure elements are displayed above the sequences and described in Figure 3.4 B.



cancer and rheumatoid arthritus and more specifically, S100A6 has been found in complex with another S100 family member in human melanoma cells [88]. We compare the structures of the Ca^{2+}-free (PDB 1K9P) and Ca^{2+}-bound (PDB 1K9K) S100A6 calcium sensor states and discuss our findings compared to the reported conformational changes in [76].

The S100 family of proteins are hetero or homodimeric in structure and the monomers consist of 2 EF hand motifs defined as α-helix loop α-helix [88]. The N-terminal EF hand consists of helices I and II and the loop between them, and conversely, the C-terminal EF hand contains helices III and IV. The two Ca^{2+} binding loops link the helices in each of the EF hands.

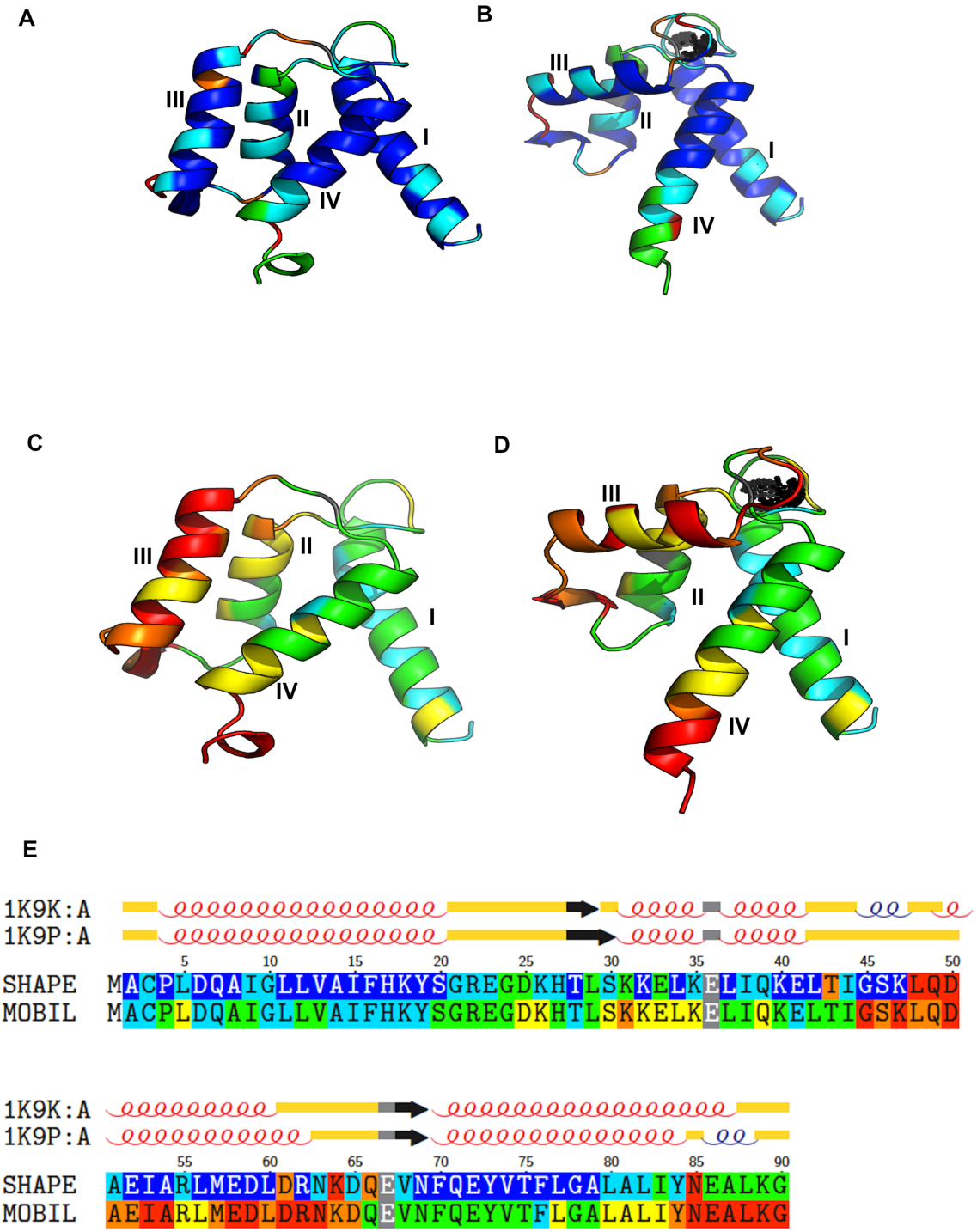
According to [76], a major difference between the Ca^{2+} bound and apo structures for the S100A6 sensor is the "dramatic reorientation" of helix III, which has both hydrophobic and hydrophilic side chains. After superimposing the 2 structures, the largest C_α deviations were reported to be in the Ca^{2+} binding loops, helices II and III with the linker between them, and the C-terminal end of helix IV [76]. These major differences affect residues 31-66 and 80-90.

Using our measurements, Figure 3.6 (C)(D) & (E) shows helix III, the linker between helix II and III, the linker between helix III and IV, and the C-terminal end as having the highest mobility scores; similar to the reported deviations. [76] describes Tyr84 as the beginning of the unwinding of helix IV, whereas our shape pliability scores indicate residue Asn85 as the largest contributer to the unwinding. Other areas of high shape pliability include the linkers between helices II and III and between helices III and IV. Our results confirm that these areas of shape pliability contribute to the extreme mobility of helix III.

3.3.3 Protein 3: HIV-1 Protease

In §4.2, the highlights of the findings for HIV-1 protease are described in detail. Here we note that the discussion of [65] in comparing Multi Drug Resistant (MDR) HIV-1 protease (1RPI) to the wildtype (3PHV) mentions specific differences for residues 7, 50, 80, and the "flaps", which typically include the region of residues 43-59. Also mentioned in the discussion is the fact that their comparison yields very different results when comparing the dimers (1.84 $\mathring{A}$) versus the individual

Figure 3.6: **Protein 2 S100 calcium sensor Ca^{2+}** free: A & C (1K9P:A) and bound to Ca^{2+}: B & D (1K9A:A). A & B are colored by shape pliability, C & D are colored by mobility. The 4 helices are labelled consecutively from the N to C terminus. The top lines of E are secondary structure definitions [46] and the bottom lines are shaded by the per residue shape pliability and mobility scores. Residues shaded gray have missing coordinates in the PDB file.



monomers (1.18 $\mathring{A}$) as is done for our measurements. Our flexibility findings include residues 16, 21, 27, 40, 48-53, 60-61, 64,7 9-80 and 95. The low sensivity score (40 %) is directly related to comparing individual monomers instead of the dimers, however, our further research showed some interesting findings using our flexibility scores to compare these monomers. Images for this protein are included with the discussion in 4.2.

3.3.4 Protein 4: β-lactoglobulin

β-Lactoglobulin is a commercially important whey protein present in the milks of many species, including all ruminants [50]. It is a member of the lipocalin family of proteins, and normally exists as a dimer. The monomers each consist of 8 β-strands forming a β-Barrel and the protein binds hydrophobic ligands, although its biological function is still unclear [50]. More specifically, the central cavity is called the calyx, and the β-barrel made up of 4 β-strands (A-D) forming one sheet, and a second sheet formed from strands E-H. We compare β-Lactoglobulin bound to palmitate (PDB 1B0O:A) [109] with its unbound form (PDB 1BEB) [14]. .

Access to the binding cavity, or calyx, is accomplished by the repositioning of Glu89 and the EF loop [109], which is in agreement with the highest mobility scores in Figure 3.7:E and Table 3.6. Additionally, [109] describes the GH loop as highly flexible. We measure a great deal of shape pliability for the GH loop residues (positions 109 to 117) but only residues Ser110 and Glu115 measure mobility. Another crystal structure determined during the same timeframe as 1BO0 for β-Lactoglobulin bound to 12-Bromododecanoic acid is described in [80], and this reference mentions mobility in loops CD (residues 61 to 65) and EF. This agrees with our measurements of high shape pliability in residues 61-65 and high mobility in residues 63 and 64. In [14], which is a reference for the determination of the unbound structure (1BEB); the authors also describes loop CD as a mobile surface loop, and missing coordinates for residues 1-4 and 161-162.

The high shape pliability scores we find in residues 33-34 may be explained by the strain in the structure described in [14] between the nitrogen of Ala34 and the δ oxygen of Asp33. We find similar shape pliability (crankshaft type) motion in residues 38-39, also in the AB loop, with no

definitive explanation in the structure references.

3.3.5 Protein 5: Che Y

The structure of magnesium bound Che Y was determined and compared in [7] (PDB 1CHN) with the previously determined wildtype Che Y [104] (PDB 3CHY). According to [7], Che Y functions as a response regulator within a system of bacterial chemotactic signal transduction. The system responds to changes in environmental chemical concentrations by alterations in swimming behavior such as tumbling versus smooth swimming. Magnesium binding in Che Y protein is required for its autokinase and autophosphitase activities, and also causes significant conformational changes.

We analyzed the differences in Mg^{2+} bound Che Y versus the unbound structure resulting in Figure 3.8 (A) and (B) showing the unbound and bound conformations respectively and colored by shape pliability and (C) and (D) likewise colored by mobility. Figure 3.8 (E) clearly shows that the fourth alpha-helix and the loop preceding exhibit the largest amount of mobility and shape pliability, and this is in agreement with the largest backbone conformational changes discussed in [7]. [7] also specifically mentions small backbone $C^{\alpha}\Delta$'s at residues 87,109 and 110 (all ¡ 1 $|AA$) and dihedral angle differences for residues 12 and 13. Except for residue 12, these small backbone changes are consistent with our analysis. We find higher mobility (and shape pliability) at the N-terminus, as well as slighly higher shape pliability at residues 37,44,45 and the C terminus. These residues are not mentioned specifically in [7], but the missing coordinates for residues 2 and 3 (in 1CHN) may be indicative of disorder at the N terminus. Also, [104] describes the backbone regions with the highest temperature factors as the two termini and the loops following α-helices 2 and 3. This is consistent with the flexibilities we see at both terminii, although we don't observe any flexibility in the loop regions with high temperature factors.

Figure 3.7: **Protein 4 β-Lactoglobulin:** A & C: unbound (1BEB:A) and B & D: bound to palmitate (1B0O:A). A & B are colored by shape pliability; C & D are colored by mobility; (E) Sequences with secondary structure for 1BEB and 1B0O above; sequence shading indicates shape pliability scores (top) and mobility scores (bottom).

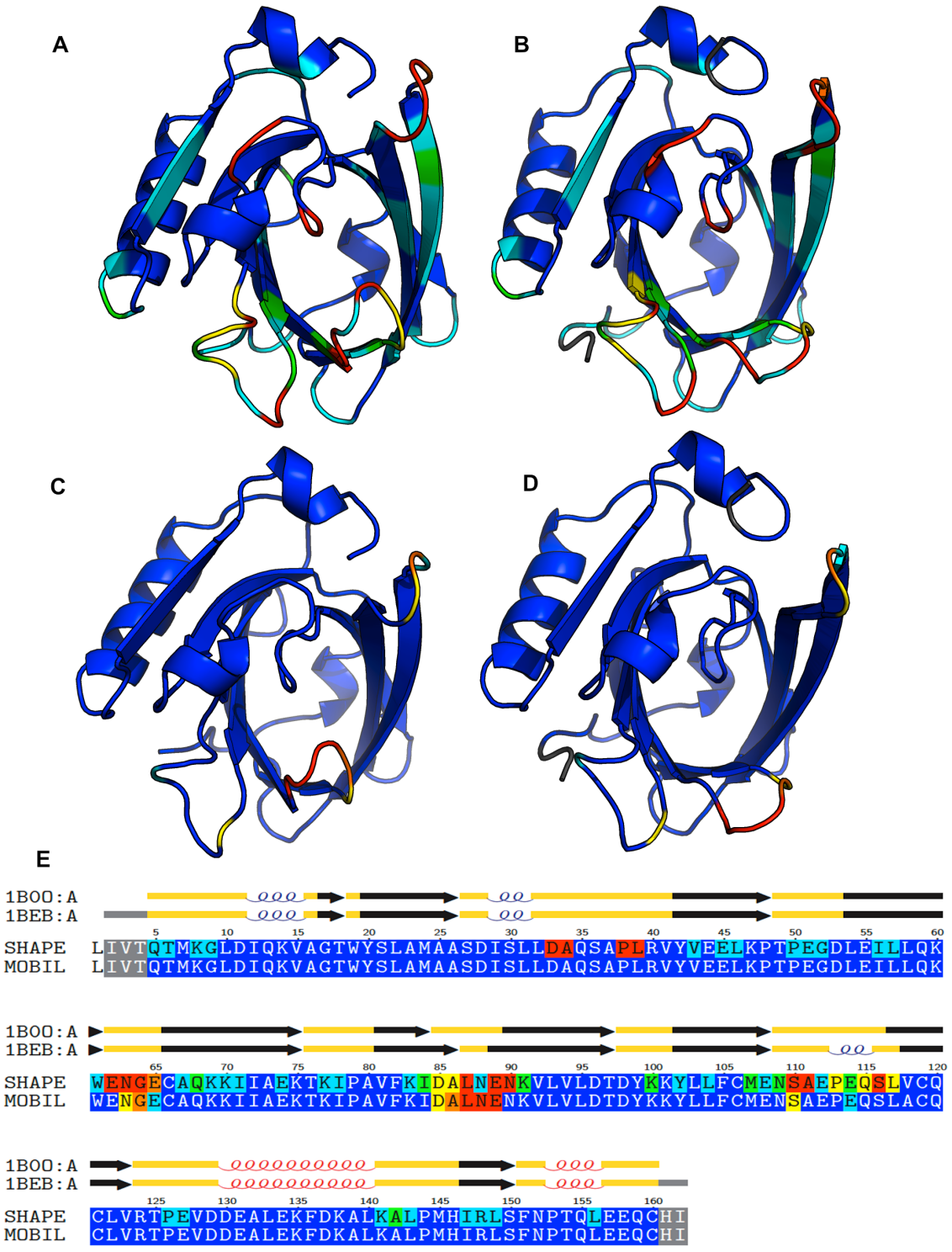
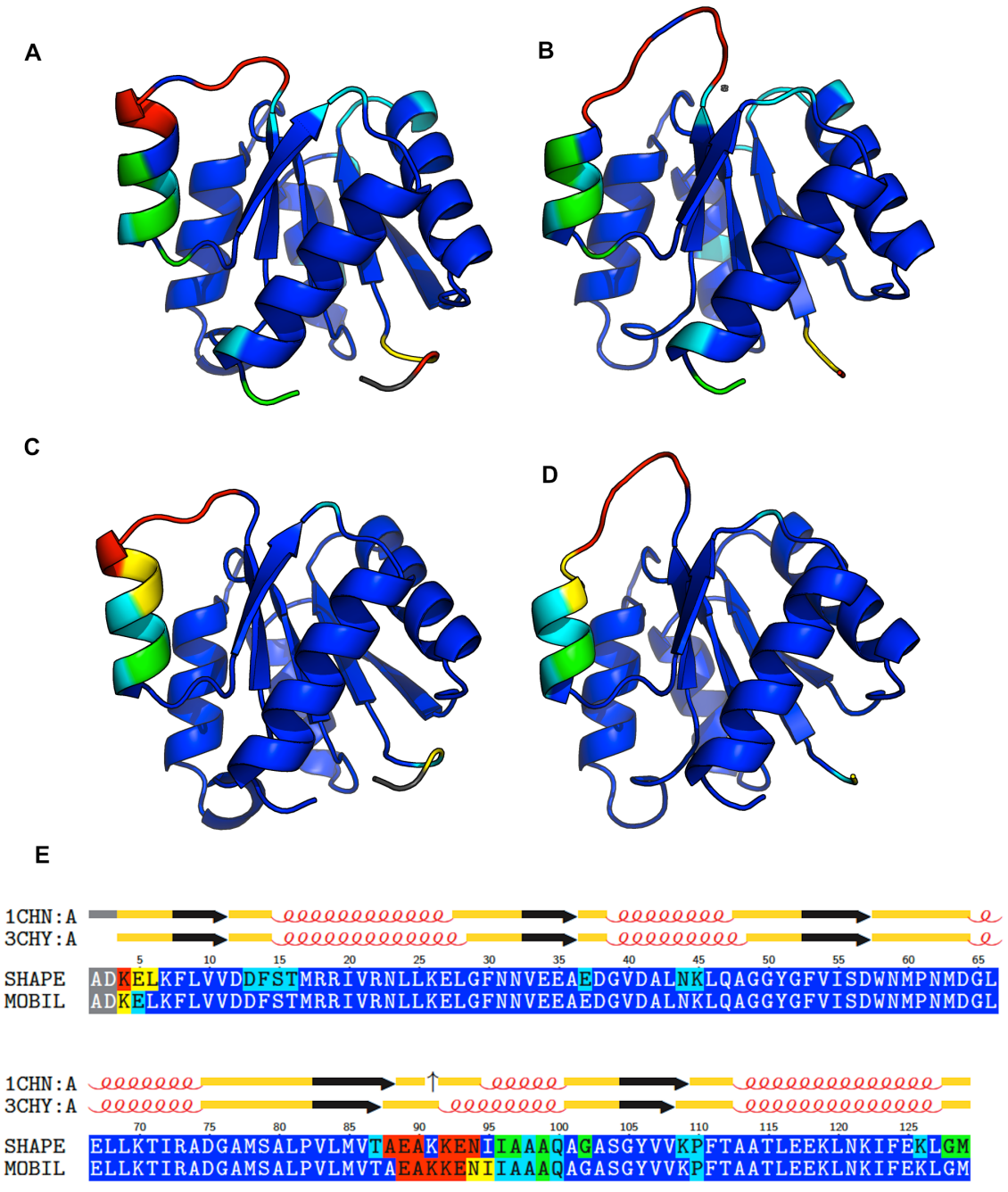


Table 3.6: **Protein 4 β-Lactoglobulin:** Measurements for residues exhibiting a high degree of shape pliability or mobility comparing PDB structures 1BEB:A versus 1B0O:A.

ResID	High Shape Pliability	High Mobility	S^3 Dist ($\AA$)	Δ Phi (Deg.)	Δ Psi (Deg.)	DDMP	SS	Ave. Rel. ASA
ASP33	√		0.46	-21.30	-175.11	0.07	S S	0.64
ALA34	√		0.15	171.73	-30.63	0.10	S S	0.45
PRO38	√		0.58	-22.67	-170.92	0.11	T T	0.50
LEU39	√		0.54	-160.78	-28.49	0.06	T T	0.12
GLU62	√		0.61	17.89	151.07	0.11	- -	0.42
ASN63	√	√	1.71	-161.24	91.05	0.32	S S	0.93
GLY64	√	√	2.38	-44.96	-81.76	0.44	S S	0.68
GLU65	√		1.50	94.50	15.51	0.32	S S	0.66
ASP85	√	√	1.60	-53.11	39.06	0.44	E S	0.74
ALA86	√	√	0.62	-61.69	-30.32	0.79	E S	0.41
LEU87	√	√	1.67	130.63	-97.63	0.87	T T	0.55
ASN88		√	1.54	-29.53	-18.60	0.81	T T	0.74
GLU89	√	√	2.10	-136.27	84.81	0.84	E -	0.30
ASN90	√		0.56	-53.46	-128.30	0.15	E E	0.21
SER110	√	√	1.68	27.67	-155.07	0.21	T S	0.34
ALA111	√		1.26	-137.60	-22.34	0.15	T S	0.89
PRO113	√		0.03	9.36	77.54	0.09	G T	0.33
GLN115	√		0.53	-34.69	52.75	0.19	G T	0.51
SER116	√		0.81	-0.88	-128.17	0.15	T S	0.14
LEU117	√		0.27	84.28	3.08	0.05	- E	0.05

Figure 3.8: **Protein 5 Che Y protein** (A) unbound (3CHY:A) and (B) bound to magnesium (1CHN:A) andcolored by shape pliability; (C) unbound and (D) bound to magnesium and colored by mobility; (E) Sequences with secondary structure for 3CHY:A and 1CHN:A above; shading around sequence letters indicate shape pliability scores (top) and mobility scores (bottom).



3.3.6 Protein 6: Cytochrome P450BM-3

The cytochrome P450 superfamily of heme proteins catalyzes the monooxygenation (enrichment with oxygen) of organic molecules and is found in all eukaryotes, most prokaryotes, and Archaea. This superfamily has been intensely studied because the biologically important P450 reactions include drug metabolism and fatty acid metabolism [91]. Cytochrome P450BM-3 is a fatty acid monooxygenase from the prokaryotic species *Bacillus megaterium*, but is similar in structure and function to eukaryotic P450's [35].

We analyze the structure of the heme domain of P450BM-3 in complex with N-palmitoylglycine (pdb 1JPZ:B) compared to the substrate-free heme domain (PDB 1BU7:B). The B chains were selected for comparison because Haines and colleagues found the largest structural differences in the B molecules between the substrate bound and free molecules, and this comparison is discussed in the structure determination paper for the substrate bound complex [35]. According to [35] the largest conformational changes are located in the "lid" domain consisting of helices F (residues 172-187), G (residues 199-226) and the loop between them. As demonstrated in figures 3.9 and 3.10, our analysis finds a great deal of mobility and some shape pliability in this region. In addition we find flexibility (both mobility and some shape pliability) in the loops that surrounding this region and while these residues are not specifically mentioned as flexible, the loop after helix G shows significant displacement in a figure depicting P450BM-3 movement in [35]. According to [35], other isolated areas of flexibility are located mainly in solvent exposed turns between secondary structures. Additionally, a clam shell movement to trap solvent molecules is described as involving the previously mentioned lid domain, the B'-helix (residues 73-82) and the amino-terminal area. Our measurements show agreement here, although the flexibilities for the B'-helix are not substantial. [35] describes changes to the conformation of the I-helix due to the addition of a water molecule between residues 263-267 within the I-helix, and these flexibilities are in agreement with our measureements (especially residues 265-266). The high shape pliability scores at residues 436-437 are due to large changes in the dihedral angles of $\Delta\psi_{436}$ and $\Delta\phi_{437}$, in addition to the C_α atom

measurements for our S^3 method showing that these residues do not closely superimpose. However, the region containing these highly shape pliable residues (including residues 436 and 437) is not mentioned in the P450BM-3 structure papers referenced in the PDB [35, 91] for 1JPZ or 1BU7.

3.3.7 Protein 7: Adenylate Kinase

Adenlylate kinases are abundant and functionally important in providing energy for the numerous cellular reactions involving ATP, ADP or AMP [72]. The specificity of the enzyme's function is accomplished by two domains closing tightly over their bound substrates. The binding domains include AMPbd, the AMP-binding domain (residues 30-59), and the INSERT domain (residues 122-159) which binds ATP [71]. The core of the protein consists of five strands forming a β-sheet, surrounded by α-helices.

We compare PDB coordinates 1AKE [71] adenylate kinase from *Escherichia coli* which has been ligated with a two-substrate-mimicking inhibitor ($P^1, P^5-bis(adenosine-5'-)pentoaphospate$) to the unligated adenlylate kinase also from *Escherichia coli* (4AKE) [72]. The classic comparison of these two adenylate kinase structures demonstrates how substrate binding leads to large domain movements. Our mobility measurements confirm the obvious large domain movement, and highlight movement in other much smaller regions as well. Following AMPbd, the α-helix occupying residues 60-73 shows a moderate amount of mobility, and we measure even greater mobility for residues 10-13 and 175-177. The shape pliability scores and mobility scores more accurately reflect the flexible residues described in [72] when we look at scores ≥ 65 (*i.e.* those residues colored orange and red in Figure 3.11) instead of 50. In Figure 3.12 we can see how the DDMP scores are high for much of the protein because so much is mobile, but the 2 largest regions of high DDMP scores occur at substrate binding domains ADPbd and INSERT.

3.3.8 Protein 8: Calmodulin

Calmodulin is a highly promiscuous protein and its structural flexibility allows it to bind to over 300 different target proteins in the cell [105]. The ubiquitious 148-residue protein binds

Figure 3.9: **Protein 6 Cytochrome P450BM-3** (A) unbound (1BU7:B) and (B) bound to substrate N_palmitoylglycine (1JPZ:B) colored by shape pliability; (C) 1JPZ:B in bold and 1BU7:B transparent, superimposed (using Pymol) and colored by mobility.

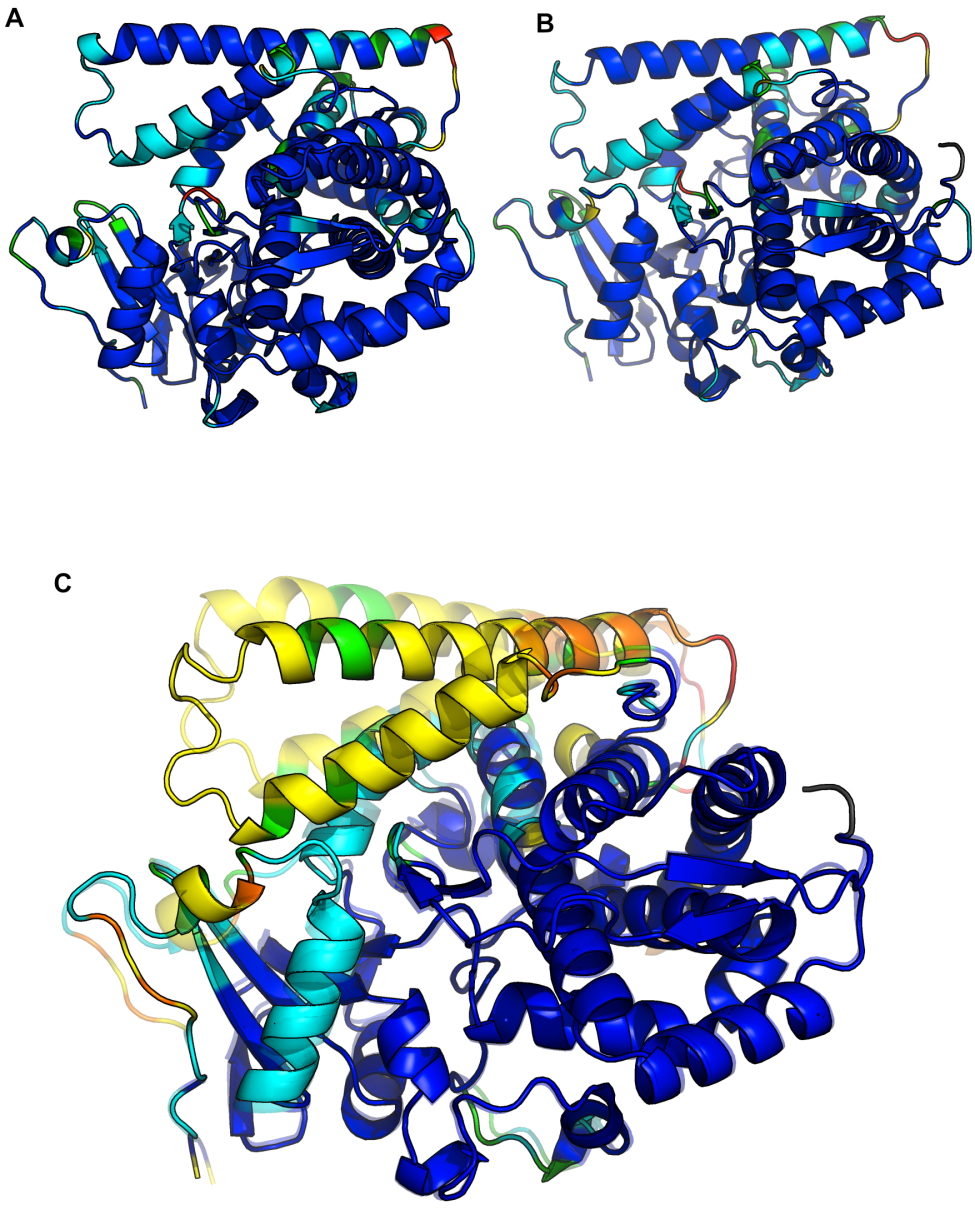


Figure 3.10: **Cytochrome P450BM-3** Sequences with secondary structure for 1BU7:B and 1JPZ:A above; shading around sequence letters indicate shape pliability scores (top) and mobility scores (bottom).

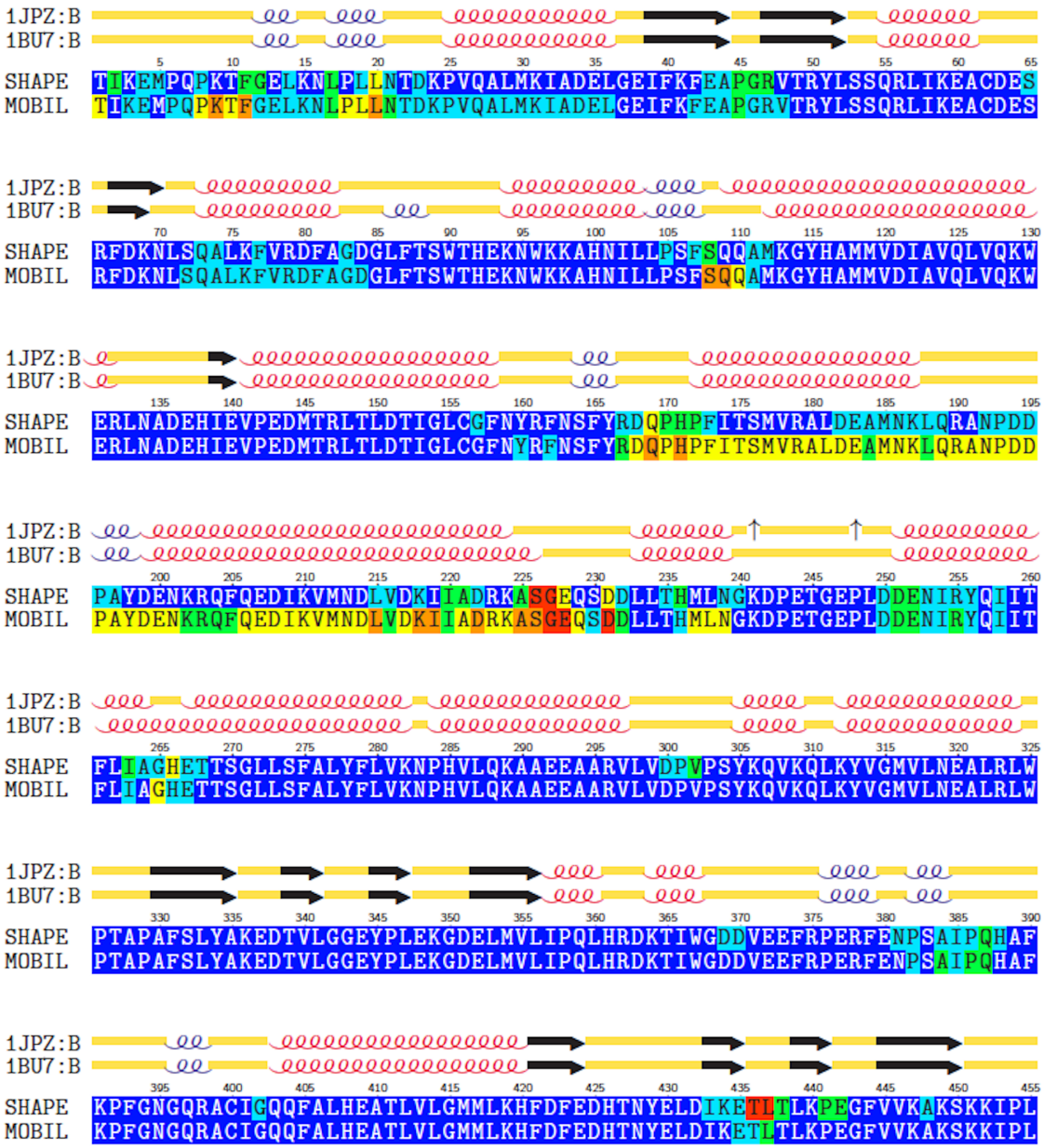


Figure 3.11: **Protein 7 Adenylate Kinase** Sequences with secondary structure for 1AKE:A and 4AKE:A above; shading around sequence letters indicate shape pliability scores (top) and mobility scores (bottom).

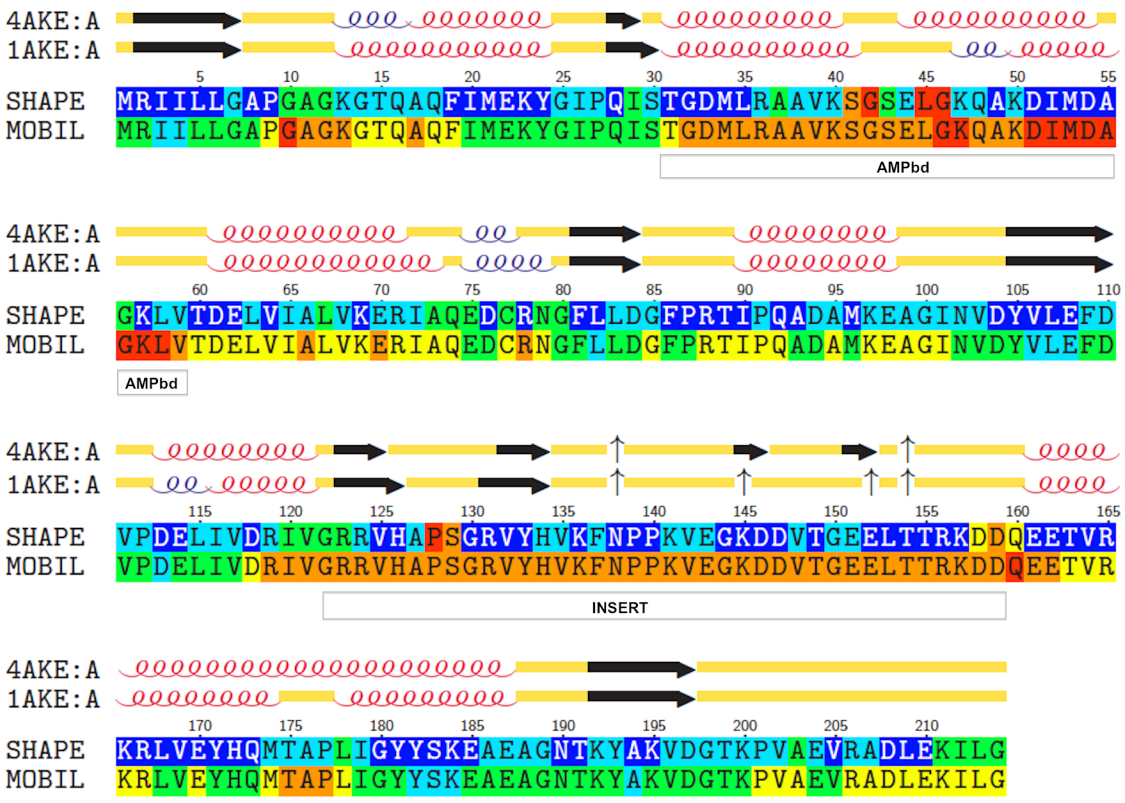
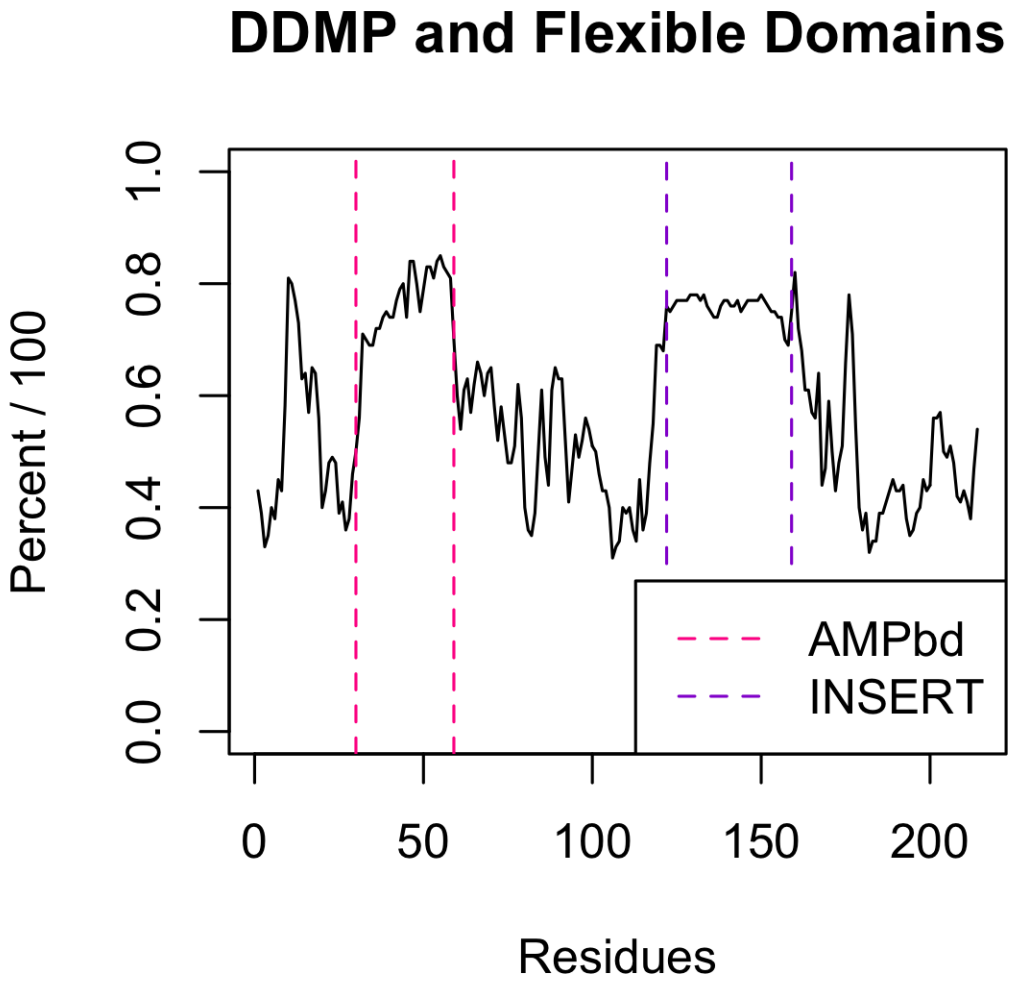


Figure 3.12: **Protein 7 Adenylate Kinase** DDMP measurements for 1AKE:A and 4AKE:A with the flexible domains indicated.



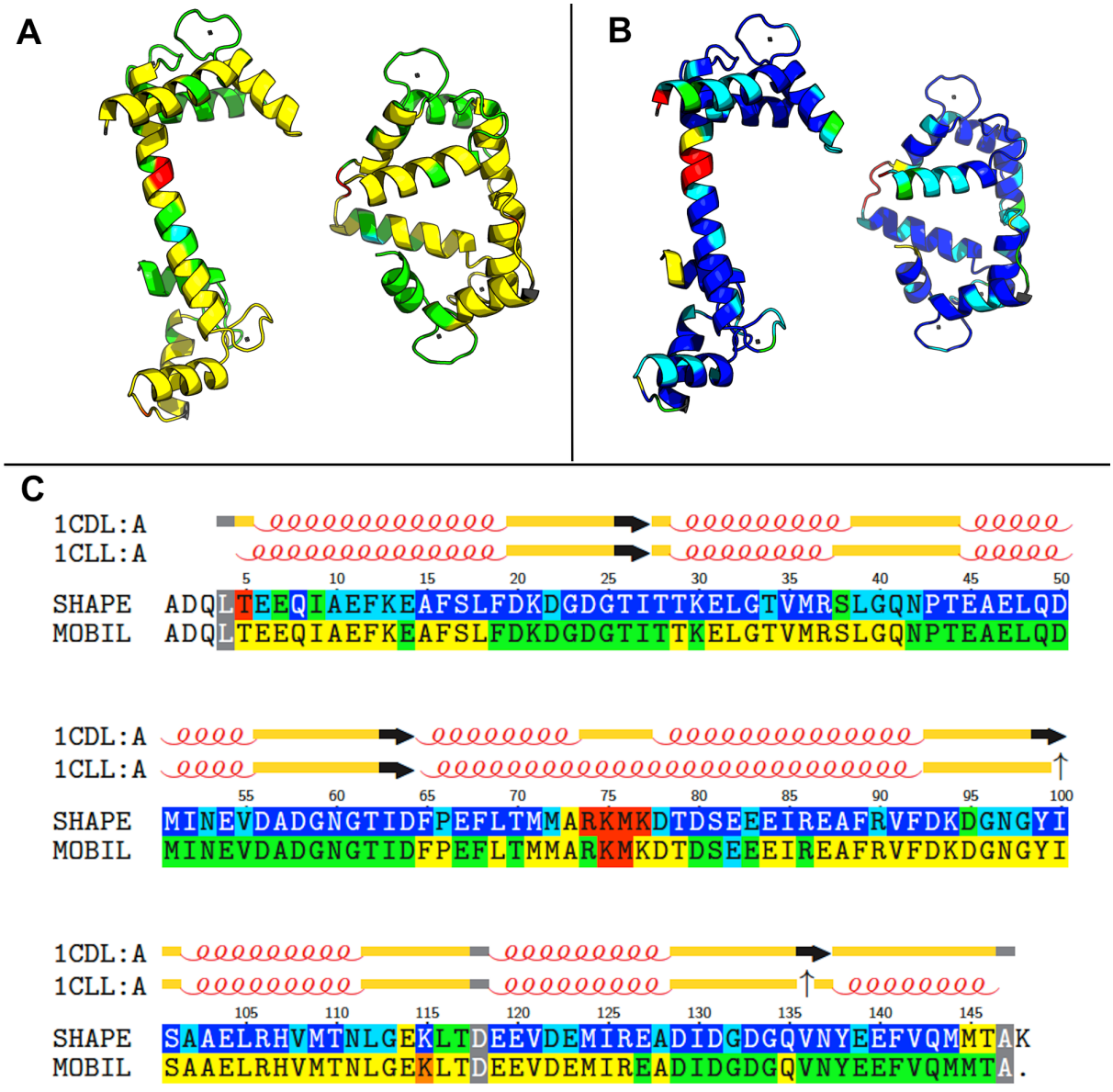
calcium and its involvement in many calcium dependent signaling pathways is coincident with its regulation of the activities of a wide variety of proteins including protein kinases, calcium pumps and proteins involved in cell motility [43]. Its structure and flexibility has been highly studied, and in Chapter 2 we display its structural complexity in Figure 2.2 as an example of how different two conformations of the same protein can appear. When studying the flexibility of calmodulin by analyzing its mobility and shape pliaiblity scores, what we find fascinating is that the large structural differences between the two conformations previously displayed can be explained mainly by 4 residues with extremely high shape pliability scores.

We compare calcium bound calmodulin (1CLL) [17] with calcium bound calmodulin which is also bound to a peptide (1CDL) [68]. These structures were solved in the same laboratory so their discussion of the conformational changes between the two structures relates directly to our comparison. As we described in Chapter 2, calmodulin when bound to a peptide differs significantly from the structure without a peptide in the central α-helix, while the lobes differ only slightly. Calmodulin without a peptide is dumb-bell shaped, with similar lobes (or hands) connected by a central α-helix. Each lobe contains 3 α-helices and 2 loops binding Ca^{2+} The large conformational change in calmodulin upon peptide binding is mainly due to changes in the dihedral angles of 4 residues (73-77) in the central α-helix [17, 68]. Therefore, it is not surprising that we find very high shape pliability scores for those 4 residues. This is clearly demonstrated in Figure 3.13. We also see a couple of extra residues with high flexiblity that are not mentioned in the articles, but these are of far less importance to the large conformational change.

3.3.9 Protein 9: G-protein G_α subunit

Research in enhancing the affinity between G-Protein G_α subunit and GoLoco peptides [12] provided some of the original motivation for this entire investigation of protein flexibility. Hence, we include the comparison between the uncomplexed $G_{\alpha i1} \cdot GDP \cdot Mg^{2+}$ protein conformation (1BOF) and the complex consisting of $G_{\alpha i1} \cdot GDP$ bound to the GoLoco region of G-protein RGS14 (1KJY) [49]. The GoLoco peptide binding is described as altering 4 conformationally flexible switches (I-

Figure 3.13: **Protein 8 Calmodulin** A. Mobility of calmodulin bound to calcium and a peptide (left, 1CDL) versus calmodulin bound to calcium without a peptide (right, 1CLL). B. Same conformations as A, colored by shape pliability. C. The sequences of the 2 conformations colored by shape pliability (top) and mobility (bottom). The secondary structure is depicted on top of the sequences, and the break in the central helix at residues 74-77 coincides with the high shape pliability (colored red) of those residues.



IV) consisting of residues 177-187, 199-219, 231-242 and 111-119 respectively [49, 70]. We find flexibility in all 4 switch regions but do not match the residues exactly. In Figure 3.14, we can see that switches II and III contain residues that are mainly disordered (colored gray, with weak electron density) in 1BOF, and we don't always find high flexiblity in the residues adjacent to the disordered regions. On the other hand, residues in switch I and IV appear to be very flexible by our metrics. We also find flexibility at the N-term and C-term regions adjacent to the disordered terminii of $G_{\alpha i1} \cdot GDP \cdot Mg^{2+}$. Furthermore, we find flexibility at residues 147,149 and 150, and although these are not described as flexible, they are important contact residues in binding of the GoLoco peptide [49].

3.3.10 Protein 10: Gelsolin docked with Actin

Gelsolin binds with the protein actin, forming a complex that is involved in important cell functions such as cell movement, cytokinesis and apoptosis, but the activation of Gelsolin requires calcium. Gelsolin consists of six domains (G1-G6) which are similar in structure, and elevated calcium levels produce large shifts in the relative positioning of the six domains, which in turn allows gelsolin to bind to actin [18]. We compare the crystal structure of domains G4-G6 bound to calcium and actin (1H1V) [18] with calcium-free gelsolin (1D0N) [15]. The constituent residues for the 6 domains are given as S1-S6 in [15], and we show these in Figure 3.16, depicting S4-S6 with colored arrows. In later citations, the domains of gelsolin are referred to as G1-G6, and the labels in Figure 3.15 use this later notation because our figure is comparable to one in [86].

Conformational differences between unbound Gelsolin and Gelsolin in complex with actin [18, 86] include a positional rearrangement of G6 relative to G4 and G5. This is shown by our high mobility scores (orange and red) for G6 in Figure 3.15. Other flexibilities [18] include the structurally variable regions going into G4 and G6, namely residues 394-412 and 620-639, as well as the C-terminal region following G6.

Figure 3.14: **Protein 9 G-protein G_α Subunit** : The sequences of 1KJY:A and 1BOF:A colored by shape pliability (top) and mobility (bottom), and gray colored residues are missing coordinates in one or both of the PDB files. Magenta-colored dashed arrows below the sequences indicate residues within flexible switches (I-IV) [70].

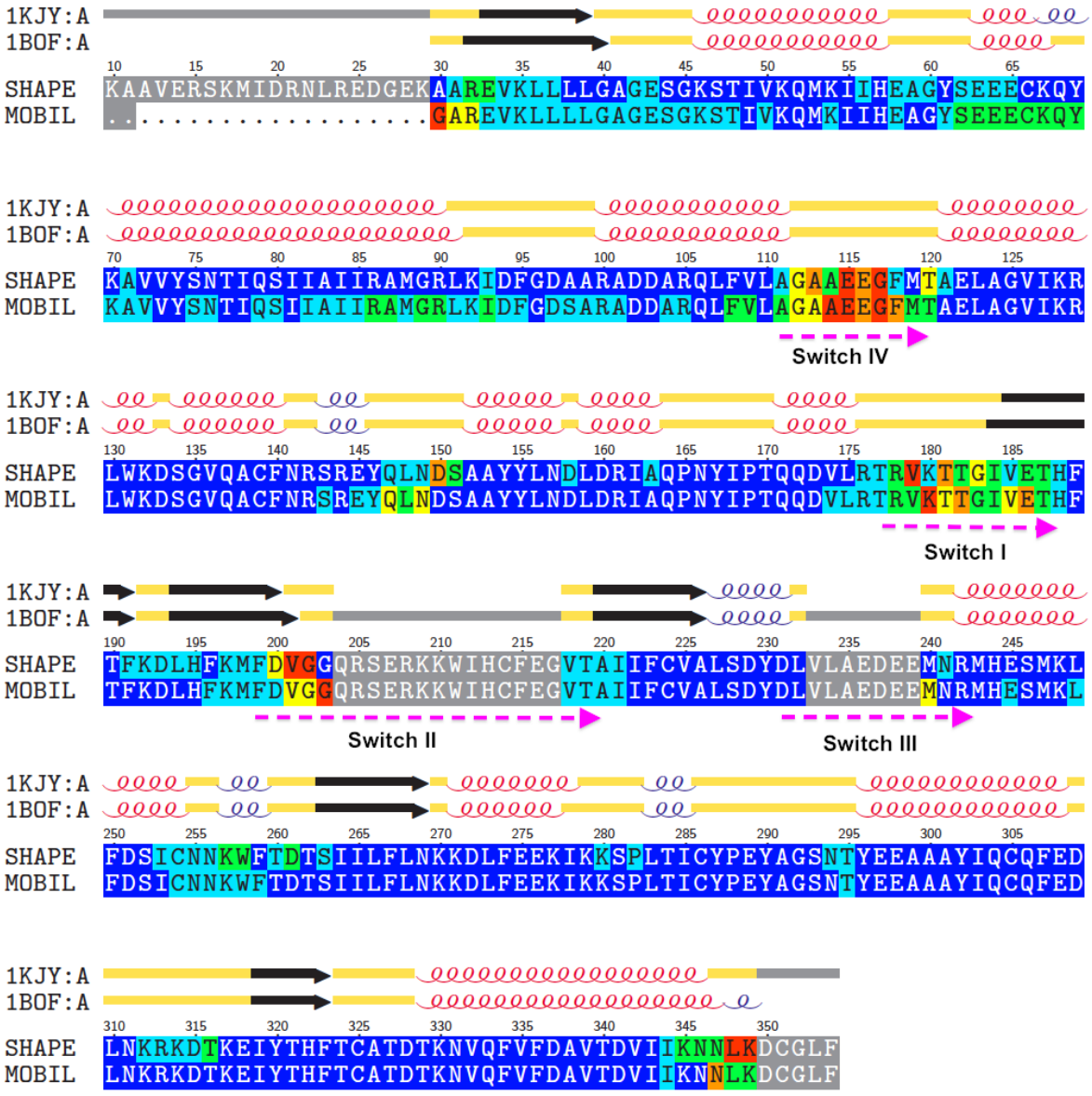


Figure 3.15: **Protein 10 Gelsolin** without Calcium and with Calcium docked to protein actin. A: apo Gelsolin (1D0N:G) and B: Gelsolin complexed with calcium and actin (1H1V:B), labeled with domains G4-G6 [86] and colored by mobility. C & D: Same conformations and domains as A and B respectively, colored by shape pliability. The main regions of shape pliability are found in loops throughout and the helix of G6. [86].

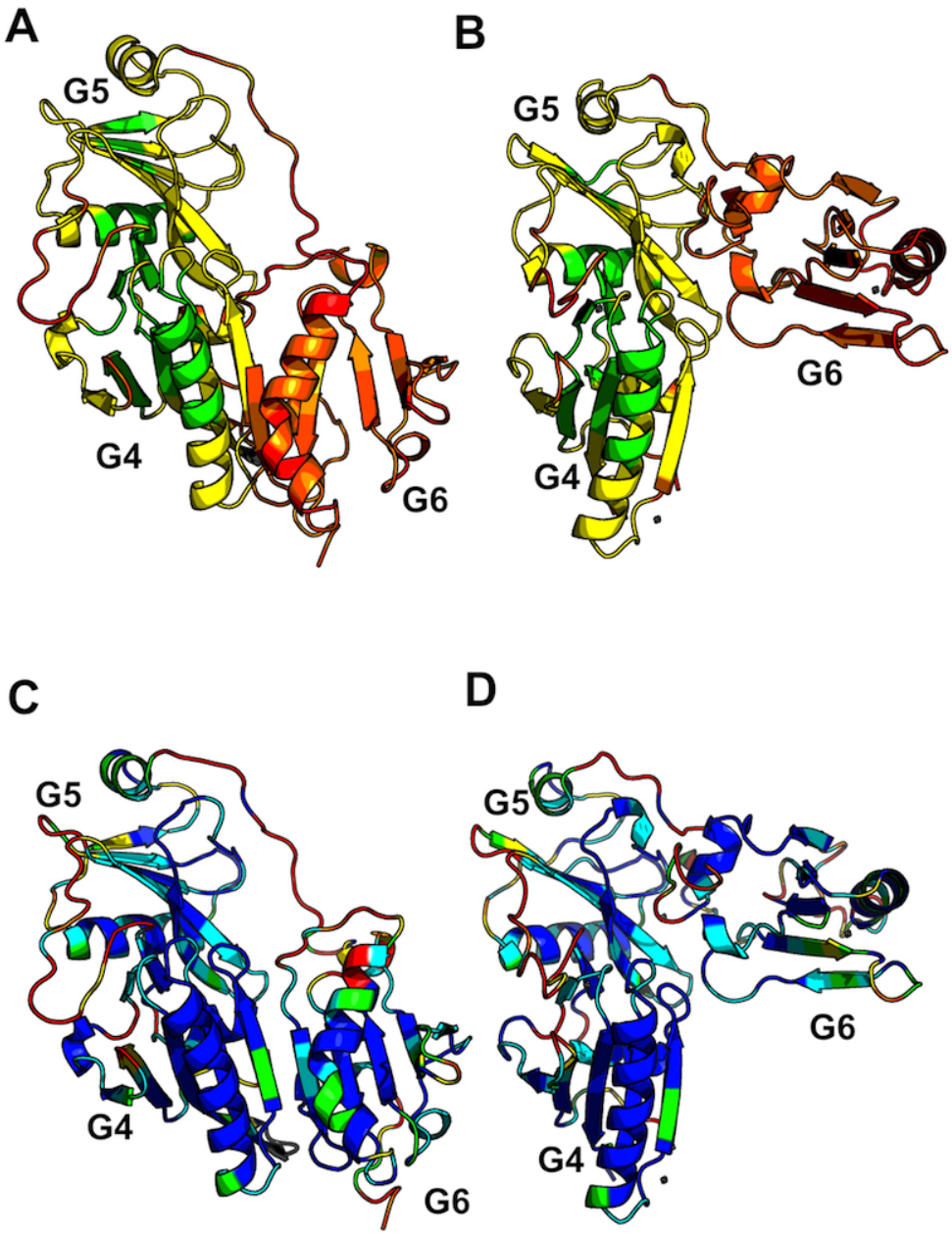
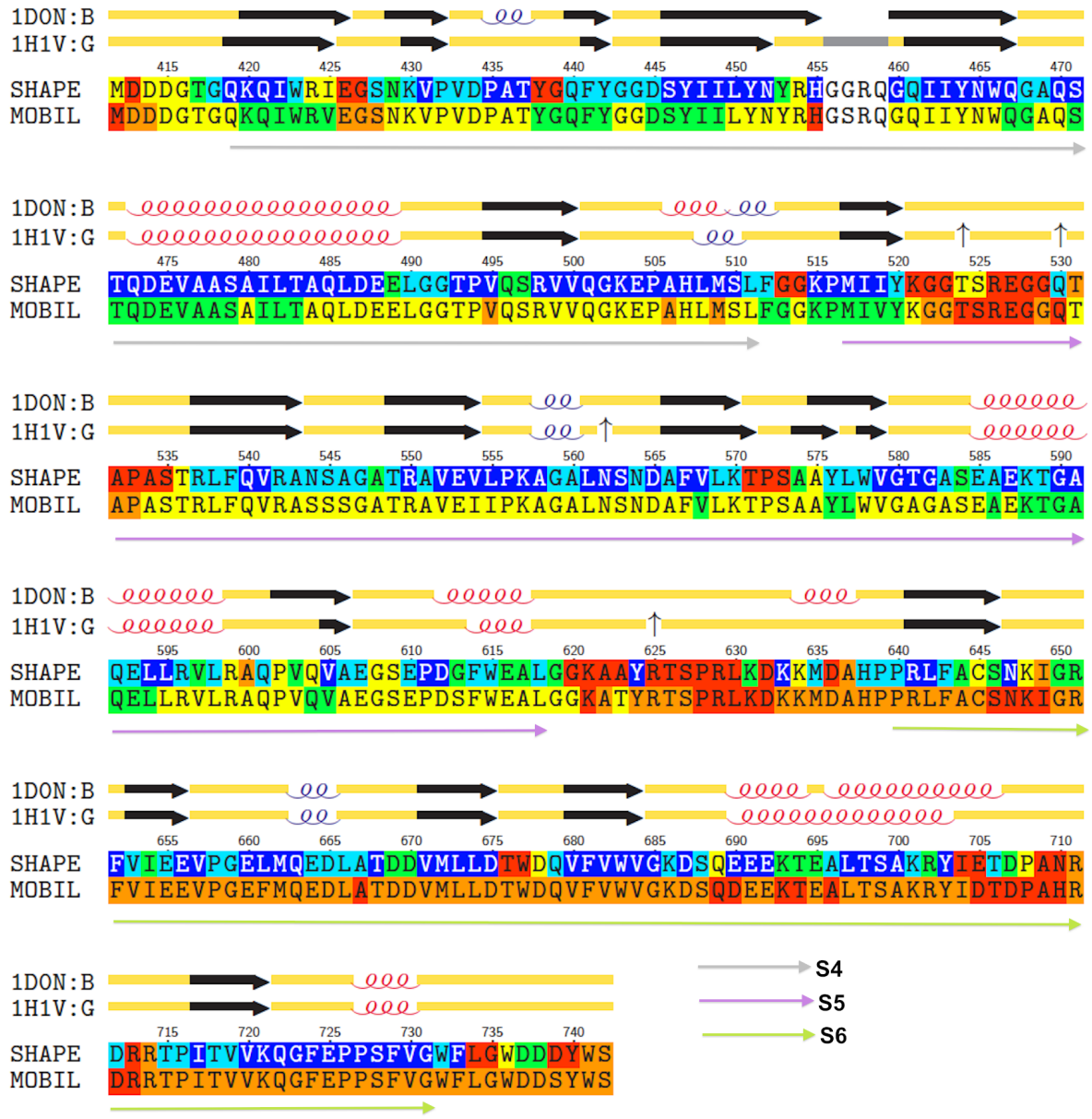


Figure 3.16: **Protein 10 Gelsolin** without Calcium and with Calcium docked to protein actin: The sequences of 1H1V:B and 1D0N:G colored by shape pliability(top) and mobility(bottom), with colored arrows below indicating the domains S4-S6, defined with their respective residue content in [15].



3.4 Discussion and Summary

For the validation using the docking benchmark, we compared *individual* protein monomers to the categories defined by complexes in the database. In the future, our measurements obtained by comparing conformational pairs may be used to predict the flexibility of individual proteins. We have learned from this validation that it may be important to select the single conformation that best represents the measured flexibility for use in a prediction algorithm. For the docking examples, this might be the unbound, rather than the bound conformation.

Mobility and shape pliability scores are continuous measures of flexiblity. The True Positive (TP) and True Negative (TN) rates use a threshold for shape pliability and mobility to be considered positive or negative. In reality, the scores are a continuum such that close scores exhibit similar flexibility regardless of whether they are above or below the cutoff. These rates were used as a way to measure how well the scoring functions compare to the literature, but they don't represent this measurement perfectly. The stories behind the comparisons of the conformations for each protein give a better account of how our measurements compare to what was reported in the literature.

Chapter 4

Applications

4.1 Introduction

In this chapter, we apply the method described in Chapter 2 to three different areas. The first two topics concern biochemistry issues that our method has allowed us to explore, and these originated from our in-depth exploration of protein conformations during the validation phase. These topics go well beyond the computer science and are included because of their relevance to drug design, in the first case, and enzyme engineering, with potential applications to biofuel production in the second. The third topic is an application to protein structure prediction that only begins to scratch the surface of the possible research in that area. Protein structure prediction and the related protein folding problem remains one of the outstanding challenges in molecular biology today and our method may aid researchers in this area by providing insights about the quality of their predictions.

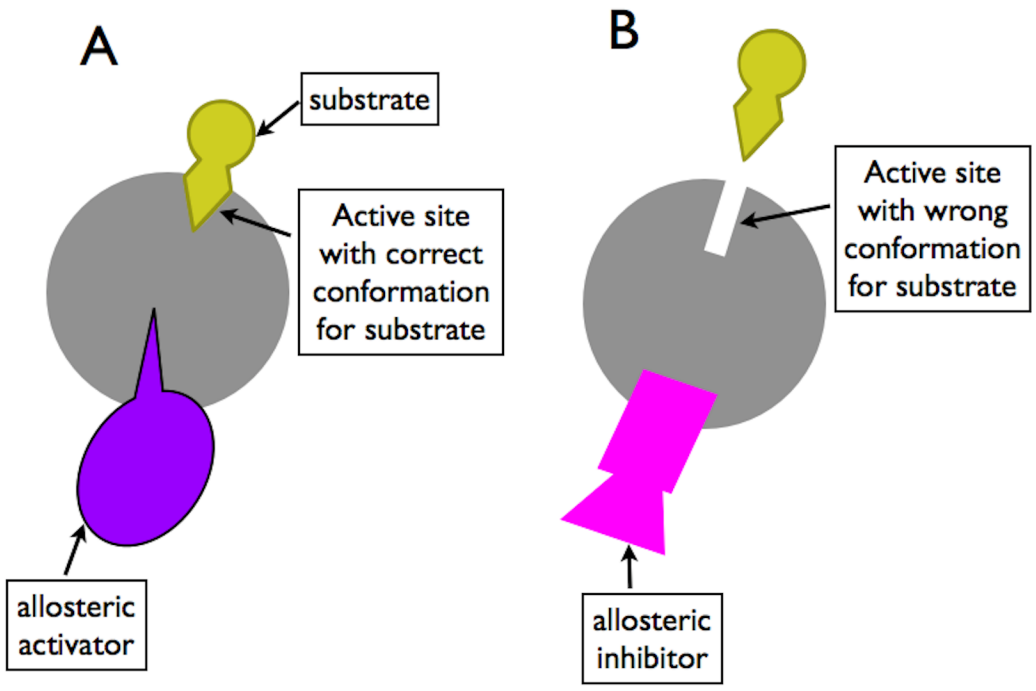
The biochemistry involved in the first application, which is related to drug design, involves allostery. Allosteric regulation is defined as "the binding of a regulatory molecule to a protein at one site that affects the function of the protein at a different site" [83], where the word *allostery* comes from the Greek words *allos* for "other" and *stereos* meaning "solid or "three dimensional". The interaction between the separated sites depends on a conformational change in the protein caused by binding to the allosteric site [1]. It follows that allosteric *inhibition* prevents a reaction from occuring such that binding to the allosteric site causes a conformational change to the active site that prevents the normal binding interaction from taking place. We include Figure 4.1, modified

from a public domain image, to illustrate these concepts.

The second application with biochemical implications is enzyme engineering. In particular, thermostable enzymes are extremely desirable in a number of industries, including biofuel production, because thermostability allows proteins to be exposed to the high temperatures found in industrial settings without denaturing [26]. Thermodynamic stability is defined by the enzymes free energy of stabilization and by its melting temperature Tm, the temperature at which 50% of the protein is unfolded. Organisms that have adapted to high temperature environments have enzymes that display higher levels of thermostability than their homologues from organisms existing in cooler environments [102]. The current hypothesis, with support from experimental data, is that enzymes from organisms surviving at the highest temperature ranges are more rigid than their homologues from species living at cooler temperatures, and that rigidity is a prerequisite for high protein thermostability [102].

The final application area involves protein structure prediction. We use data from the 9th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP9), an experiment that took place during the summer of 2010. The stated goal of CASP is "to obtain an in-depth and objective assessment of our current abilities and inabilities in the area of protein structure prediction" [30]. During the prediction season, target proteins whose coordinates have been recently determined but not yet published are released to the predictors with a time limit on predictions. At the close of the prediction season, assessors for different categories of prediction determine the relative quality of all the predictions for each target. In July 2012, during the prediction season for CASP10, one of the assessors and I discussed some of the difficulties associated with assessing the relative quality of the numerous predictions received for each target protein. The goal of this application of our methodology is to develop a tool to help the CASP assessors. This present work is informative to predictors in terms of "where they got it wrong", assuming not too much is wrong. The qualitative comparison over all predictions for a particular target is future work.

Figure 4.1: A reaction regulated by an allosteric activator is shown in A and the allosteric inhibition of the same reaction is shown in B. The conformational change to the active site by the binding of the allosteric inhibitor prevents the substrate from binding.



4.2 Allostery in HIV-1 protease

The protease protein in the human immunodeficiency virus (HIV) is crucial to the life cycle of the virus, and consequently has been a primary target for drug development in fighting the virus. [11]. HIV-1 protease is a homo-dimer (formed by two identical protein monomers) with flap regions (residues 43-59) at the top of each monomer controlling entry into the active-site pocket. The flap tips (residues 49-53) are known to be extremely flexible [11].

Our analysis of this protein examines the differences found in conformations from a study of a multi-drug resistant (MDR) HIV-1 protease derived from a patient failing inhibitor-based therapies (PDB 1RPI) with the wildtype (WT) protease (PDB 3PHV). The mutations in the MDR protease are at sequence positions 10, 25, 36, 46, 54, 62, 63, 71, 82, 84 and 90 [65].

Our validation results in Chapter 3 for this comparison showed low specificity (40 %) because we compared only one monomer of the protease dimer, whereas the comparable flexibililties described in [65] result from comparing changes to distances in the dimer. The PDB coordinates for the WT protease we used in our comparison included only the single monomer, but the MDR protease PDB file includes coordinates for both monomers. We include the additional monomer in Figure 4.2 A and B (colored gray) to help visualize the dimer perspective and specifically, the active site cavity formed by the two monomers. Figure 4.2 shows mobility measurements between the superimposed MDR protease (bold colored) and the wildtype protease (somewhat transparent) in A and the shape pliability measurements for the MDR protease in B. The sequence differences between the MDR protease (1RPI, top) and the wildtype protease (3PHV, bottom) in C. are shaded by shape pliability (top) and mobility (bottom) scores. The measurements for the highest shape pliability and mobility scores are given in Table 4.1.

According to the study of MDR HIV-1 protease by Logsdon and colleagues [65], resistance of existing therapeutics to HIV-1 protease are believed to be caused by the expansion of the active site cavity disrupting inhibitor binding affinity. The active site cavity can be seen in figures 4.2 A. and B. as the large space between the 2 monomers, just below the flexible flap tips at the top and center

Figure 4.2: Multiple Drug Resistant (MDR) HIV-1 protease (1RPI:A) is bold and superimposed on the (more transparent) wildtype (3PHV:A), colored by A. mobility and B. shape pliability, with no WT protease. Both figures include the other monomer from the MDR HIV-1 protease (1RPI:B) in gray, so that the active cavity located just below the flexible flap tips (at the top and center for both images) can be readily seen. In C. the sequences 1RPI (top) and 3PHV (bottom) are shown with shading around sequence letters indicate shape pliability scores (top) and mobility scores (bottom). Above the sequence, the secondary structure elements are drawn with symbols described previously, and subtle differences in secondary structure between the mutated and wildtype proteins are visible.

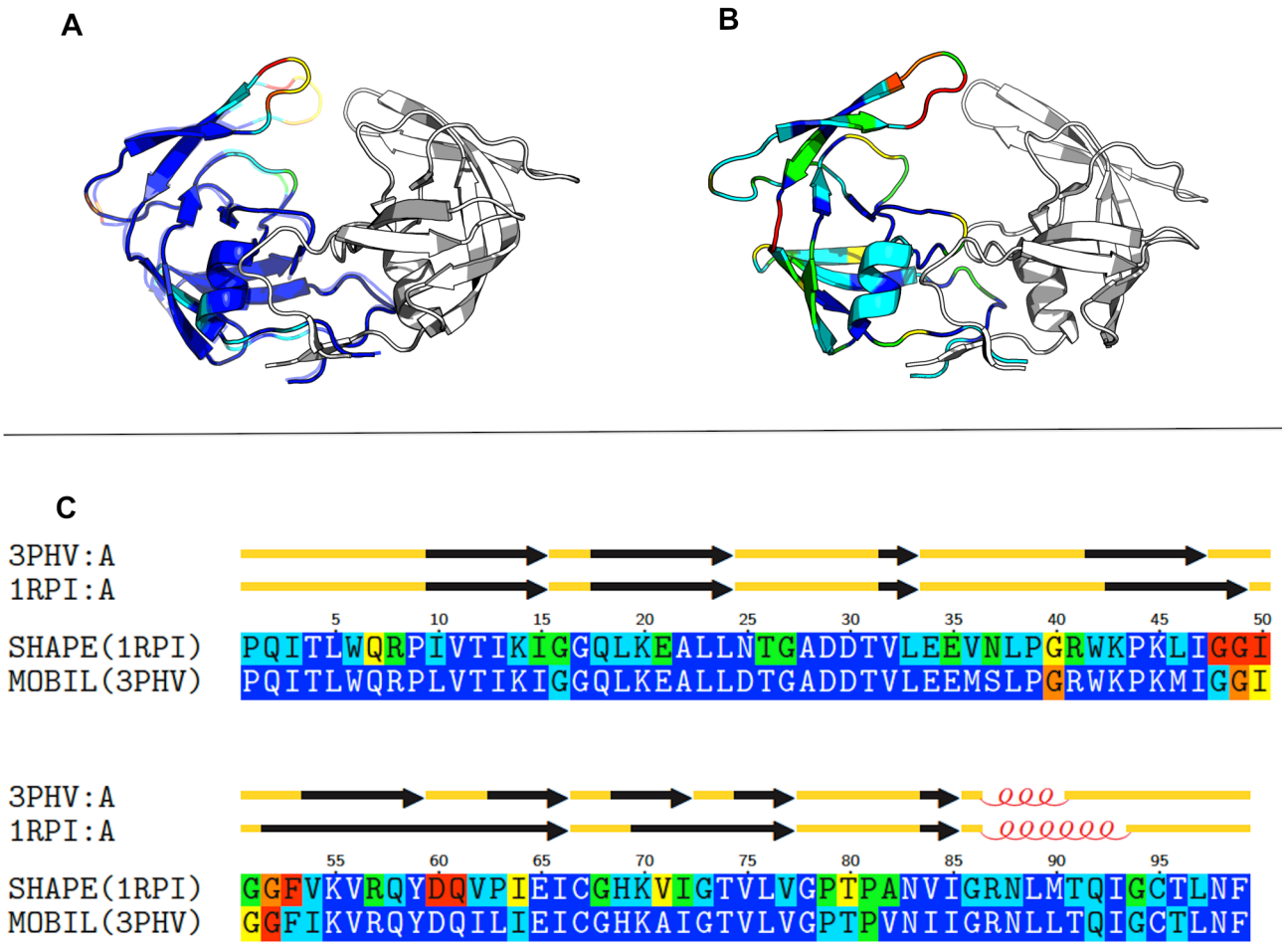


Table 4.1: HIV-1 Protease: Measurements for residues exhibiting a high degree of shape pliability or mobility comparing PDB structures 1RPI:A (MDR) versus 3PHV:A (WT).

ResID	High Shape Pliability	High Mobility	S^3 Dist ($\mathring{A}$)	Δ Phi (Deg.)	Δ Psi (Deg.)	DDMP	SS	Ave. Rel. ASA
GLN7	$\surd$		0.51	48.54	33.27	0.06	S S	0.65
GLY40	$\surd$	$\surd$	1.52	-63.65	14.76	0.73	S S	0.74
GLY48	$\surd$		1.40	-112.21	69.00	0.28	E -	0.68
GLY49	$\surd$	$\surd$	2.39	-37.78	-147.64	0.63	E S	0.39
ILE50	$\surd$	$\surd$	1.54	-130.19	35.24	0.62	T S	1.00
GLY51		$\surd$	1.24	-31.76	37.20	0.60	T S	1.00
GLY52	$\surd$	$\surd$	1.91	38.61	-73.67	0.82	E -	0.39
PHE53	$\surd$		0.90	-56.64	82.47	0.25	E -	0.76
ASP60	$\surd$		0.14	5.28	-166.76	0.14	E S	0.47
GLN61	$\surd$		0.45	-179.01	7.88	0.14	E S	0.79
ILE64	$\surd$		0.84	-64.08	14.80	0.22	E E	0.02
VAL71	$\surd$		0.25	32.81	45.59	0.06	E E	0.10
THR80	$\surd$		0.91	-58.48	-28.67	0.21	- -	0.29

of each figure. Specifically, residues mentioned that are involved in this expansion (within the first monomer of the homodimer) are 50, 80, and residues near position 7. Both Table 4.1 and Figure 4.2 show high shape pliability for all 3 residues mentioned as involved in the expansion of the active site cavity, with residue 50 (located in the flap tip) also containing a high mobility score. Additionally, a decrease in the sizes of amino acid sidechains for mutations in positions 82 and 84 changes the active site, but these are not found in our analysis because we focus only on backbone measurements. However, in the review article of [11], molecular dynamics simulations have identified flexible regions of the protease such as the fulcrum(residues 11-21), the flap tips (residues 49-53), the flap elbow (residues 38-42) and the cantilever (residues 64-74). Furthermore, they note that several of the "compensatory mutations" for restoring catalytic activity in drug resistant mutants are found in or near regions of above-average mobility. These observations essentially validate all of the remaining residues that we find to have high shape pliability or mobility scores, with the exception of residues 60 and 61, which are close to the mutations at positions 62 and 63.

Although residues 60 and 61 have very high shape pliability scores, they exhibit almost no mobility. This is because the large dihedral angle change in the ψ angle of residue 60 is compensated for by the similarly large ϕ angle change of residue 61, and thus these large dihedral angle changes do not significantly effect the overall backbone positioning. This type of motion (often referred to as a crankshaft motion [29]) is rarely highlighted in structure references because the change to the respective C_α positions is minimal. We observe high shape pliability score for these residues because the change to the backbone involves a peptide plane shift, and in turn, the position of the carbonyl oxygen rotates by 180 degrees.

A proposed mechanism for drug design using allosteric inhibitors instead of inhibitors targeted to the active site is decribed in [41] . In [58], the crystal structure for a fully open MDR protease conformation is believed to be stabilized in the open position by crystal contacts for each flap tip buried between the fulcrum and elbow regions of a symmetry related neighber, with residues 39, 41, 60, 61 and 72 enclosing the tip. As in [41], [58] notes that this may provide experimental evidence for a proposed allosteric inhibition. The allosteric inhibitoin could prevent the substrate

from entering the active site cavity by targeting the flexibility that allows the protease to change conformations between fully opened (to the active site cavity), partially closed, and closed. Based on our measurements and the proposed allosteric inhibition, the planar flexibility we observe at residues 60 and 61 perhaps plays a role in the flexibility that enables conformational change, and provides further evidence that allosteric inhibition at this location could potentially be useful in resolving this particular set of resistance issues for drugs targeting HIV-1 protease.

4.3 Analysis of Mesophilic and Thermophilic Adenlylate Kinase

The large-scale motion of Adenlyate Kinase (Adk) has been studied in a number of different species using a variety of experimental and computational methods. The dynamics and catalysis of Adk have been compared in mesophilic *E. coli* and thermophilic *Aquifex aeolicus* [108, 39] because these homologs have very similar structures but different rates of catalysis. Thermophiles exist in higher temperature environments than mesophiles, and thus the proteins of hyperthermophilic and thermophilic species are inherently more thermostable than their mesophilic homologs. The work of Bae and colleagues [3] compares thermal stabilities of modifed Adenylate Kinase (Adk) sequences, with mutations constructed by combining portions of different sequences from a mesophile and thermophile. The selection of residues to mutate is based on flexibility characteristics, and their ability to design more stable variants of the mesophilic Adk that retained their catalytic activity is especially exciting from a protein engineering viewpoint. In this section, we examine the hypothesis that the more thermostable Adenylate Kinase (thermoAdk) is also more rigid than its mesophilic counterpart (mesoAdk).

We study the flexibilities of Adk from mesophilic *E. coli* and thermophilic *Aquifex aeolicus* by examining differences in mesophilic coordinate files for the apo (4AKE) and bound (1AKE) conformations compared to the differences found in thermophilic conformations; apo (2RH5) and bound (2RGX). In both cases, the apo (unligated) form is the open state, whereas the more closed structure is bound to an inhibitor which mimics the binding of both AMP and ATP at different sites. More specifically, the structure of Adk is comprised of 3 domains; AMP, LID and CORE.

The AMP and LID are widely distanced in the apo structures but are in much closer proximity in the bound structures. The CORE domain consists of all residues not in AMP or LID, and is generally more static than the other 2 domains. The large-scale motion of AMP and LID for both species is apparent in the higher average mobility scores shown in Table 4.2 for AMP and LID domains versus the CORE domain. The more local shape pliability score averages do not vary much between domains. The images in Figure 4.3 C and D show the orange and red highlighted domains (indicative of the high mobility scores) in very different positions on the left (open) and right (closed) sides for mesoAdk and thermoAdk respectively. The same images also capture the similarities for the gross movements between mesoAdk and thermoAdk.

Table 4.2: Adenylate Kinase: meso and thermo mobility and shape Pliability mean, median and minimum/maximum scores per domain. The 2 highly mobile domains AMP and LID show much higher mean and median mobility scores than the less mobile CORE. There is not very much variability in the shape pliability scores of the 3 domains.

	MesoAdk			**ThermoAdk**		
Mobility	**Mean**	**Median**	**Min/Max**	**Mean**	**Median**	**Min/Max**
AMP	76.8	76.17	50.0/99.7	74.6	74.0	55.5/85.6
LID	75.8	76.17	69.2/78.0	73.8	76.7	48.5/ 82.2
CORE	50.2	48.1	31.3/81.8	47.4	43.6	27.7/89.1
Shape Pliability	**Mean**	**Median**	**Min/Max**	**Mean**	**Median**	**Min/Max**
AMP	30.1	22.0	4.5/100.0	21.6	24.4	7.5/43.3
LID	23.5	16.3	4.8/82.6	21.2	20.25	4.3/40.0
CORE	22.3	21.35	3.1/62.1	18.9	14.0	1.6/100.0

Residues with relatively high mobility scores are typically located in regions of well-documented motion in other proteins. This is clearly the case for the motion of the AMP and LID domains, where highly mobile residues cover these domains in both species as seen in Table 4.3. Overall mobility similarities between the Adk structures of the 2 different species are apparent in Figure 4.3. The graphs in Figure 4.3 (A and B) demonstrate the same regions of high mobility scores in the AMP and LID domains, and the p-loop. In the structural images in Figure 4.3 (C and D) , the extremely mobile LID and AMP regions can be seen as completely separate in the apo

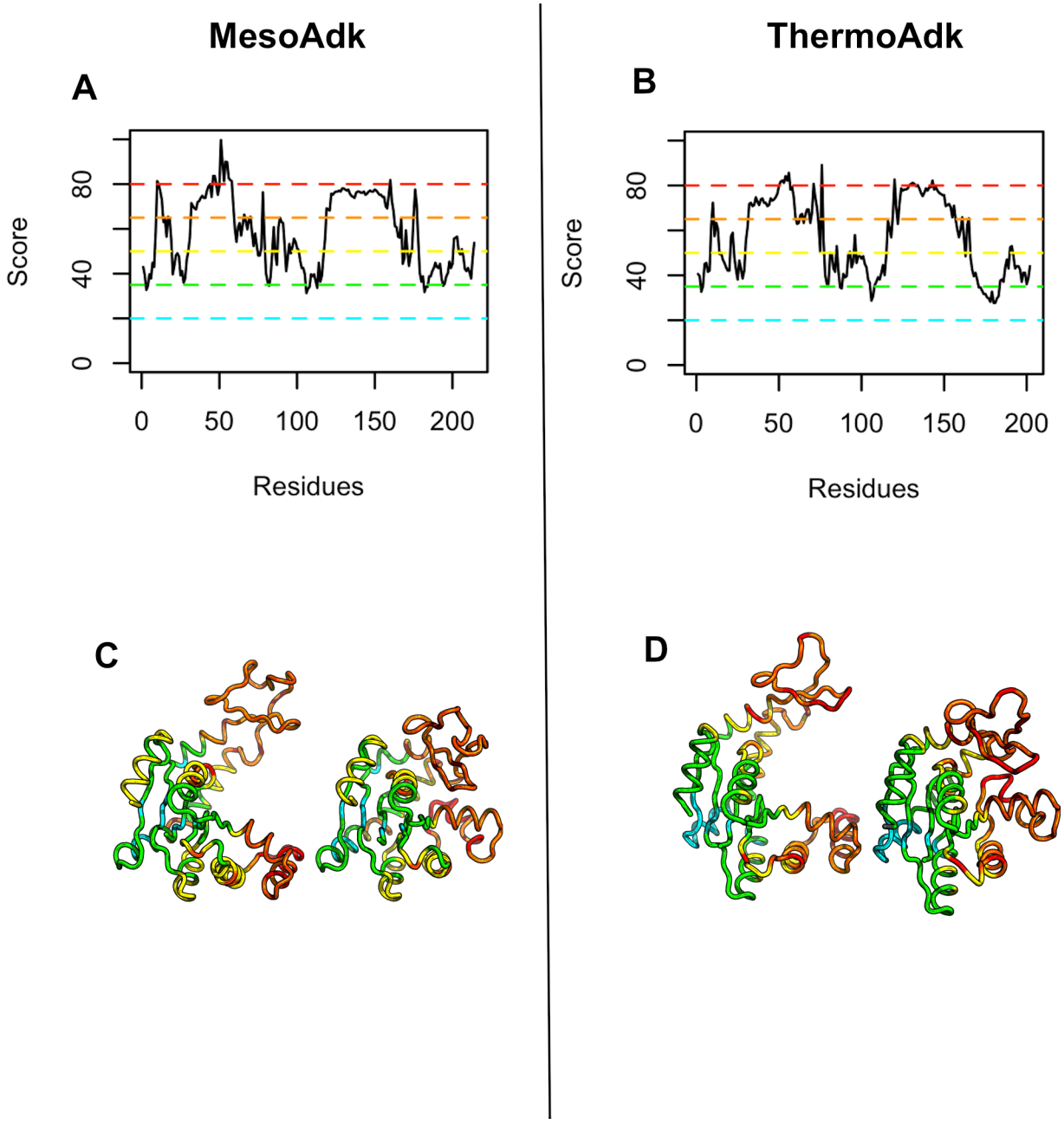
structures and then much closer in distance in the bound structures. However, because our mobility scores represent differences in measurements of intramolecular distances or superpositioning distances found by comparing x-ray crystal stuctures, high mobility scores are not restricted to large domain motions, and are also found in smaller regions such as certain loop regions between secondary structure elements. For example, the P-loop is a binding motif (GXXGXGK) found in both meso and thermo Adk's where ATP phosphates are partially bound [92]. Table4.3 shows higher scoring mobility residues within the P-loop of mesoAdk. The P-loop in thermoAdk contains one additional proline residue, which may be the reason it is both less mobile and less shape pliable in thermoAdk than MesoAdk. Likewise, hinge 8 also shows less flexibility in thermoADK than mesoADK. α-helix-7 splits into two halves to form a loop in the middle in the bound form of mesoAdk, but not thermoAdk. ThermoAdk, on the other hand, shows higher mobility in the residues within hinge 4, which is the area described by Müller and colleagues as one of the counterweights that change flexibility in the apo versus bound forms of mesoAdk [72]. The differences between meso and thermo Adk's are even more striking in the shape pliabilities scores, and we discuss this next.

Table 4.3: Comparing highly mobile residues in mesoAdk and thermoAdk to reported features in the literature. H4 and H8 are two of the eight hinge regions that have been compared between mesoAdk and thermoAdk in [39].

MesoAdk		**ThermoAdk**		
High Mobility Residues	Max Mobility Residue (Score)	High Mobility Residues	Max Mobility Residue (Score)	**Feature** (Residues)
10-13	G10 (81.3)	10	G10 (72.3)	P-loop (7-13) [92]
32-59	D51 (99.7)	32-60	G56 (85.6)	AMP: (30-59) [92]
78	R78 (76.3)	70-73,76	G76 (89.0)	H4: (M:71-81, T:71-78) [39]
118-162	Q160 (81.8)	119-156	R120(82.7)	LID: (M:122-159,T:123-152) [92, 39]
176-177	A176 (77.6)			H8: (M:172-178, T:166-172) [39]

Shape pliability includes but is not limited to hinges, however, the hinge regions have been

Figure 4.3: ADK Mobility. A and B plot the mobility scores for the residues of MesoAdk and ThermoAdk respectively. C: apo (left) versus bound (right) mesoAdk, colored by mobility scores, and likewise D: apo(left) versus bound (right) thermoAdk.



extensively studied in comparing meso and thermo Adk. In [39] 8 hinges and a kink region are described for MesoAdk and compared to similar regions found in ThermoAdk. Our method for scoring shape pliability is based on ϕ and ψ dihedral angle differences and C_α distances computed by our S^3 algorithm, whereas [39] compares a completely different set of angles. Table 4.4 compares the high shape pliability scores of residues using our scoring technique with those described in [39], and highlights (in bold) the highest shape pliability scores. [39] mentions that in NMR experiments, higher order parameters (*i.e.* atoms with more rigidity) were found for thermoAdk hinges than in mesoAdk. Furthermore, four hinges (2, 3, 4 and 7) contain the amino acid Proline in thermo but not meso, and aromatic ring stacking interactions near hinges 4 and 5 in thermo may also rigidify. We find the highest shape pliability scores for mesoAdk in H2 and H7, as well as in loop between β_5 and β_6 within the LID domain. In contrast, the single shape pliability score spike in thermoAdk occurs in hinge 4, although this is one of the hinges containing a PRO that is believed to rigidify it in comparison to mesoAdk without the PRO. Interestingly, this particular hinge region is a counterweight described by [72], and also has high mobility scores. These findings are illustrated in Figure 4.4 showing graphs A and B of the shape pliability scores of mesoAdk and thermoAdk respectively, with the high peaks circled and numbered, and then displayed in the partial structure images C, D, and E of the AMP, CORE and LID domains respectively.

The general conclusions of this section comparing flexibilities in mesoAdk and thermoAdk are the following:

(1) The mobilities of the Adk's from the 2 species are quite similar, as expected.

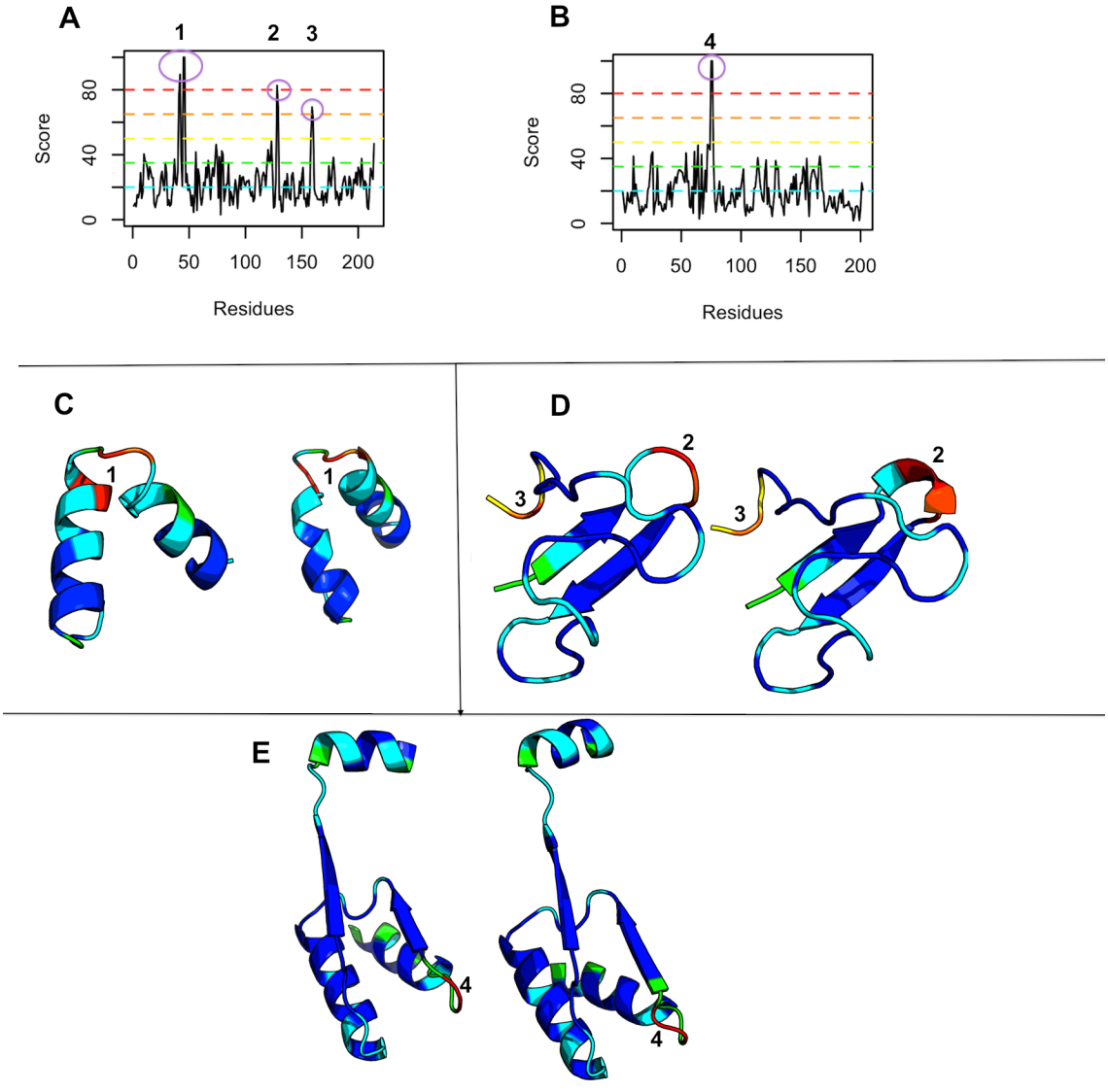
(2) The eight hinges described by [39] are all found in or near regions of relatively high-scoring shape pliability residues by using a lower cutoff value (35.0 instead of 50.0) for shape pliability residues of interest. This is consistent with the values of the angular differences for hinges in [39]. However, in [39], there is no comparable discussion of the *relative* flexibilities of the hinges that we highlight in the analysis of our measurements.

(3) There are more very high shape pliability regions in mesoAdk (3) than in thermoAdk (1).

Table 4.4: Adenylate Kinase: Shape Pliability versus Hinges for MesoAdk and ThermoAdk conformations.

MesoAdk			**ThermoAdk**		
High Shape Pliability Residues	Max Shape Pliability (Score)	Hinge Residues [39]	High Shape Pliability Residues	Max Shape Pliability (Score)	Hinge Residues [39]
10-12	G10 (40.4)		10	G10 (41.4)	
			25-26	F26 (43.9)	
29	I29 (35.0)	H1:29-31	30	S30 (35.5)	H1:28-31
37	A37(37.2)				
41-43; 45-46	**G42 (89.5) L45,G46 (100.0)**	H2:42-50	54	E54(43.3)	H2:49-51
56	G56 (41.6)	H3:59-61	61	D61 (44.3)	H3:59-62
			64	I64 (48.2)	
67	L67 (36.4)		67	L67 (42.5)	
73-75,77 79-80	Q74 (46.2) D79 (42.3)	H4:71-81	72-77	**H75,G76 (100.0)**	H4:71-78
		H5:110	114	D114 (40.4)	H5:111-114
		Kink:115-116			
120-123	R123 (48.3)	H6:120	121	L121 (39.0)	H6:120-122
128-129	**P128 (82.7)**		129	E129 (38.6)	
158-160	**D159 (69.4)**	H7:157-161	155	P155 (40.0)	H7:151-155
178	L178 (38.4)	H8:172-178	165-166	R166 (41.4)	H8:166-172
203	A203 (37.7)				
214	G214 (46.8)				

Figure 4.4: ADK Shape Pliability. A and B show the Shape Pliability scores for mesoAdk and thermoAdk respectively. C: mesoAdk AMP domains (residues 30-59) with the highest scoring peak (1) from A, D shows LID domains (residues 122-159) from mesoAdk structures with the remaining high shape pliability peaks (2) and (3) from A. E: CORE domain (residues 60-122) of thermoAdk with the highest shape pliability peak(4) from B.



Perhaps this provides evidence that thermoAdk is more rigid than mesoAdk, as theory would suggest.

(4) The locations of the high shape pliability regions in the enzymes of the two different species are also quite different. The high shape pliability regions in mesoAdk suggest locations where mutations could be applied to rigidify the protein, but there is no evidence that we are aware of in the research to date suggesting mutations at these locations or their effect on catalysis.

4.4 Protein Structure Prediction Error Analysis

The CASP experiment is a biennial event, with the prediction season occurring during the summer of every other year. The website [30] reports on all targets with predictions, ranks and the native structure PDB code for each. The ranking is done primarily using a measure called the GlobalDistanceTest_TotalScore (GDT_TS), which is a measure of the similarity between two protein structures with identical sequences. It is defined as $(GDT\_P1+GDT\_P2+GDT\_P4+GDT\_P8)/4$ where each GTD_Pn denotes the percent of residues under a distance cutoff $\leq n\mathring{A}$. For each CASP experiment, the GDT_TS is computed for each prediction compared to the coordinates of the native structure published in the PDB.

Our flexibility measurements and the shape pliability and mobility scores are based on comparisons between different conformations of a protein with the same or very similar sequences. We also measure differences and similarities between the two structures, and describe these differences in terms of flexibility. But the measurements could also be representative of the errors found in predicting the structure. Mobility is representative of the global error in positioning the parts of the protein relative to the remainder, and could be thought of the error in the global fold. Shape pliability represents changes to individual, local elements of the protein, and could be thought of as the error in the local (secondary structure elements) shape.

We applied our scoring functions to a very small number of predictions to demonstrate the feasibility of using our method in error analysis. Here, we highlight our finding with a single target, and two predictions of very different quality. The native structure is an α/β protein shown in Figure 4.5 A and is colored by spectrum over the various secondary structure elements, as is customary when comparing CASP structure predictions.

Figure 4.5 B and E are colored using the same scheme, which highlights the fact that the predicted strucure of B is very close to the native structure and E is not. In fact, Figure 4.5 B, C, and D all represent the same prediction submitted by the group "Foldit", a group comprised of expert players at using the online game of the same name [20]. This prediction had the highest

GDT_TS score (70.48) over all predictions for this target. The biggest error in the prediction is the helix colored red in both Figures C and D, and does not appear at all in the native structure in A. The other areas colored yellow and orange in these two figures represent more subtle errors in the prediction and locate residues that have smaller differences in the backbone from the native structure.

In stark contrast, the prediction shown in the bottom row, Figures 4.5 E, F and G had rank 129, with GDT_TS score 16.43. The fold has little resemblance to the native fold, and the mobility scores in Figure 4.5 G. are all at the high end of the mobility spectrum. The shape pliability scores reprented in F show some rigidity, or smaller differences, because by predicting almost all helices, a few of them happened to be close to the correct postitions. The scores for this prediction are less useful in identifying which residues have errors in their backbone positioning because there is so little signal for the S^3 alignment and DDMP distance comparisons.

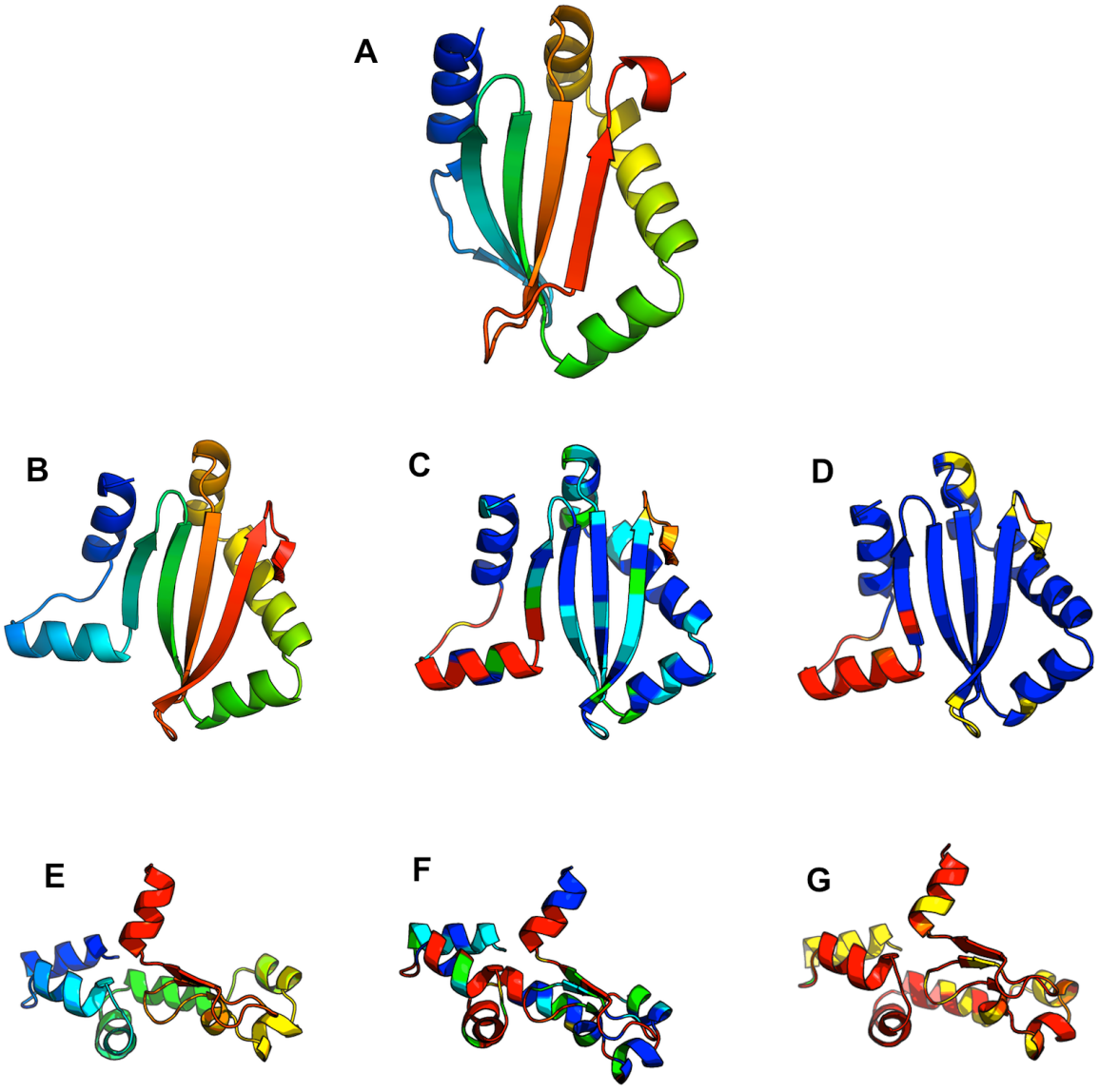
To summarize, this application of our method is particularly effective in highlighting specific errors for predictions with structures close to that of the target protein, although not as illuminating in cases where the predicted structures do not resemble the target protein. The next phase is to combine the scores for mobility and shape pliabiliy in some way to produce a single score, and compare our ranking of predictions to those given by GDT_TS.

4.5 Discussion and Summary

In the first two applications of our method, shape pliability scores pointed out flexibilities that were not highlighted in the literature, but may have biochemical significance.

In the third application, the visualization of the differences for good structure predictions can give insights into where the prediction is more or less successful, and pehaps lead to a better understanding of the relative merits of different structure predictions and structure prediction algorithms.

Figure 4.5: CASP9 target T0581 showing the native structure in A compared to the prediction with the highest GDT_TS score for that target (B, C, D) and a prediction without much accuracy (E, F, G). Specifically, T0581_D1 is shown as: A: the native structure for T0581 with secondary structure elements colored by specturm, B: spectrum colored T0581 prediction 170_1, C: T0581 prediction 170_1, colored by shape pliability scores, D: T0581 prediction 170_1 colored by mobility scores, E: Prediction 20_1, spectrum colored, F: Prediction 20_ colored by Shape Pliability, and G: Prediction 20_1 colored by mobility.



Chapter 5

Analysis of Backbone Flexibility

5.1 Introduction

The previous chapters describe two characterizations of residue-level protein backbone flexibility we find implicit in protein conformational differences. These characterizations involve global mobility of the residues with respect to the whole structure, and shape pliability of the local backbone structure. Now, we ask the question: "Are there attributes of the structure and sequence of a *single* protein conformation that correlate to the flexibilities described in the previous chapters?". The goal of this analysis is to determine which, if any, of these investigated attributes could be used to predict backbone flexibility using only local information. This type of information would be useful in the context of protein design or other protein modeling algorithms where locality of flexible residues is important. We assess the hypothesis that atomic crowding or lack-of-crowding is associated with backbone flexibility. We investigate which specific chemical or energetic interactions correlate to flexibility, and use the scores from our mobility and shape pliability scoring functions to examine these questions.

In chapter 3, we used the protein-protein docking benchmark [27, 42] to validate our method of scoring mobility and shape pliability by comparing the interface residues of bound and unbound monomers. In this chapter, we use the same benchmark proteins but focus on the scores for all residues in the comparisons, not just those residues found on the interface. This provides a collection of over 17,000 residues from the comparisons of 81 protein monomers. Unless otherwise specified, all statistics apply to the collection of residues as a whole, and not to the individual proteins or

docking categories described in §3.2.

We investigate the correlation of energetic terms to flexibility using the full atom energy or scoring function from the Rosetta molecular modeling suite of programs [61]. The Rosetta energy terms are empirically derived from analyzing observed geometries of protein structures from the PDB [19], and historically, were developed to predict protein structure for an *ab initio* (meaning from "first principles") protein structure prediction algorithm. The all-atom scoring function is comprised of weighted individual terms that are primarily knowledge-based potentials, with their underlying function based on Newtonian physics-based energy terms similar to those found in other force fields such as CHARMM [13] or Amber [79]. We use the Rosetta energy scoring functions for this analysis because our findings are mainly targeted to researchers in the areas of protein and interface design and protein-protein docking, many of whom use Rosetta for their research.

5.2 Methods

Scoring functions for shape pliability and mobility were described in detail in §2.2.5, and we use the same parameters for scoring as previously described.

Spearman rank correlation is used to estimate a rank-based measure of association between two variables without any assumption regarding the underlying normality of the distributions of the variables [54, 10]. The implementation of this correlation test in [96] computes the corresponding p-values using an asymptotic t approximation. The correlaton coefficient value is given as ρ.

Secondary structure classifications per residue were obtained by using the "Dictionary of protein Secondary Structure Program" or DSSP [46]. The DSSP program calculates hydrogen bonds and classifies residues by the hydrogen bonding patterns found. The residue classification of a Helix, (Beta) Sheet or Loop is based on the following DSSP characterizations of the hydrogen bonding patterns:

- **Helix:** The hydrogen bonding of residues in the sequence that are 3, 4, and 5 residues apart are given the DSSP symbols G, H and I for participation in a 3_{10}, α or π helix,

respectively.

- **Sheet:** a single pair $\beta-sheet$ hydrogen bonding forms a beta bridge (B) and extended strand hydrogen bonding is symbolized by E.

- **Loop:** All characterizations of residues not included in Helix or Sheet classifications. These include DSSP symbols T, S, and blank for the hydrogen bonded (T)urn, (S) for a bend region of high curvature (non hydrogen bonded) and otherwise unclassified as blank or space.

By modifying the *Rosetta++* suite of programs [61], we obtained neighboring atom counts used in the structural analysis, as follows:

- **All atom neighbors:** For each residue we compute the all atom count by comparing the distance from each residue atom to to all other atoms in the structure, and counting the atoms in all other residues within a cutoff distance of $5\mathring{A}$. We also add to that the number of hetero atoms from the PDB within the cutoff distance. Hetero atoms included are metals such as calcium and magnesium and atoms from small compounds such as ATP, but not included are the additional water molecules found in the PDB model coordinate files.

- **Aromatic neighbors:** For each residue, we count the residues such as Phe, Tyr or Trp that have an atom neighboring that residue.

- **Nonpolar neighbors:** For each residue, we count the residues that are nonpolar but not aromatic, including residues such as Ala,Cys,Gly,Ile,Leu,Met,Pro, and Val that have an atom neighboring that residue.

- **Uncharged Polar neighbors:** Same as aromatic neighbors counts but for residues His, Aln, Gln, Ser and Thr.

- **Charged Polar neighbors:** Same as aromatic neighbor counts but for residues Asp, Glu, Lys and Arg.

We obtained per-residue energy values for each term by reading the PDB coordinate files into the *Rosetta++* suite of programs, and simply reporting the Rosetta full-atom energy per-residue calculation.

For all the graphs created in this chapter, we used the R language and environment for statistical computing [96].

5.3 Analysis

Our analysis consists of three broad classes of attributes: structural, energetic and sequence terms. The structural analysis of individual monomers includes the secondary structure classification of each residue, the number of atomic neighbors per residue and the number of neighboring residues for specific chemical types for each conformation. The energetic analysis contains the per-residue energy terms from the Rosetta scoring function [61] applied to each protein conformation, and the sequence analysis correlates properties of amino acid types relative to average mobility and shape pliability per type of residue.

5.3.1 Structural analysis

The question of which sequence and structural features distinguish rigid and flexible binding sites in protein-ligand interactions was addressed by Gunasekaran and Nussinov, and surprisingly, contact density between residues was not a structural feature that was associated with flexibility although specific types of interactions did modulate conformational changes [34]. An earlier study found that packing density values in protein structures varied approximately linearly with respect to solvent accessibility levels [5] and a later study concluded that residue flexibility is strongly influenced by relative solvent accessibility when flexibility is quantified by B factors [112]. This seems to present contradictory conclusions concerning whether or not there is an association between residue density and flexibility. In this section, we explore the correlation between residue density and flexibility, different types of chemical contact neighbors and the association of secondary structure classifications with shape pliability and mobility.

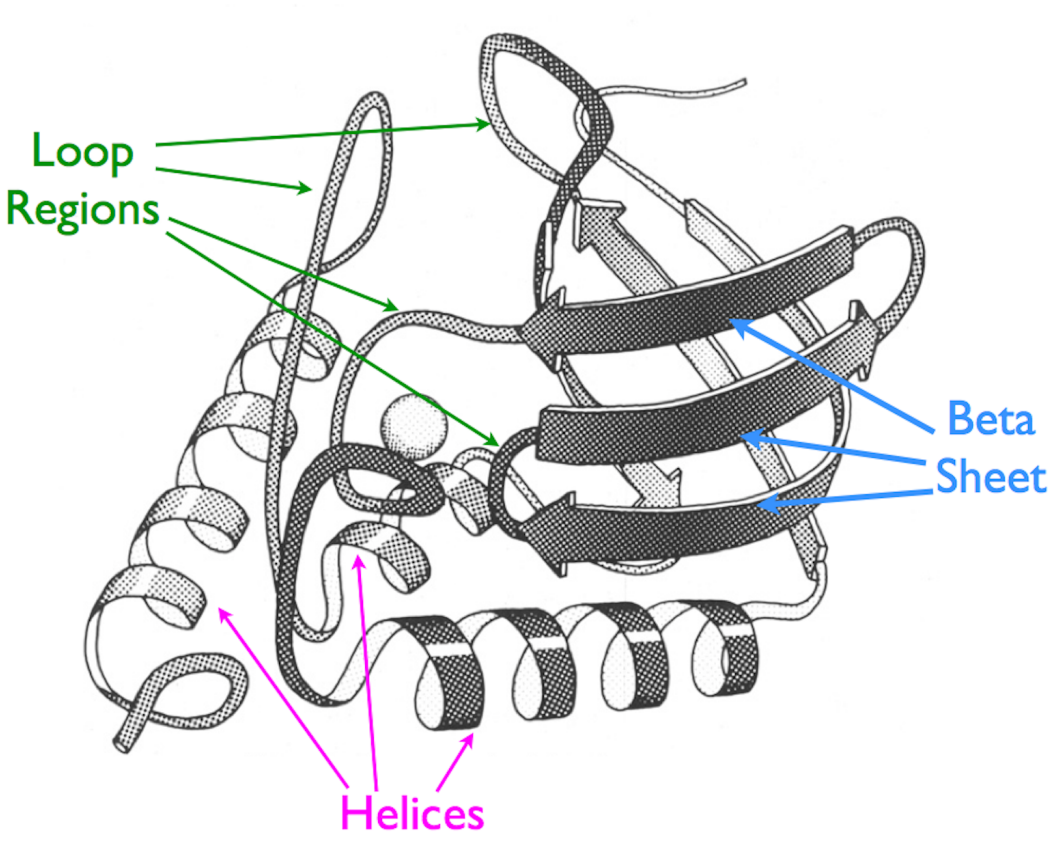
5.3.1.1 Secondary Structure

Each residue is classified according to secondary structure as defined by the DSSP, described in §5.2. Figure 5.1 shows a photograph of a drawing by Jane Richardson of a protein using ribbon diagrams (invented by Richardson) to display secondary structure elements [84]. The figure is labelled to illustrate a beta sheet, helices and loop regions in a protein. We compare the measurements for shape pliability and mobility to the residue classifications of Sheet, Helix and Loop in Table 5.1. The percentages for the most rigid bin for both mobility and shape pliability (0-20) has the highest percentages of residues for all secondary structure classifications. Looking at the highly flexible ranges (60-80 and 80-100), the highest percentages are for Loop mobility and shape pliability, whereas the lowest percentages appear in the helix shape pliability scores. The structure of helices and sheets are rigidified by hydrogen bonds, whereas loop regions do not tend to have the same rigidifying force, and hence, are more flexible. Shape pliability is a local measure, and helices are stabilized by hydrogen bonding to other residues 3, 4 or 5 apart in sequence, also a local force. For this reason, one could expect the helix shape pliaiblity percentage in the higher score range (60-100) to be lower than the corresponding mobility percentage and for the sheet classification where the hydrogen bonds may connect non-local strands, and this is indeed what is seen.

Table 5.1: Each residue is defined by its secondary structure in the protein, and we report the percentage of residue shape pliability snd mobility scores within each range of flexibility (defined by increments of scores of 20) for each secondary structure classification.

Score	Sheet		Helix		Loop	
Ranges	Mobility	Shape Pliability	Mobility	Shape Pliability	Mobility	Shape Pliability
0-20	73.1	68.8	59.1	70.6	64.2	53.5
20-40	15.7	22.5	26.4	23.5	17.7	27.6
40-60	6.04	4.29	9.5	3.21	8.06	6.85
60-80	3.31	1.73	3.44	1.14	5.41	3.06
80-100	1.8	2.75	1.56	1.55	4.67	9.03

Figure 5.1: Photograph of a 1980 pen and ink hand drawing by Jane Richardson [84], inventor of the Ribbon Diagram. The labels were added for helices, loop regions and beta sheets to show the secondary structure classifications.



5.3.1.2 Neighboring Atom Counts

Using a Spearman rank correlation, we compare the total number of neighboring stoms and number of different chemical types of neighboring atoms with the per-residue shape pliability and mobility scores. Table 5.2 shows that the all atom counts are inversely correlated to shape pliability with a correlation of -.31 ($p < 10^{-15}$). This is shown in Figure 5.2 for shape pliability score percentages in ranges colored by bins with divisions of 20. The mobility scores show less correlation, and are shown in Figure A.1 in Appendix A. The full atom counts include a few proteins with very high numbers of atomic neighbors (in the 80-100 range) because we include hetero atoms, described in our full atom counting in §5.2. The graphs present percentages for each flexibility bin, because if this were used in the context of a prediction algorithm, we would most likely bin the data in a manner similar to what is shown here. The actual counts show the frequency of different numbers of neighbors, but this is less relevant than the percentage comparison between the bins of different flexibilities versus the ranges of neighbors counts. We also show similar graphs for the percentages of nonpolar residue neighbors and aromatic residues neighbors in Figures 5.3 and 5.4 repectively. The respective mobility graphs are again shown in Appendix A, because the correlations are not as strong as those for shape pliability percentages.

Table 5.2: Spearman correlation ρ and P-values for associating shape pliability and mobility with neighboring atom counts.

`Neighboring Atom (Residue) Counts`	Shape Pliability		Mobility	
	ρ	P value	ρ	P value
All atoms	-.31	$< 2.2e-16$	-.12	$< 2.2e-16$
Polar residues (C_α atoms)	-.09	$< 2.2e-16$	-.026	0.0001869
Nonpolar residues (C_α atoms)	-0.28	$< 2.2e-16$	-0.12	$< 2.2e-16$
Aromatic residues (C_α atoms)	-0.19	$< 2.2e-16$	-0.16	$< 2.2e-16$

The answer to the question regarding the association of residue density with flexibility is not completely clear from this analysis. The correlations are stronger for shape pliability than mobility,

across the board. This is reasonable in that neighborhood density and shape pliability are both local phenomena, whereas the global flexibility of the mobility scores is not. The p values show confidence that there is close to zero possiblity that the associations are occurring due to random chance, but the correlation coefficients are weak. Perhaps the best evidence that this correlation is signficant is in the graphs, which show that the highest ranges of shape pliability values (scores of 60-100) clearly have higher percentages of the total at the densities (for all chemical types) with the fewest neighbors.

Figure 5.2: This graph shows the shape pliability score percentages for the total number of atomic neighbors per residue, measured by distance cutoff of $5\mathring{A}$ from the residue's C_α atom.

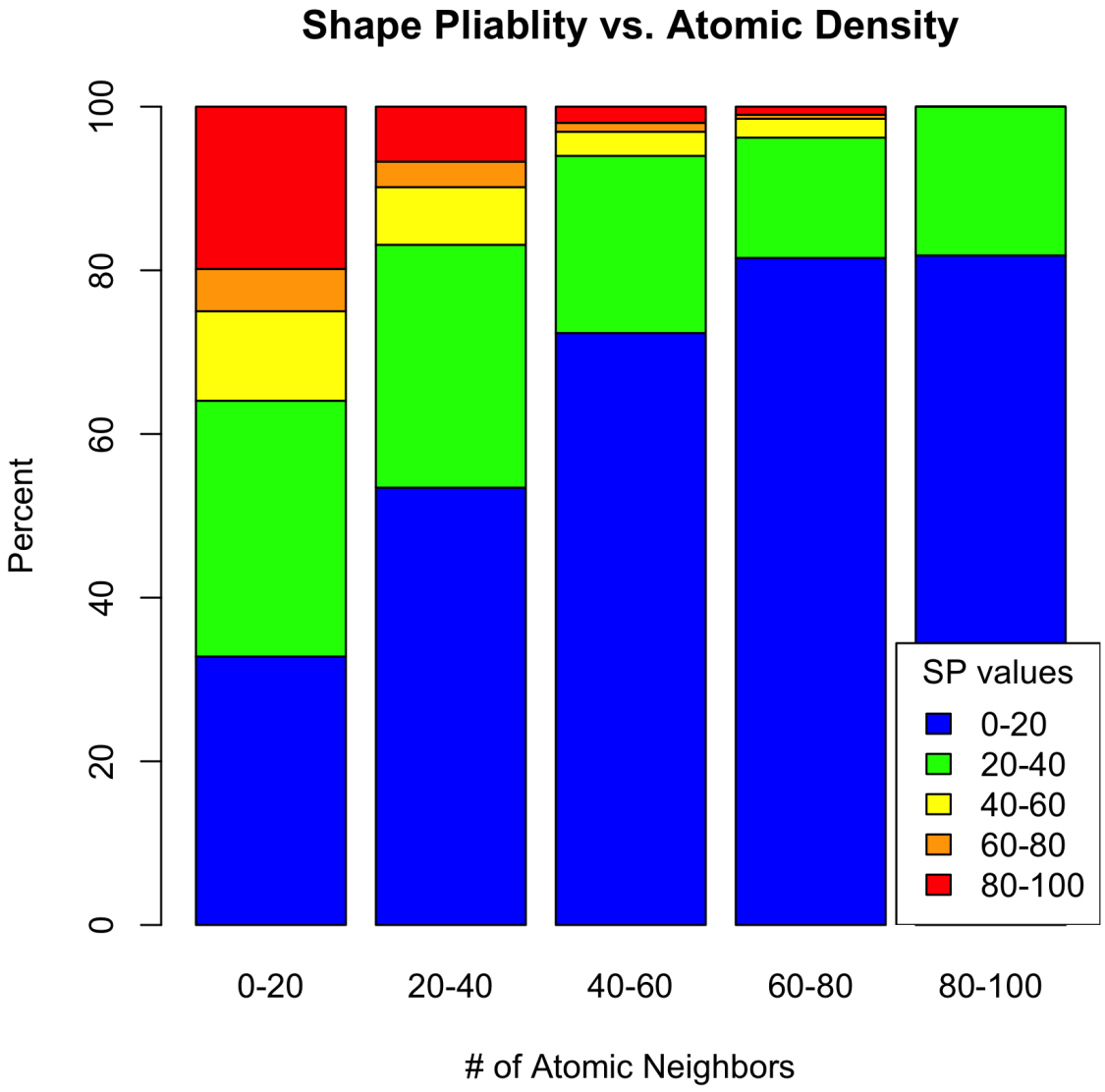


Figure 5.3: This graph shows shape pliability score percentages for ranges of the number of nonpolar residue neighbors per residue.

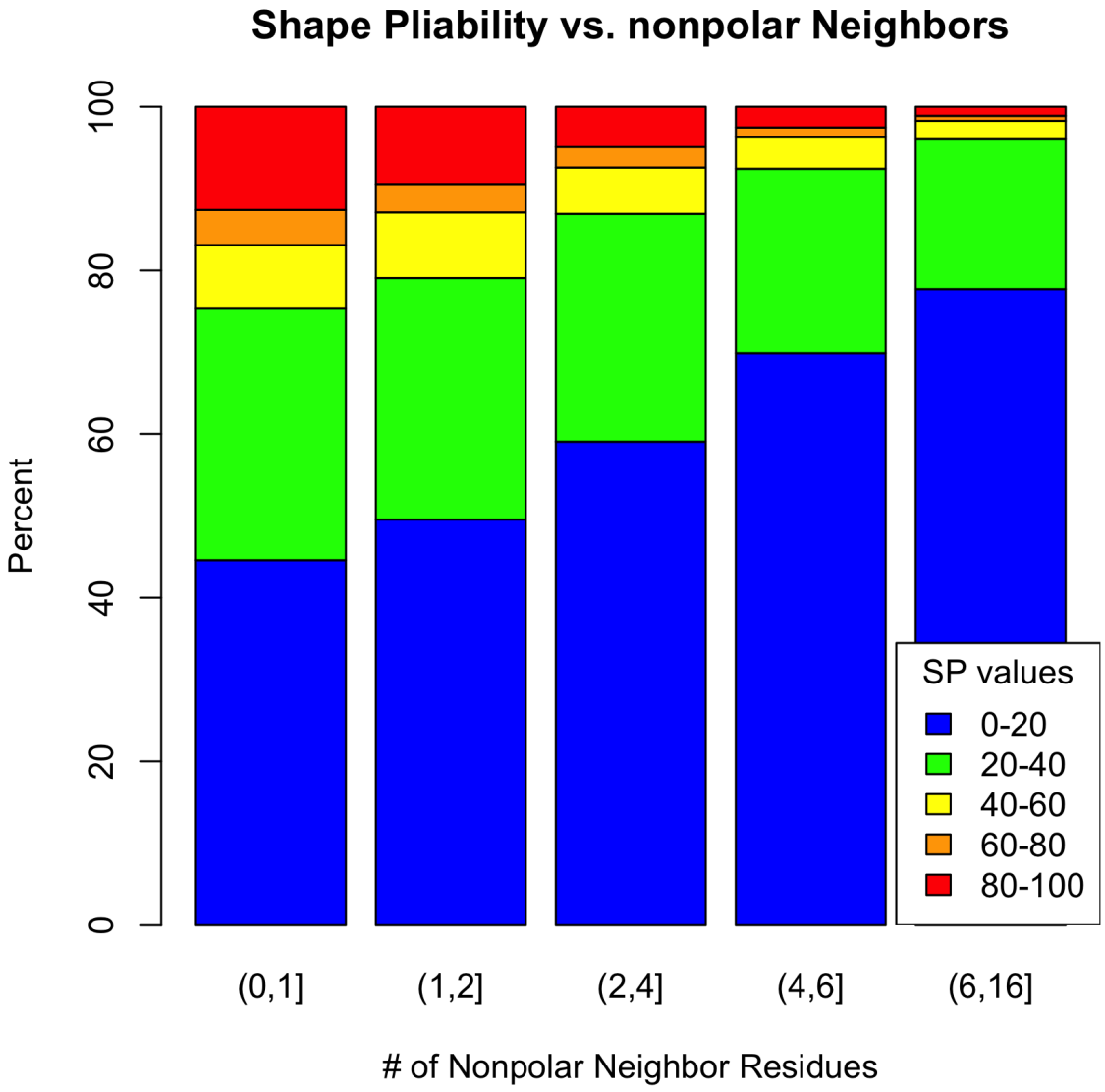
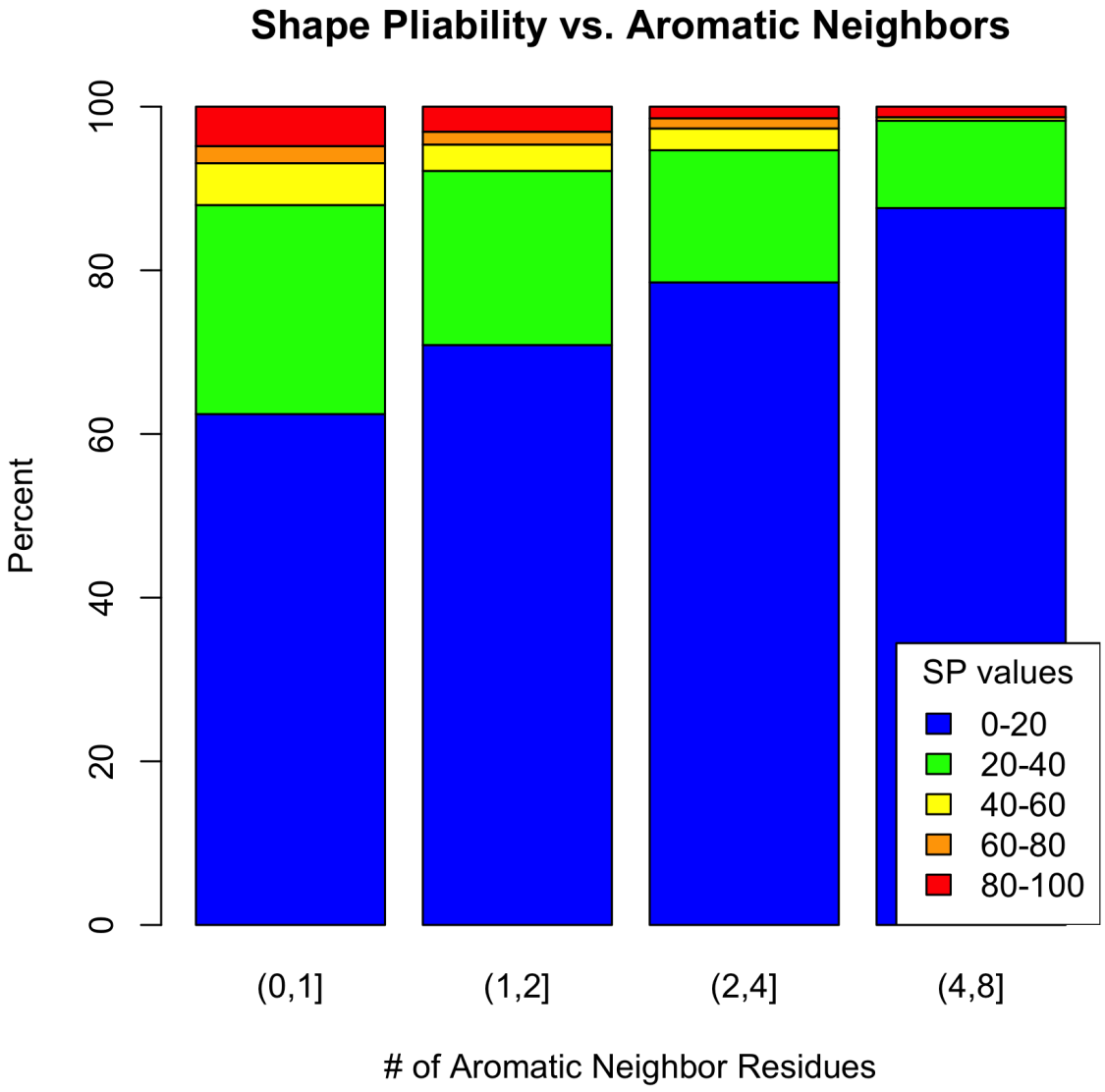


Figure 5.4: This graph shows shape pliability score percentages for ranges of the number of aromatic residue neighbors per residue.



5.3.2 Analysis of Rosetta Energy Terms

We examine the Rosetta energy and scoring function terms that are related to how well the protein is packed [60], because there is not an explicit scoring function term in Rosetta for measuring flexibility. These terms include the following:

- E_{pair}: pairwise electrostatic term derived from statistics over proteins in the PDB; favors salt bridges, or interactions between atoms with opposite charges.
- E_{atr}: Lennard Jones attractive force; rewards close contacts.
- E_{rep}: Lennard Jones repulsive force; penalizes contacts that are too close.
- E_{sol}: Lazaridis-Karplus solvation energy models water implicitly and penalizes the burial of polar atoms [59].
- $SASA_{pack}$: measures the quality of packing by penalizing voids that are too small to accomodate a water molecule; compares the measurement to expected packing observed in structures in the PDB. Negative values are favorable and indicate that this residue is more tightly packed than what is found on average in the PDB [64].
- E_{hbnd}: an orientation dependent hydrogen bonding potential for long and short range hydrogen bonding [51]; we ignore sidechain-sidechain hydrogen bond energy, but include all backbone hydrogen bond related energies.

Table 5.3: Spearman correlation ρ and P-values for associating shape pliability and mobility with Rosetta energy terms.

`Energy Term`	Shape Pliability		Mobility	
	ρ	P value	ρ	P value
Electrostatic	-0.009	0.21	-.001	0.87
LJ attractive	0.30	$< 2.2e-16$	0.13	$< 2.2e-16$
LJ repulsive	.04	4.074e-10	.10	$< 2.2e-16$
Solvation	-.15	$< 2.2e-16$	-.07	$< 2.2e-16$
SASA-pack	.07	$< 2.2e-16$	.07	$< 2.2e-16$
Hydrogen bond	.28	$< 2.2e-16$	.11	$< 2.2e-16$

In Table 5.3, the highest correlation for the energetic terms are for the Lennard Jones attractive force and the hydrogen bond energy, both related to shape pliability. These correlations

are also visible in Figures 5.5 and 5.6. The solvation energy term has the next highest correlation, and from Figure 5.7 we see that this correlation is less evident. It does not seem surprising that the E_{atr} and E_{hbnd} forces are associated with rigidity, at least locally, in the structures. Both terms reward well-formed close contacts that are likely rigidifying, as we have already seen in the secondary structure and atomic density correlations in §5.3.1. The solvation energy term is also related to local packing and its correlation to shape pliability is expected.

The structural data from the PDB are representative of overall low energy, native structures and the E_{rep} term is not significant here. It doesn't mean that this term isn't relevant to flexibility, but only that it isn't a factor in the structures we are analyzing. Likewise, $SASA_{pack}$ is a very useful term for protein structure prediction, where candidate structures may lack the good packing qualities inherent in crystal structures, but in our data the native structures analyzed are presumably well-packed. The surprising result, however, is the lack of any correlation of flexibility to the electrostatic E_{pair} term. According to [53], salt bridges appear to constrain flexibility and motion and are rarely found across regions that are joined by flexible hinges. This anomaly in the correlation results requires further investigation.

Figure 5.5: This graph shows shape pliability score percentages for ranges of E_{atr} per residue.

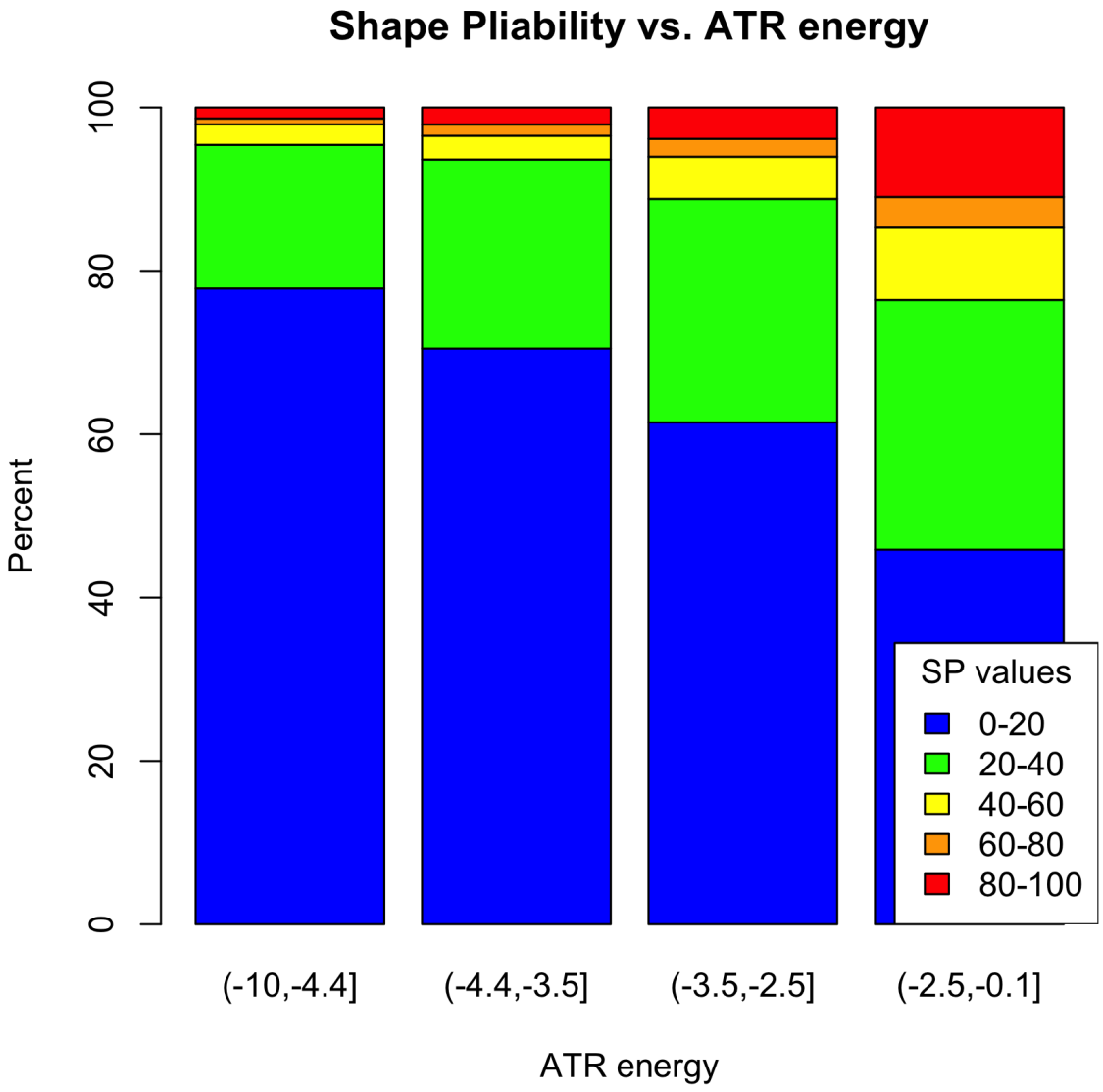


Figure 5.6: This graph shows shape pliability score percentages for ranges of hydrogen bond energy per residue.

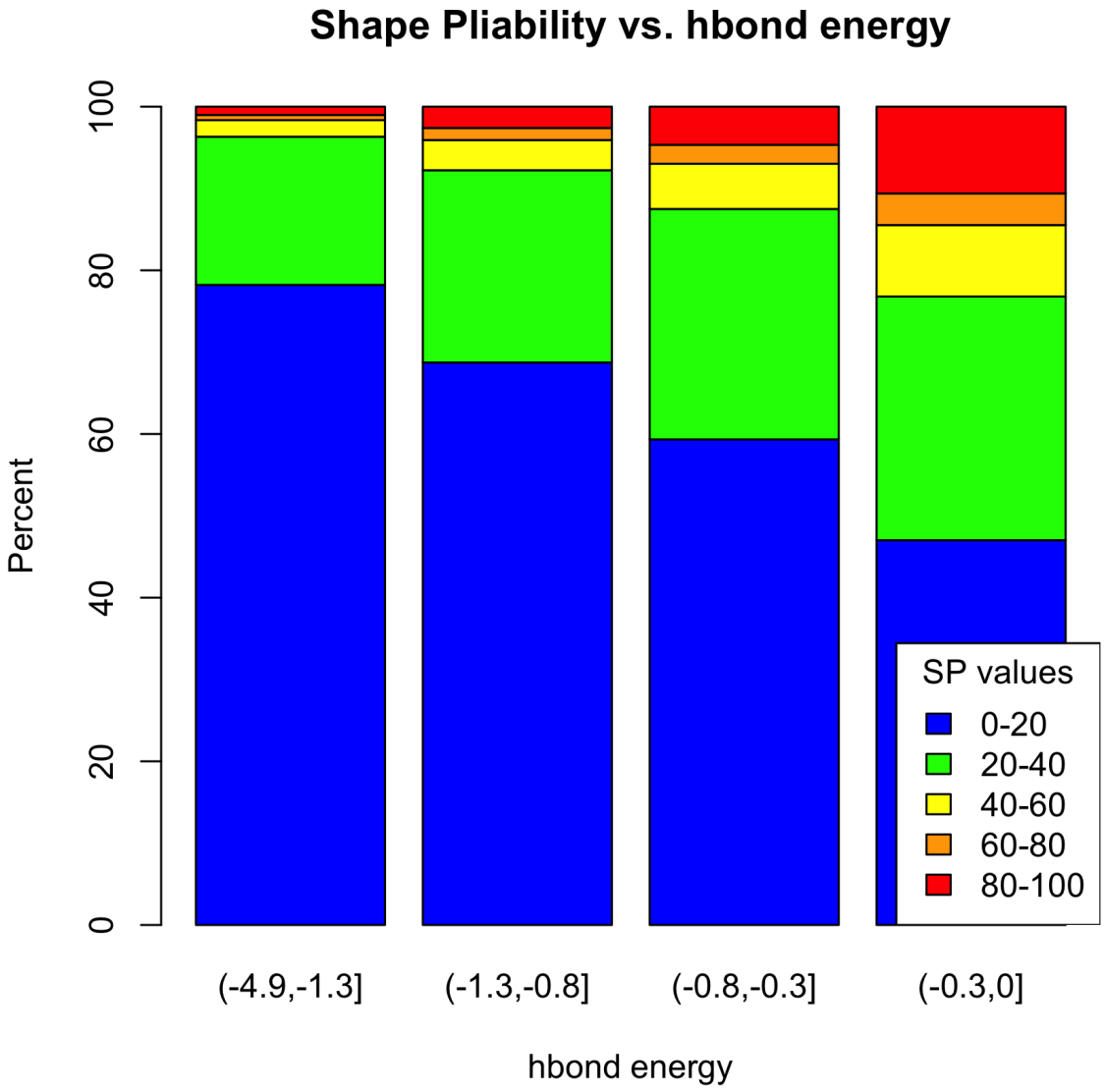
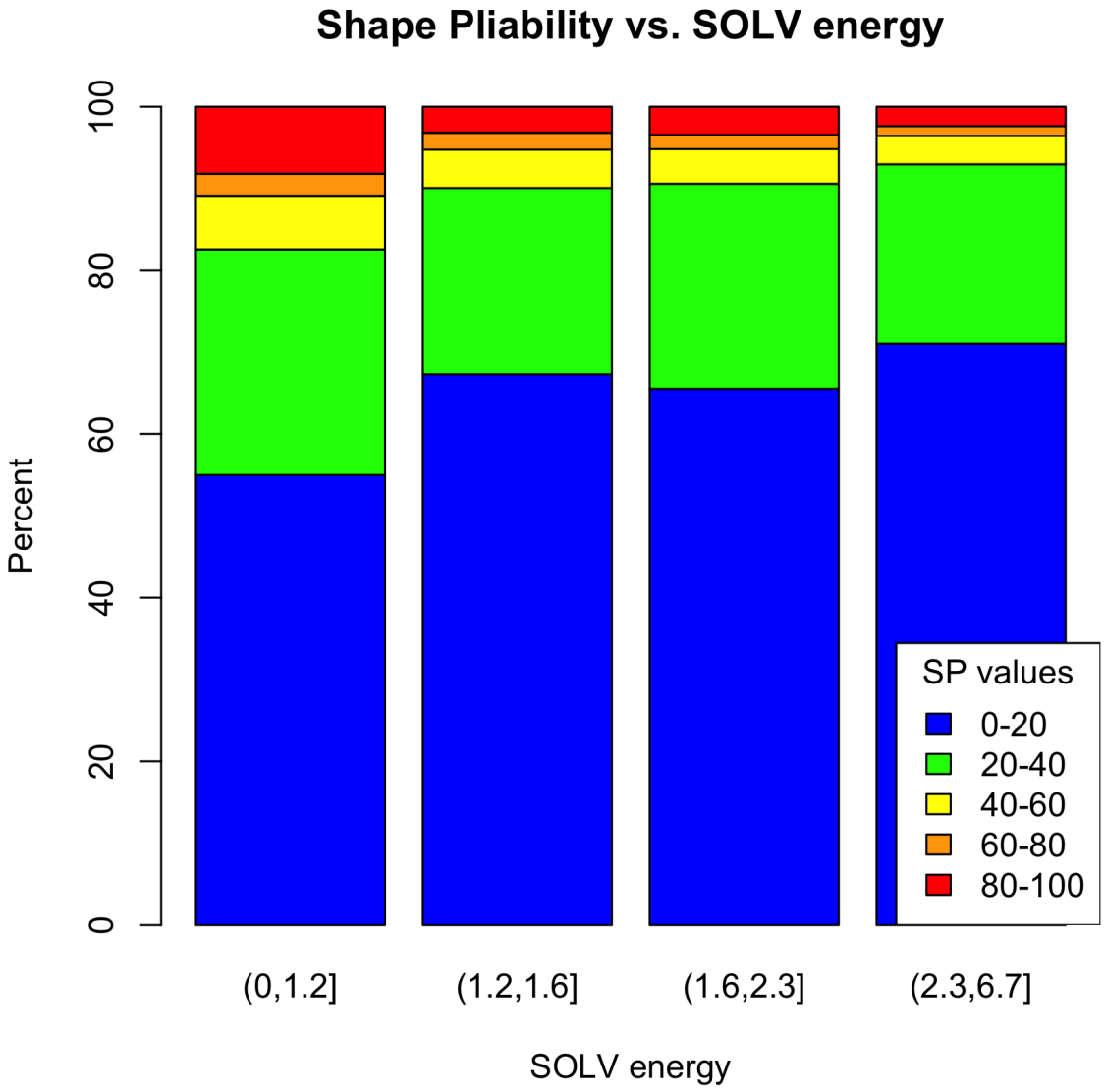


Figure 5.7: This graph shows shape pliability score percentages for ranges of values of Lazaridis-Karplus solvation energy [59] per residue.



5.3.3 Sequence analysis

Protein sequences and instrinsic disorder (at the extreme end of protein flexibility) have been extensively studied [81], and sequence data have been used for prediction of motion in protein-protein interactions [40]. We are foremost interested in the predictive qualities of protein structural data, but sequence information has been shown to play a decisive role in flexibility and therefore must be considered within the context of this analysis.

We looked at 2 different indices related to the characteristics of particular residues. Table 5.4 reports molecular weights and hydropathicity per residue. The hydropathicity index goes from -4.5 for the extremely hydrophilic (water loving) to 4.5 for extremely hydrophobic. Charged polar residues are the most hydrophilic because the charged atoms seek neutralization in water, whereas hydrophobic atoms are nonpolar and typically locate in the protein core away from the aqueous solution surrounding the surface. The Spearman correlation of these indices to shape pliability and mobility is shown in Table 5.5. Hydropathicity shows higher correlation to mobility and shape pliability than any of the structural or energetic terms, but the p values are not as persuasive because here we are looking at only 20 data points each.

For the molecular weight index, Glycine (GLY) has the lowest weight (75 g/mol) and Tryptophan (TRP) is highest (204 g/mol). Both shape pliability and mobility have higher percentages of flexible glycines and rigid trptophan's, but the correspondance between weight and flexibility is not linear between the endpoints. This is visible in Figure 5.10, with the TRP and GLY residues circled because their mean shape pliability is almost perfectly inversely correlated to molecular weight.

For the hydropathicity index, the shape pliability score percents show somewhat more flexibility when hydrophilic and less when hydrophobic, with some huge exceptions like Glycine and Trytophan, which correlate better to molecular weight than hydropathicity. Again, the data points for these residues are circled in Figure 5.9. The highest correlation is evident in the mean mobility versus hydropathicity correlation coefficient from Table 5.4, and this correlation is shown

graphically in Figure 5.8.

In Figure 5.11, we look at the shape pliability percentages for each bin as shown in previous sections, but now with regard to residue type. Glycine (GLY) residues have much higher percentage (9.5%) of very high shape pliability (80-100) than all others, with the next closest Aspartic Acid (ASP) at 6.5%). Glycine has no heavy atoms in its sidechain so the backbone can be sampled in many positions without clashing with other sidechain atoms. Phenylalanine (PHE), Trptophan (TRP), and Valine (VAL) have higher percentages of low shape pliability (0-20) than the remainder, whereas for mobility (not shown), the largest precentage of rigid scores (0-20) are for Tryptophan (TRP) and Cysteine (CYS), and no single residue stands out as having a higher percentage of high mobility scores.

5.4 Discussion and Summary

The attibutes that show correlation to flexibility could be useful in *predicting* the backbone flexibility for a single protein conformation, and the prediction, in turn, could be used within the context of molecular modeling algorithms such as protein design or protein-protein docking to guide sampling and monitor changes in structural flexibility. Even with weak correlations, using a kernel to train and test a machine learning prediction method will combine the input attributes of each residue, and the weak correlations may still be useful in that context. We used Rosetta because of its successes in the research areas of protein and interface design; these are algorithms where prediction of flexibility could be useful for selecting specific mutations to explore.

Table 5.4: For each amino acid residue, the molecular weight and hydropathicity index [55] is shown.

AA	Mol. weight (g/mol)	Hydropathicity Index	Mean Mobility	Mean Shape Pliablity
ALA (A)	89.00	1.80	18.00	21.50
ARG (R)	174.00	-4.50	20.94	23.45
ASN (N)	132.00	-3.50	16.99	22.83
ASP (D)	133.00	-3.50	19.64	25.01
CYS (C)	121.00	2.50	14.86	20.08
GLN (Q)	146.00	-3.50	18.77	21.71
GLU (E)	147.00	-3.50	23.17	24.14
GLY (G)	75.00	-0.40	19.06	27.40
HIS (H)	155.00	-3.20	21.13	23.43
ILE (I)	131.00	4.50	17.16	18.34
LEU (L)	131.00	3.80	17.40	18.58
LYS (K)	146.00	-3.90	21.33	23.31
MET (M)	149.00	1.90	19.47	21.58
PHE (F)	165.00	2.80	16.79	18.60
PRO (P)	115.00	-1.60	19.90	21.49
SER (S)	105.00	-0.80	17.61	24.74
THR (T)	119.00	-0.70	19.09	23.57
TRP (W)	204.00	-0.90	14.17	17.40
TYR (Y)	181.00	-1.30	16.80	20.34
VAL (V)	117.00	4.20	16.51	17.81

Table 5.5: Mobility and shape pliability averages per residue type related to hydrophathicity and molecular weight.

`AA Index`	Mean Shape Pliability		Mean Mobility	
	ρ	P value	ρ	P value
Hydropathicity	-0.583	0.007	-0.611	0.004
Molecular weight	0.029	0.902	-0.187	0.429

Figure 5.8: Mean residue type mobility versus hydropathicity

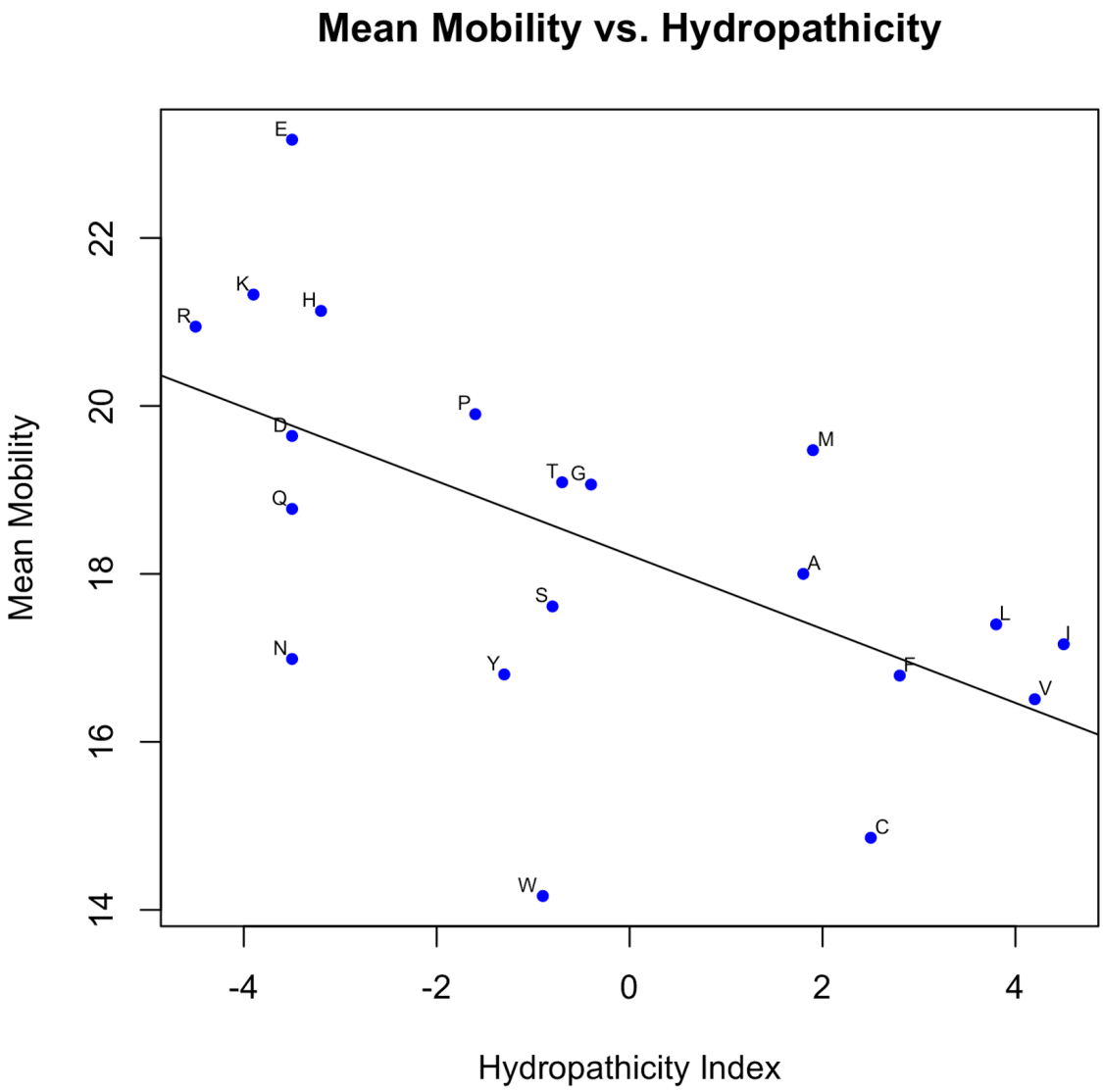


Figure 5.9: Mean residue type shape pliability versus hydropathicity

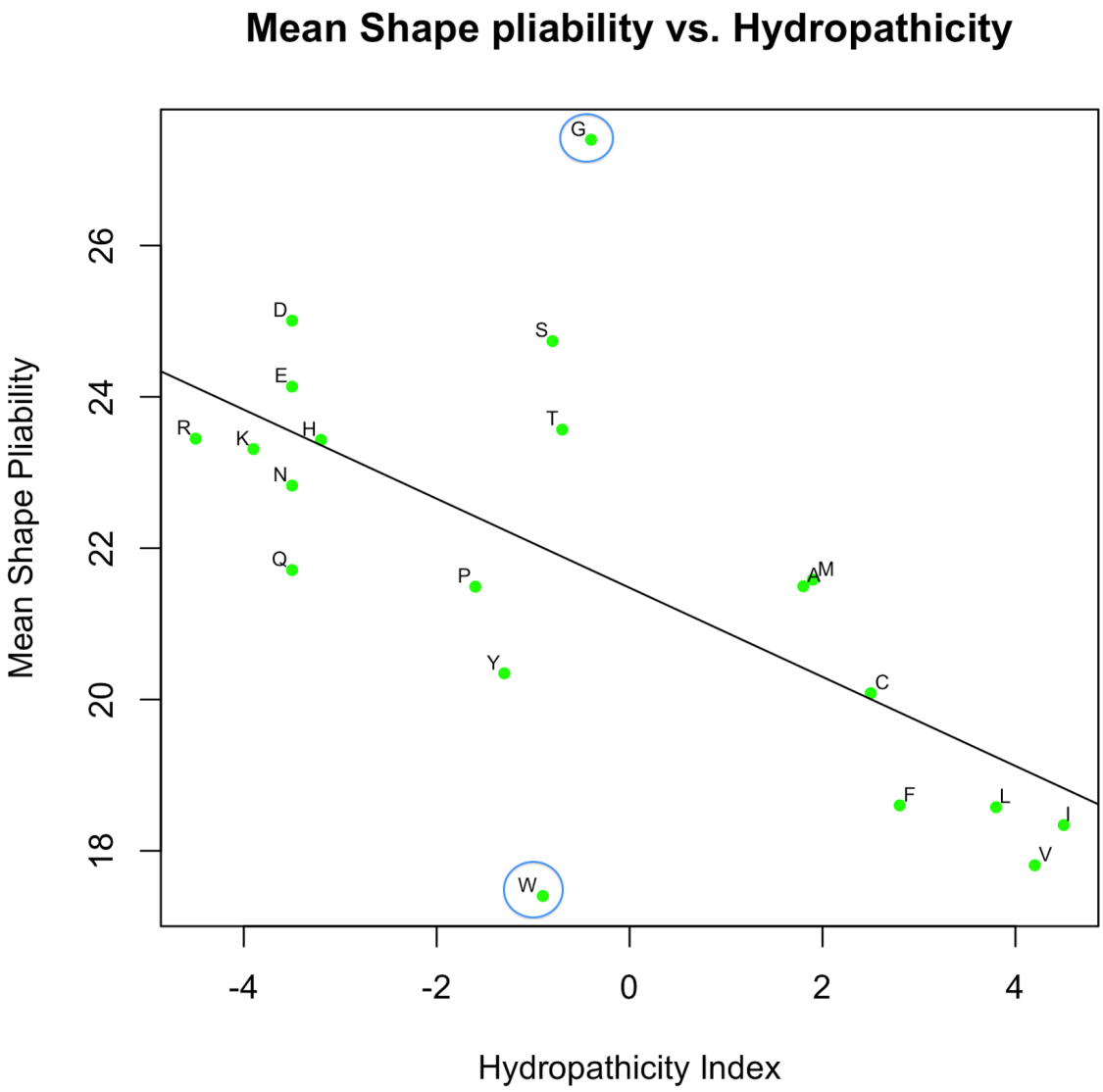


Figure 5.10: Mean residue type shape pliability versus molecular weight

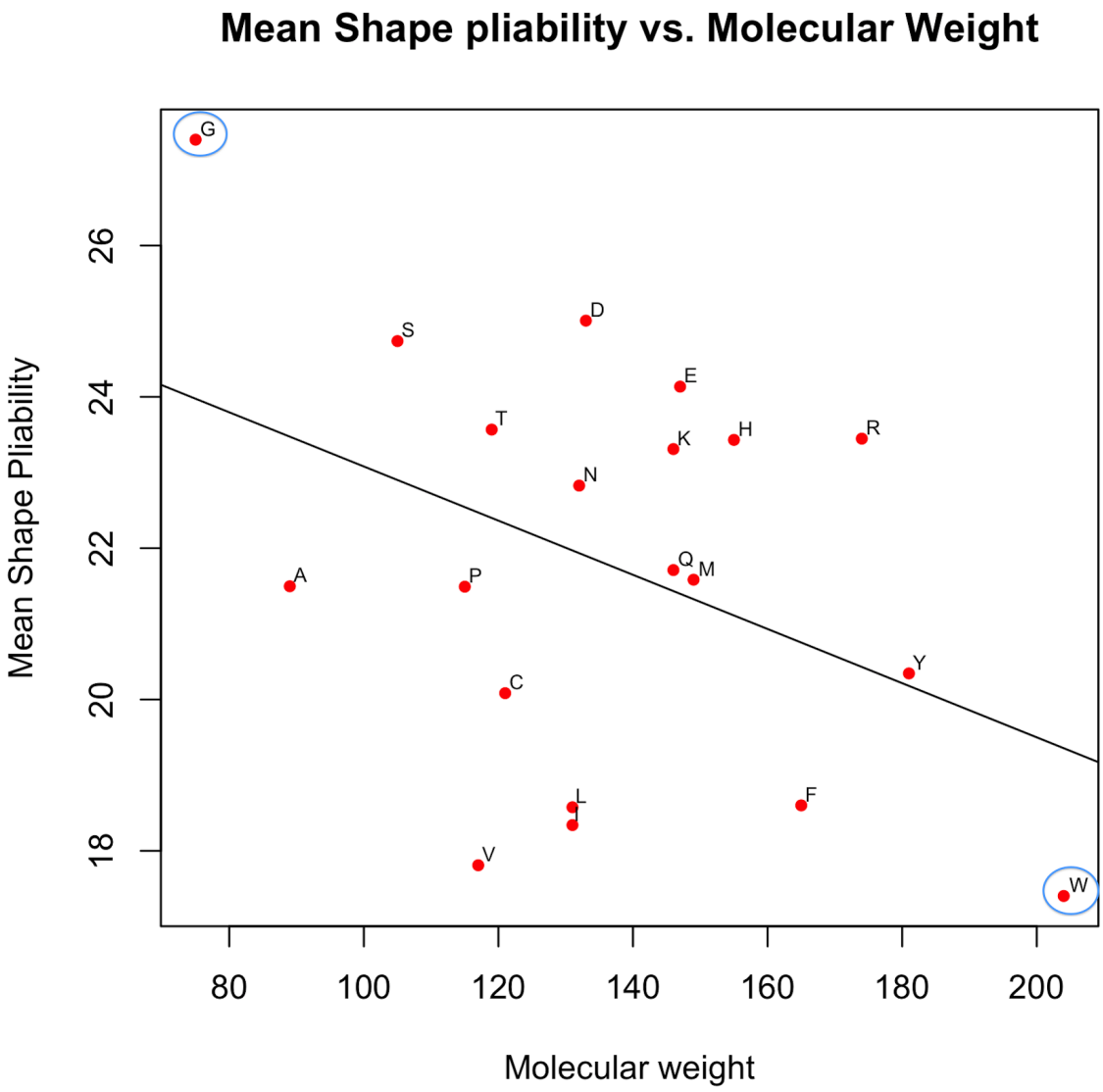
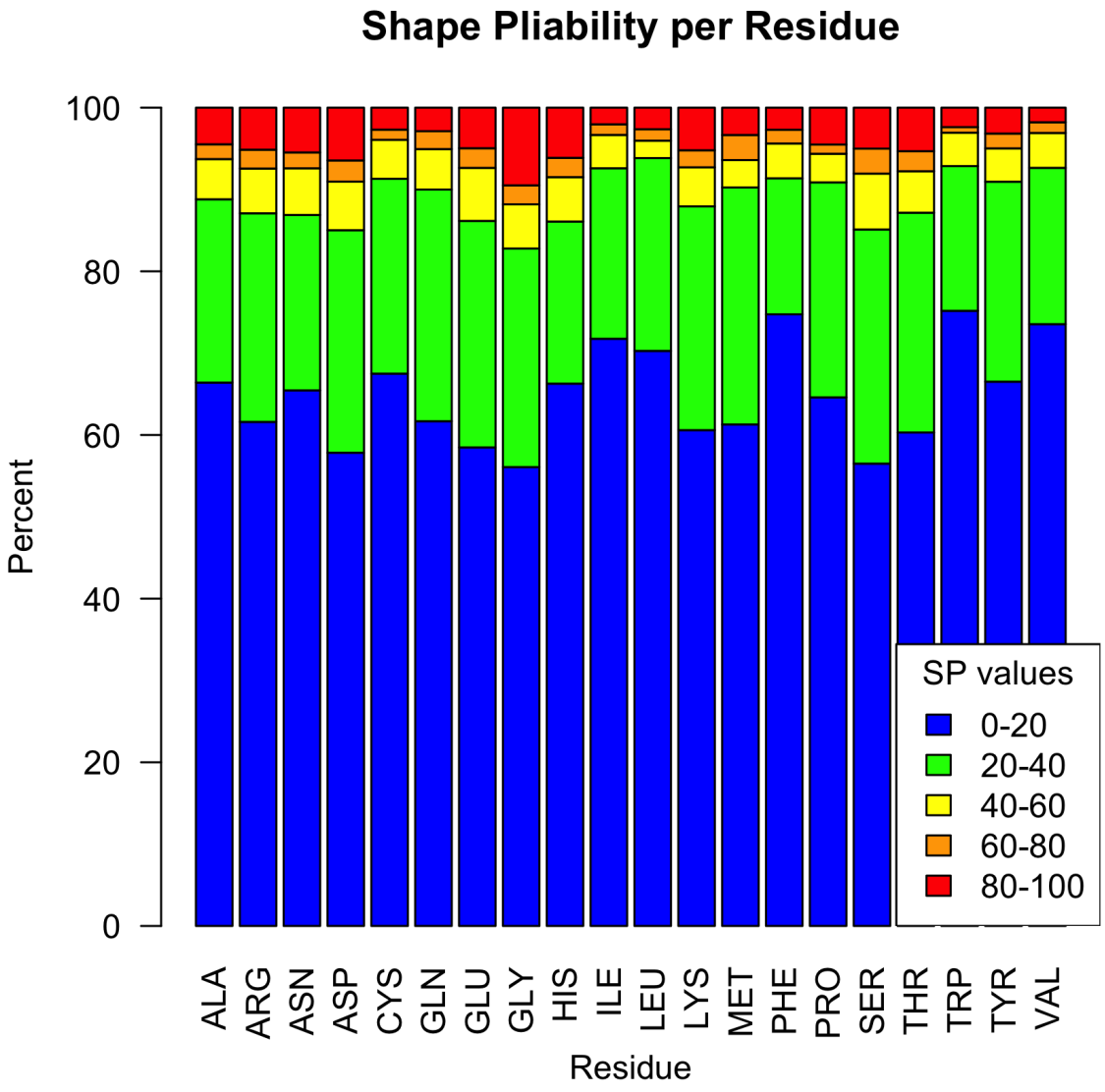


Figure 5.11: Shape pliability score percentages for the 20 naturally occurring residues in proteins.



Chapter 6

Conclusion

6.1 Thesis Contributions

- We designed and implemented a new algorithm to align proteins by secondary structure elements and minimize distances between corresponding residues when secondary structure elements align closely. The goal of this algorithm is to find distances between corresponding aligned residues that relate to flexiblity, such that rigid regions of the protein align more closely than flexible regions. The comparison to an existing method known to handle flexibility in its alignment shows the differences in how the methods produce measurements for specific residues that are known to be flexible versus those that are less flexible.

- We defined a characterization of protein backbone flexibility as two distinct types of flexiblity with per-residue scoring functions for each.

- We validated the scoring methods over a wide variety of protein conformational changes, showing the generality and accuracy of the method.

- We applied the flexibility scoring functions to three important, disparate research areas. These include two biochemisty areas pertaining to drug and enzyme design, and a third application to error analysis in protein structure predictions.

- We analyzed the correlation of our flexibility scores to structural, energetic and sequence attributes that may influence (or predict) flexibility.

6.2 Future Directions

- The development of an evaluation method for protein structure prediction based on our two flexibility scoring functions may be useful for assessing the quality and ranking of protein structure predictions.

- The implementation of a machine learning prediction method for prediction of shape pliability, using the attributes that appeared to have some correlation in Chapter 5, could be used to predict local flexibility in individual protein structures.

- The correlation of evolution to residue flexiblity will provide information in the likelihood of mutation in rigid or flexible residues of protein structures.

- The generalization of the scoring functions to handle comparisons of more than two structures could apply to NMR coordinate files with multiple conformations, populations of more than two x-ray crystal conformations, or modeling algorithms with multiple sample points containing different backbone coordinates for a particular protein structure.

Bibliography

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Molecular Biology of the Cell. Garland Science, New York, N.Y., fourth edition, 2002.

[2] Christian B. Anfinsen. Principles that govern the folding of protein chains. Science, 181(4096):223–320, 1973.

[3] Euiyoung Bae and Jr. George N. Phillips. Roles of static and dynamic domains in stability and catalysis of adenylate kinase. PNAS, 103(7):2132–2137, 2006.

[4] Cyril Barinka, Graham Perry, Jennifer Callahan, David E. Shaw, Alice Kuo, Khalil Bdeir, Douglas B. Cines, Andrew Mazar, and Jacek Lubkowski. Structural basis of interaction between urokinase-type plasimogen activator and its receptor. J. Mol. Biol., 363:482–495, 2006.

[5] Franck Baud and Samuel Karlin. Measures of residue density in protein structures. PNAS, 96(22):12494–12499, 1999.

[6] Eric Beitz. TeXshade: shading and labeling multiple sequence alignments using LaTeX2e. Bioinformatics, 16:135–139, 2000.

[7] Lluis Bellsolell, Jesus Prieto, Luis Serrano, and Miguel Coll. Magnesium binding to the bacterial chemotaxis protein CheY results in large conformational changes involving its functional surface. J. Mol. Biol., 238:489–495, 1994.

[8] Jeremy M. Berg, John L. Tymoczko, and Lubert Stryer. Biochemistry. W. H. Freeman and Company, New York, sixth edition, 2007.

[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I.N Shinkyalov, and P. E. Bourne. The protein data bank. Nucleic Acids Research, 28:235–242, 2000.

[10] D. J. Best and D. E. Roberts. Algorithm AS 89: The upper tail probabilities of Spearman's Rho. Journal of the Royal Statistical Society. Series C (Applied Statistics), 24(3):377–379, 1975.

[11] David D. Boehr, H. Jane Dyson, and Peter E. Wright. An NMR perspective on enzyme dynamics. Chem. Rev., 106:3055–3079, 2006.

[12] Dustin E. Bosch, Adam J. Kimple, Deanne W. Sammond, Robin E. Muller, Michael J. Miley, Mischa Machius, Brian Kuhlman, Francis S. Willard, and David P. Siderovski. Structural determinants of affinity enhancement between GoLoco motifs and g-protein α subunit mutants. J. Biol. Chem., 286:3351–3358, 2011.

[13] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. Charmm: A program for macromolecular energy, minimization and dynamic calculations. Journal of Computational Chemistry, 4(2):187–217, 1983.

[14] S. J. Brownlow, J. H. M. Cabral, R. Cooper, D. R. Flower, S. J. Yewdall, I. Polikarpov, A.C.T. North, and L. Sawyer. Bovine $\beta-$lactoglobulin at 1.8 angstrom resolution-still an enigmatic lipocalin. Structure, 5:481–497, 1997.

[15] Leslie D. Burtnick, Edward K. Keopf, Jonathan Grimes, E. Yvonne Jones, David I. Stuart, Paul J. McLaughlin, and Robert C. Robinson. The crystal structure of plasma gelsolin: Implications for actin severing, capping and nucleation. Cell, 90:661–670, 1997.

[16] Esther Caballero-Manrique, Jenelle K. Bray, William A. Deutschman, Frederick W. Dahlquist, and Marina G. Guenza. A theory of protein dynamics to predict NMR relaxation. Biophys. J., (4128-4140), 2007.

[17] Rajagopal Chattopadhyaya, William E. Meador, Anthony R. Means, and Florante A. Quiocho. Calmodulin structure refined at 1.7$\mathring{A}$ resolution. J. Mol. Biol., 228:1177–1192, 1992.

[18] Han Choe, Leslie D. Burtnick, Marisan Mejillano, Helen L. Yin, Robert C. Robinson, and Senyon Choe. The calcium activation of gelsolin: Insights from the 3$\mathring{A}$ structure of the G4-G6/Actin complex. J. Mol. Biol., 324:691–702, 2002.

[19] Steven A. Combs, Samuel L. DeLuca, Stephanie H. DeLuca, Gordon H. Lemmon, David P. Nannemann, Elizabeth D. Nguyen, Jordan R. Willis, Jonathan H. Sheehan, and Jens Meiler. Small-moleculare ligand docking into comparative models with Rosetta. Nature Protocols, 8(7):1277–1298, 2013.

[20] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popovic, and Foldit players. Predicting protein structures with a multiplayer online game. Nature, 466:756–760, 2010.

[21] Evangelos A. Coutsias, Chaok Seok, and Ken A. Dill. Using quaternions to calculate RMSD. J. Comp. Chem., 25(15):1849–1857, 2004.

[22] Peter Csermely, Robin Palotai, and Ruth Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends in Biochemical Sciences, 35:539–546, 2010.

[23] J. J. M. Cuppen. The singular value decomposition in product form. Siam J. Sci. Stat. Comput., 4(2):216–222, 1983.

[24] Kelly L. Damm and Heather A. Carlson. Gaussian-weighed RMSD superposition of proteins: A structural comparison for flexible proteins and predicted protein structures. Biophys. J., 90:4558–4573, 2006.

[25] Ron O. Dror, Robert M. Dirks, J. P. Grossman, Huafeng Xu, and David E. Shaw. Biomolecular simulation: A computational microscope for molecular biology. Annu. Rev. Biophys, 41:429–452, 2012.

[26] Mansour Ebrahimi, Amir Lakizadeh, Parisa Agha-Golzadeh, Esmaeil Ebrahimie, and Mahdi Ebrahimi. Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: A new vista in engineering enzymes. PLoS One, 6(8):1–11, 2011.

[27] Nathaniel Echols, Duncan Milburn, and Mark Gerstein. Molmovdb: analysis and visualization of conformational change and structural flexibility. Nucleic Acids Research, 31(1):478–482, 2003.

[28] Ugar Emekli, Dina Schneidman-Duhovny, Haim J. Wolfson, Ruth Nussinov, and Turkan Haliloglu. Hingeprot: Automated prediction of hinges in protein structures. Proteins, 70:1219–1227, 2008.

[29] Addi R. Fadel, Dan Q. Jin, Gaetano T. Montelione, and Ronald M. Levy. Crankshaft motions of the polypeptide backbone in molecular dynamics siumulations of human type-α transforming growth factor. Journal of Biomolecular NMR, 6:221–226, 1995.

[30] Krzysztof Fidelis, Andrly Kryshtafovych, and Bohdan Monastyrskyy. Protein structure prediction center. http://predictioncenter.org/casp9/index.cgi, 2010.

[31] Maria M. Flocco and Sherry L. Mowbray. Cα-based torsion angles: A simple tool to analyze protein conformational changes. Protein Science, 4:2118–2122, 1995.

[32] Darren R. Flower. Rotational superposition: A review of methods. J. Mol. Grahics Mod., 17:238–244, June/August 1999.

[33] Xiaoran Fu, James R. Apgar, and Amy E. Keating. Modeling backbone flexibility to achieve sequence diversity: the design of novel α-helical ligands for $Bcl - x_L$. J. Mol. Biol., 371:1099–1117, 2007.

[34] Kannan Gunasekaran and Ruth Nussinov. How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. J. Mol. Biol., 365:257–273, 2007.

[35] Donovan C. Haines, Diana R. Tomchick, Mischa Machius, and Julian A. Peterson. Pivotal role of water in the mechanism of P450BM-3. Biochemistry, 40:13456–13465, 2001.

[36] Steven Hayward and Richard A. Lee. Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. Journal of Molecular Graphics and Modelling, 21:181–183, 2002.

[37] Ulf Hensen, Tim Meyer, Jürgen Haas, René Rex, Gert Vriend, and Helmut Grubmüller. Exploring protein dynamics space: The dynasome as the missing link between protein structure and function. PLoS One, 7(5):1–16, May 2012.

[38] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. Nature, 450:964–972, December 2007.

[39] Katherine A. Henzler-Wildman, Ming Lei, Vu Thai, S. Jordan Kerns, Martin Karplus, and Dorothee Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature, 450:913–918, December 2007.

[40] Shuichi Hirose, Kiyonobu Yokota, Yutaka Kuroda, Hiroshi Wako, Shigeru Endo, Satoru Kanai, and Tamotsu Noguchi. Prediction of protein motions from amino acid sequence and its application to protein-protein interaction. BMC Structural Biology, 10(20):1–18, 2010.

[41] Viktor Hornak and Carlos Simmerling. Targeting structural flexibility in HIV-1 protease inhibitor binding. Drug Discovery Today, 12(3/4):132–138, February 2007.

[42] Howook Hwang, Thom Vreven, Joel Janin, and Zhiping Weng. Protein-protein docking benchmark version 4.0. Proteins, 78(15):3111–3114, November 2010.

[43] Mitsuhiko Ikura, G. Marius Clore, Angela M. Gronenborn, Guang Zhu, Claude B. Klee, and Ad Bax. Solution structure of a calmodulin-target peptide complex by multidimensional NMR. Science, 256:632–638, 1992.

[44] Lukasz A. Joachimiak, Tanja Kortemme, Barry L. Stoddard, and David Baker. Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. J. Mol. Biol., 361:195–208, 2006.

[45] Wolfgang Kabsch. A solution of the best rotation to relate two sets of vectors. Acta Crystallographica, 32:922, 1976.

[46] Wolfgang Kabsch and Chris Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22:2577–2637, 1983.

[47] Dana Kahra, Michael Kovermann, Christian Low, Verana Hirschfeld, Caroline Haupt, Jochen Balback, and Christian Gerhard Hubner. Conformational plasticity and dynamics in the generic protein folding catalyst slyD unraveled by single-molecule FRET. J. Mol. Biol., 411:781–790, 2011.

[48] Kevin S. Keating, Samuel C. Flores, Mark B. Gerstein, and Leslie A. Kuhn. StoneHinge: Hinge prediction by network analysis of individual protein structures. Protein Science, 18:359–371, 2009.

[49] Randall J. Kimple, Michelle E. Kimple, Laurie Betts, John Sondek, and David P. Siderovski. Structural determinants for GoLoco-induced inhibition of the nucleotide release by g_α subunits. Nature, 416:878–881, 2002.

[50] G. Kontopidis, C. Holt, and L. Sawyer. Invited review: β-lactoblobulin: Binding properties, structure and function. J. Dairy Sci, 87(4):785–96, Apr 2004.

[51] Tanja Kortemme, Alexandre V. Morozov, and David Baker. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J. Mol. Biol., 326:1239–1259, 2003.

[52] Brian Kuhlman, Gautam Dantas, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker. Design of a novel globular protein fold with atomic-level accuracy. Science, 302:1364–1368, 2003.

[53] Sandeep Kumar and Ruth Nussinov. Close-range electrostatic interactions in proteins. ChemBioChem, 3(7):604–617, 2002.

[54] Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li. Applied Linear Statistical Models. McGraw-Hill Irwin, New York, N.Y., fifth edition, 2005.

[55] Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol., 157:105–132, 1982.

[56] P. Therese Lang, Ho-Leung Ng, James S. Fraser, Jacob E. Corn, Nathaniel Echols, Mark Sales, James M. Holton, and Tom Alber. Automated electron-density sampling reveals widespread conformational polymorphism in proteins. Protein Science, 19:1420–1431, 2010.

[57] Oliver F. Lange, Nils-Alexander Lakomek, Cristophe Farés, Gunnar F. Schröder, Korvin F. A. Walter, Stefan Becker, Jens Meiler, Helmut Grubmüller, Christian Griesinger, and Bert L. de Groot. Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. Science, 320(13):1471–1475, 2008.

[58] Melinda Layten, Viktor Hornak, and Carlos Simmerling. The open structure of a multi-drug-resistant HIV-1 protease is stabilized by crystal packing contacts. J Am Chem Soc, 128:13360–13361, 2006.

[59] Themis Lazaridis and Martin Karplus. Effective energy functions for proteins in solution. Proteins, 35:133–152, 1999.

[60] Andrew Leaver-Fay, Glenn L. Butterfoss, Jack Snoeyink, and Brian Kuhlman. Maintaining solvent accessible surface area under rotamer substitution for protein design. Journal of Computational Chemistry, 28(8):1336–1341, 2007.

[61] Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis, Oliver F. Lange, James Thompson, Ron Jacak, Kristian Kaufman, P. Douglas Renfrew, Colin A. Smith, Will Sheffler, Ian W. Davis, Seth Cooper, Adrien Treuille andDaniel J. Mandell, Florian Richter, Yih-En Andrew Ban, Sarel J. Fleishman, Jacob E. Corn, David E. Kim, Sergey Lyskov, Monica Berrondo, Stuart Mentzer, Zoran Popović, James J. Havranek, John Karanicolas, Rhiju Das, Jens Meiler, Tanja Kortemme, Jeffrey J. Gray, Brian Kuhlman, David Baker, and Philip Bradley. Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol, 487(545-574), 2011.

[62] U. Lessel and D. Schomburg. Similarities between protein 3-D structures. Protein Eng., 10:1175–87, 1994.

[63] Min-Jun Li, Harry M. Greenblatt, Orly Dym, Shira Albeck, Adi Pais, Chidambaram Gunanathan, David Milstein, Hadassa Degani, and Joel L. Sussman. Structure of estradiol metal chelate and estrogen receptor complex: The basis for designing a new class of selective estrogen receptor modulators. Journal of Medicinal Chemistry, 54:3575–3580, 2011.

[64] Yi Liu and Brian Kuhlman. Rosettadesign server for protein design. Nucleic Acids Research, 34:W235–W238, 2006.

[65] Bradley C. Logsdon, John F. Vickrey, Philip Martin, Gheorghe Proteasa, Jay I. Koepke, Stanley R. Terlecky, Zdzislaw Wawrzak, Mark A. Winters, Thomas C. Merigan, and Lasislau C. Kovari. Crystal structures of a multdrug-resistant human immunodeficiency virus type

1 protease reveal an expanded active-site cavity. Journal of Virology, 78(6):3123–3132, March 2004.

[66] Rajarshi Maiti, Gary H. Van Domselaar, Haiyan Zhang, and David S. Wishart. Superpose: a simple server for sophisticated structural superposition. Nucleic Acids Research, 32:W590–W594, 2004.

[67] Paul Marakis and Martin Karplus. Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase. J. Mol. Biol., 352:807–822, 2005.

[68] William E. Meador, Anthony R. Means, and Florante A. Quiocho. Target enzyme recognition by calmodulin: 2.4$\mathring{A}$ structure of calmodulin-peptide complex. Science, 257:1251–1255, August 1992.

[69] Raul Mendez, Raphael Leplae, Leonardo De Maria, and Shoshana J. Wodak. Assessment of blind predictions of protein-protein interactions: Current status of docking methods. Proteins, 52:51–67, 2003.

[70] Mark B. Mixon, Ethan Lee, David E. Coleman, Albert M. Berghuis, Alfred G. Gilman, and Stephen R. Sprang. Tertiary and quaternary structural change in $G_{i\alpha 1}$ induced by GTP hydrolysis. Science, 270(954-960), November 1995.

[71] Christoph W. Müller and Georg. E. Shulz. Structure of the complex between adenlyate kinase from *Escherichia coli* and the inhibitor Ap_5A refined at 1.9$\mathring{A}$ resolution. J. Mol. Biol., 224:159–177, 1992.

[72] CW Müller, GJ Schlauderer, J Reinstein, and GE Schulz. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure, 4(2):147–156, 1996.

[73] Hannes Neuweiler, Soren Doose, and Markus Sauer. A microscopic view of miniprotein folding: Enhanced folding efficiency through formation of an intermediate. PNAS, 102(46):16650–16655, 2005.

[74] Wataru Nishima, Guoying Qi, Steven Hayward, and Akio Kitao. DTA: dihedral transition analysis for characterization of the effects of large main-chain dihedral angle changes in proteins. Bioinformatics, 25(5):628–635, 2009.

[75] Angel R. Ortiz, Charlie E.M. Strauss, and Osvaldo Olmea. MAMMOTH (matching molecular models obtained from theory): An automated method for model comparison. Protein Science, 11(11):2606–2621, 2002.

[76] Ludovic R. Otterbein, Jolanta Kordowska, Carlos Witte-Hoffmann, C.-L. Albert Wang, and Roberto Dominguez. Crystal structures of S100A6 in the CA^{2+}-free and CA^{2+}-bound states: The calcium sensor mechanism of S100 proteins revealed at atomic resolution. Structure, 10:557–567, 2002.

[77] Olayinka A. Oyeyemi, Kevin M. Sours, Thomas Lee, Amnon Kohen, Katheryn A. Resing, Natalie G. Ahn, and Judith P. Klinman. Comparative hydrogen-deuterium exchange for a mesophilic vs thermophilic dihydrofolate reductase at 25 °C: Indentification of a single active site region with enhanced flexibility in the mesophilic protein. Biochemistry, 50:8251–8260, 2011.

[78] Jay Painter and Ethan A. Merritt. Optimal description of a protein structure in terms of multiple groups underging TLS motion. Acta Crystallographica Section D, D62:439–450, 2006.

[79] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, III T.E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comp. Phys. Commun., 91:1–41, 1995.

[80] Bin Y. Qin, Lawrence K. Creamer, Edward N. Baker, and Geoffrey B. Jameson. 12-bromododecanoic acid binds inside the calyx of bovine β-lactoglobulin. FEBS Letters, 438:272–278, 1998.

[81] Predrag Radivojac, Zoran Obradovic, David K. Smith, Guang Zhu, Slobodan Vucetic, Celeste J. Brown, J. David Lawson, and A. Keith Dunker. Protein flexibility and intrinsic disorder. Protein Science, 13:71–80, 2004.

[82] Alexander A. Rashin, Abraham H. L. Rashin, and Robert L. Jernigan. Protein flexibility: coordinate uncertainties and interpretation of structural differences. Acta Crystallographica Section D, 65:1140–1161, 2009.

[83] Jane B. Reece, Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky, Robert B. Jackson, and Neil A. Campbell. Campbell Biology. Pearson Benjamin Cummings, San Francisco, CA, ninth edition, 2011.

[84] Jane S. Richardson. Staph nuclease B & W shaded ribbon. http://commons.wikimedia.org/wiki/User:Dcrjsr/gallery_of_protein_structure, 1980.

[85] Jane S. Richardson and David C. Richardson. The *de novo* design of protein structures. TIBS, 14:304–309, 1989.

[86] Robert C. Robinson, Marisan Mejillano, Vincent P. Le, Leslie D. Burtnick, Helen L. Yin, and Senyon Choe. Domain movement in gelsolin: A calcium-activated switch. Science, 286:1939–1942, 1999.

[87] Deanne W. Sammond, Dustin E. Bosch, Glenn L. Butterfoss, Carrie Purbeck, Mischa Machius, David P. Siderovski, and Brian Kuhlman. Computational design of the sequence and structure of a protein-binding peptide. J Am Chem Soc, 133(12):4190–4192, March 2011.

[88] Liliana Santamaria-Kisiel, Anne C. Rintala-Dempsey, and Gary S. Shaw. Calcium-dependent and -independent interactions of the S100 protein family. Biochemistry, 396:201–214, 2006.

[89] Michael Schneider, Xiaoran Fu, and Amy E. Keating. X-ray vs. NMR structures as templates for computational protein design. Proteins, 77(1):97–110, 2009.

[90] Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.3. August 2010.

[91] Irina F. Sevrioukova, Huiying Li, Hong Zhang, Julian A. Peterson, and Thomas L. Poulos. Structure of a cytochrome P450-redox partner electron transfer complex. PNAS, 96:1863–1868, 1999.

[92] Yury E. Shapiro, Michael A. Sinev, Elena V. Sineva, Vitali Tugarinov, and Eva Meirovitch. Backbone dynamics of *escherichia coli* adenylate kinase at the extreme stages of the catalytic cycle studied by N NMR relaxation. Biochemistry, 39:6634–6644, 2000.

[93] Maxim Shatsky, Ruth Nussinov, and Haim J. Wolfson. FlexProt: Alignment of flexible protein structures without a predefinition of hinge regions. Journal of Computational Biology, 11(1):83–106, 2004.

[94] Robert G. Smock and Lila M. Gierasch. Sending signals dynamically review. Science, 324:198–203, April 2009.

[95] Karsten Suhre and Yves-Henri Sanejouand. ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Research, 32:W611–W614, 2004.

[96] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[97] Douglas L. Theobald and Deborah S. Wuttke. Theseus: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics, 22(17):2171–2172, 2006.

[98] Douglas L. Theobald and Deborah S. Wuttke. Accurate structural correlations from maximum likelihood superpositions. PLoS Computational Biology, 4(2):e43, 2008.

[99] Thomas B. Thompson, Michael G. Thomas, Jorge C. Escalante-Semerena, and Ivan Rayment. Three-dimensional structure of adenosylcobinamide kinase/adenosylcobinamide phosphate guanylyltransferase (CobU) complexed with GMP: Evidence for a subtrate-induced transferase active site. Biochemistry, 38:12995–13005, 1999.

[100] M.F Thorpe, Ming Lei, A. J. Rader, Donald J. Jacobs, and Leslie A. Kuhn. Protein flexibility and dynamics using constraint theory. Journal of Molecular Graphics and Modelling, 19:60–69, 2001.

[101] Michael D. Tyka, Daniel A. Keedy, Igemar Andre, Frank Dimaio, Yifan Song, David C. Richardson, Jane S. Richardson, and David Baker. Alternate states of proteins revealed by detailed energy landscape mapping. J. Mol. Biol., 405:607–618, 2011.

[102] Claire Vieille and Gregory J. Zeikus. Hyperthermophilic enzymes: Sources, uses, and molecular mechanisms for thermostability. Microbiol. Mol. Biol. Rev., 65(1):1–43, 2001.

[103] Donald Voet and Judith G. Voet. Biochemistry. John Wiley and Sons, Inc., U.S.A., 3rd edition, 2004.

[104] Karl Volz and Philip Matsumura. Crystal structure of *Escherichia coli* CheY refined at 1.7$\mathring{A}$ resolution. J. Biol. Chem., 266(23):15511–15519, 1991.

[105] Qian Wang, Kao-Chen Liang, Arkadiusz Czader, M. Neal Waxham, and Margaret S. Cheung. The effect of macromolecular crowding, ionic strength and calcium binding on calmodulin dynamics. PLoS Computational Biology, 7(7):1–16, 2011.

[106] Sebastian Westenhoff, Elena Nazarenko, Erik Malmerberg, Jan Davidsson, Gergely Katona, and Richard Neutze. Time-resolved structural studies of protein reaction dynamics: a smorgasbord of x-ray approaches. Acta Crystallographica Section A, A66:207–219, 2010.

[107] James C. Whisstock and Stephen P. Bottomley. Molecular gymnastics: serpin structure, folding and misfolding. Current Opinion in Structural Biology, 16:761–768, 2006.

[108] Magnus Wolf-Watz, Vu Thai, Katherine Henzler-Wildman, Georgia Hadjipavlou, Elan Z. Eisenmesser, and Dorothee Kern. Linkage between dynamics and catalysis in a thermophilic-mesophilic pair. Nature structural and molecular biology, 11(10):946–949, 2004.

[109] S. Y. Wu, M. D. Perez, P. Puyol, and L. Sawyer. β-lactoglobulin binds palmitate within its central cavity. J. Biol. Chem., 274:170–174, 1999.

[110] Yuzhen Ye and Adam Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics, 19(Suppl. 2):ii246–ii255, 2003.

[111] Ada Yonath. X-ray crystallography at the heart of life science. Current Opinion in Structural Biology, 21:1–5, 2011.

[112] Hua Zhang, Tuo Zhang, Ke Chen, Shiyi Shen, Jishou Ruan, and Lukasz Kurgan. On the relation between residue flexibility and local solvent accessibility in proteins. Proteins, 76:617–636, 2009.

[113] Tuo Zhang, Eshel Faraggi, and Yaoqi Zhou. Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction. Proteins, 78:3353–3362, 2010.

Appendix A

Supplementary Information

We supply a number of figures for comparing mobility scores versus different attributes of the structure or energy of the protein. These figures are described in Chapter 5, and have weaker correlations than the correlation figures (for shape pliability) depicted within the chapter.

Figure A.1: Mobility score percentages for the total number of atomic neighbors per residue, measured by distance cutoff of $5\mathring{A}$ from the residue's C_α atom.

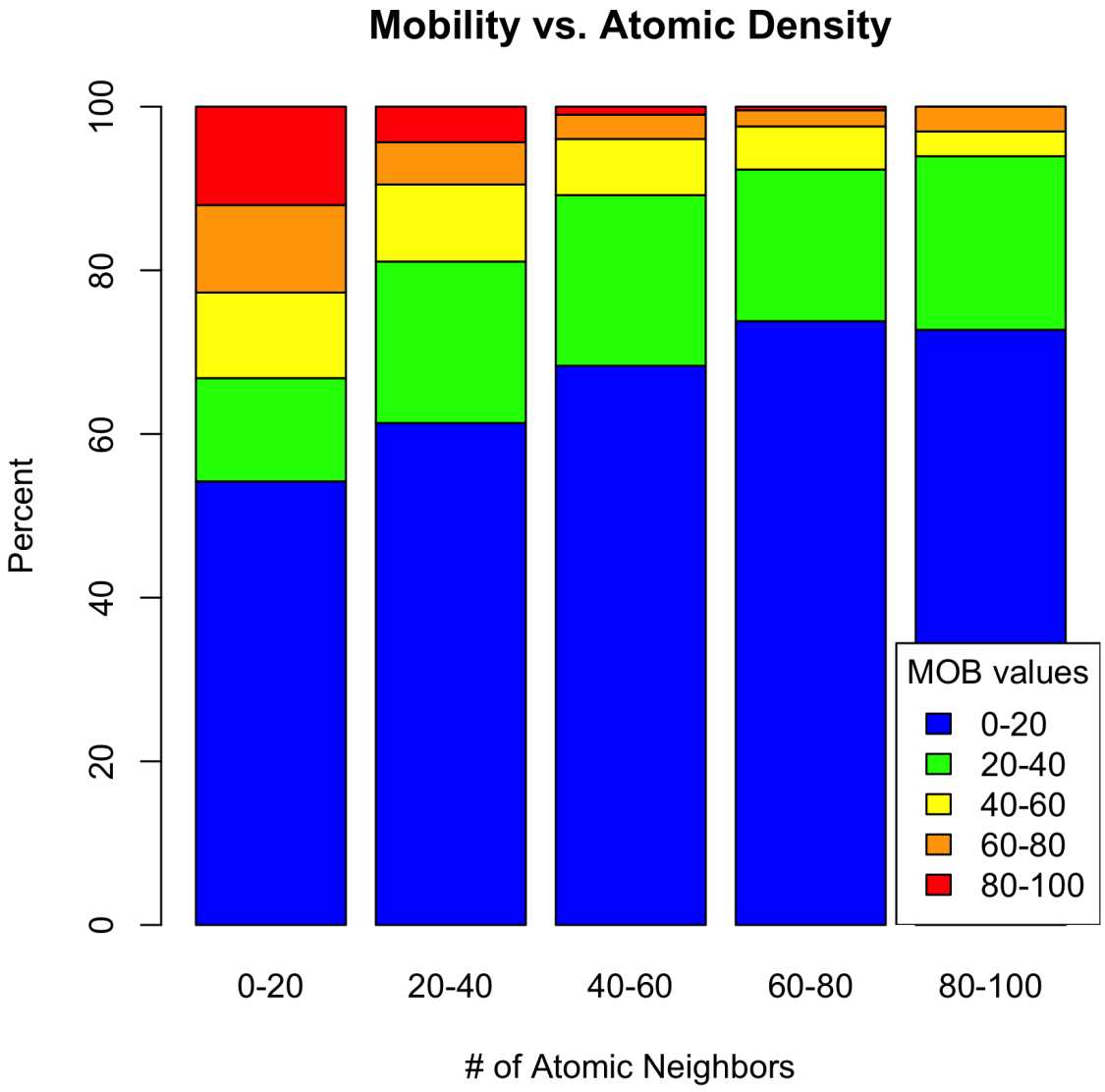


Figure A.2: Mobility score percentages for the ranges of the number of aromatic residue neighbors per residue.

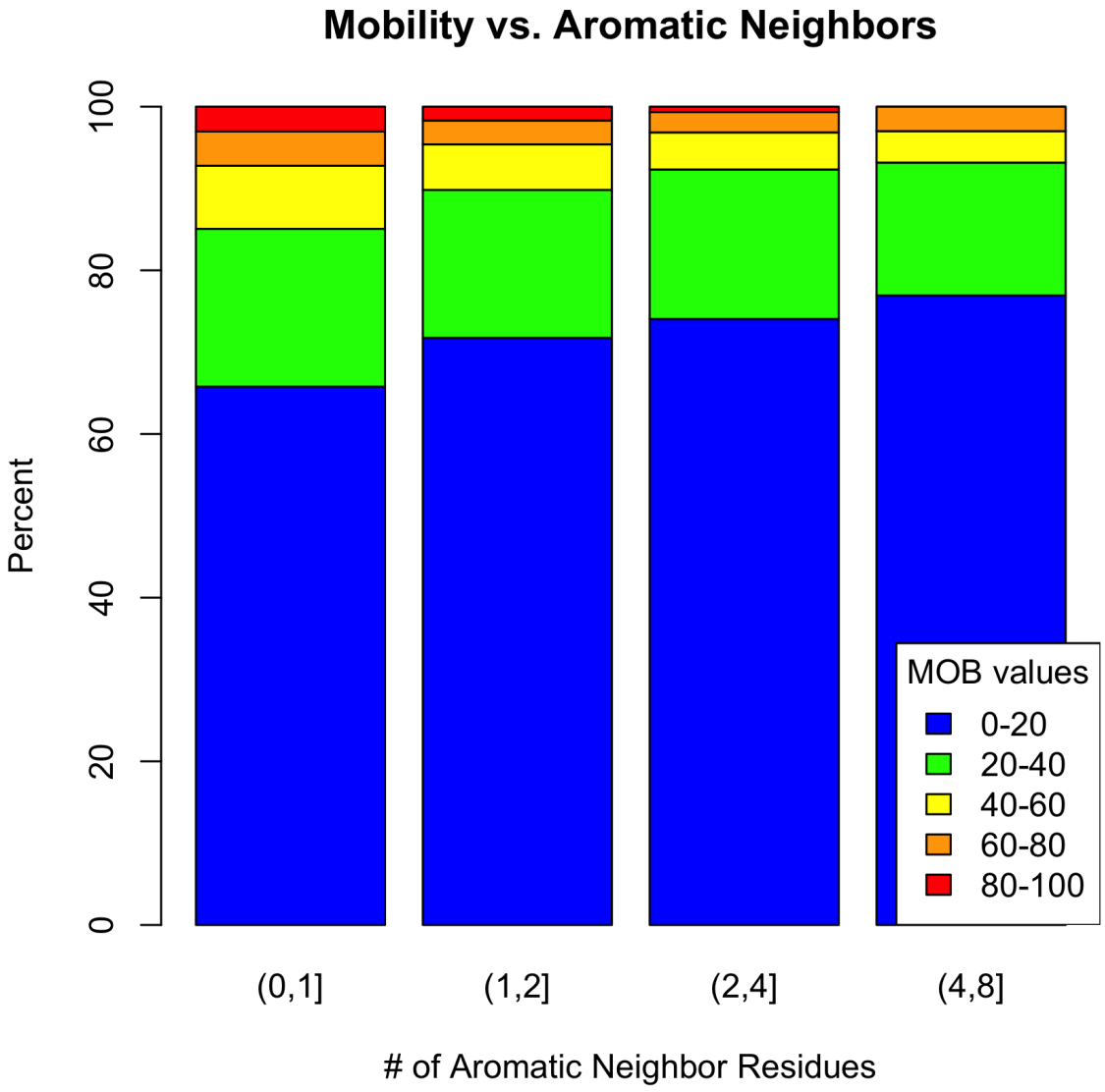


Figure A.3: Mobility score percentages for the ranges of the number of nonpolar residue neighbors per residue.

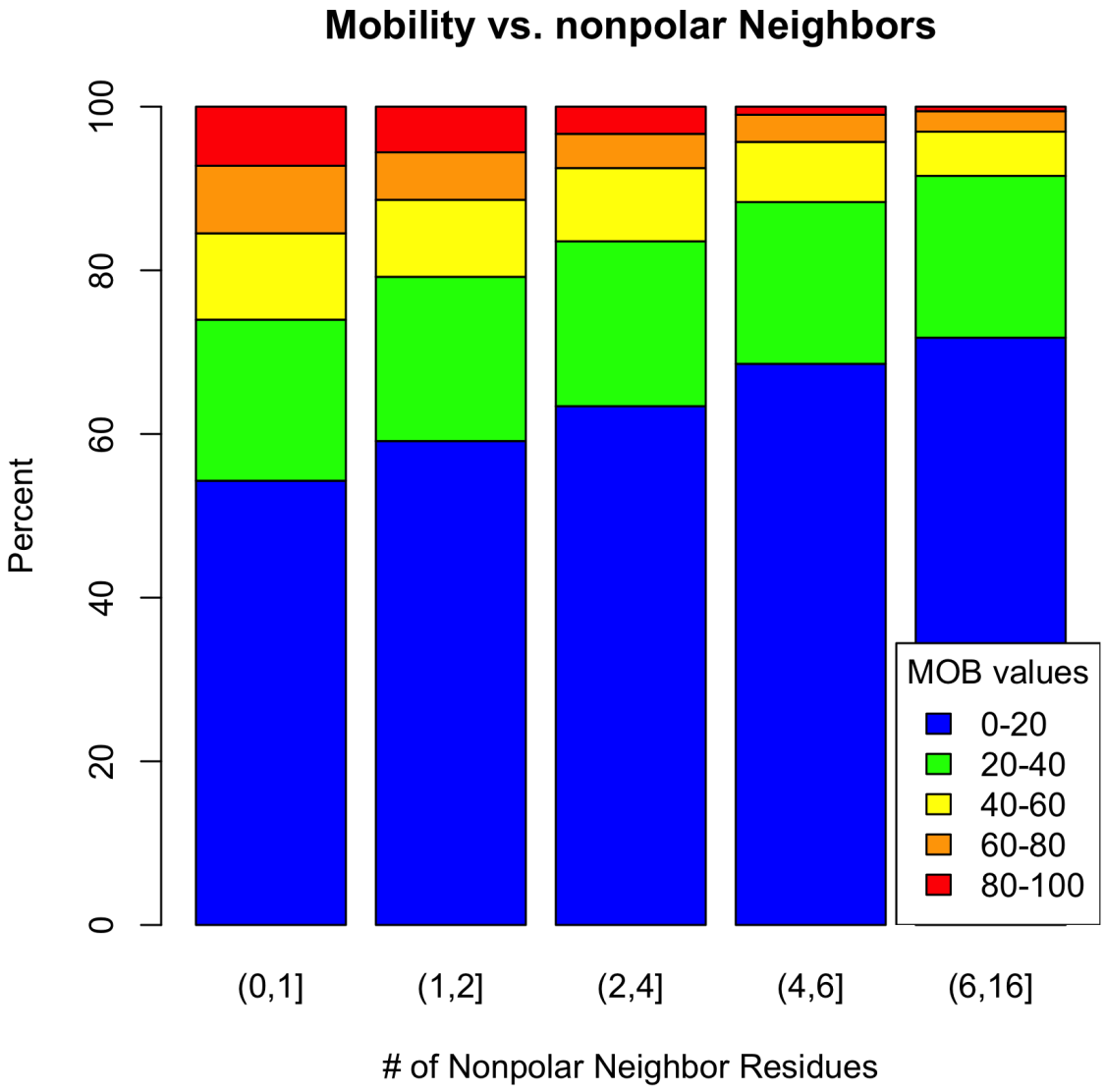


Figure A.4: Mobility score percentages for the ranges of the number of nonpolar neighbors per residue.

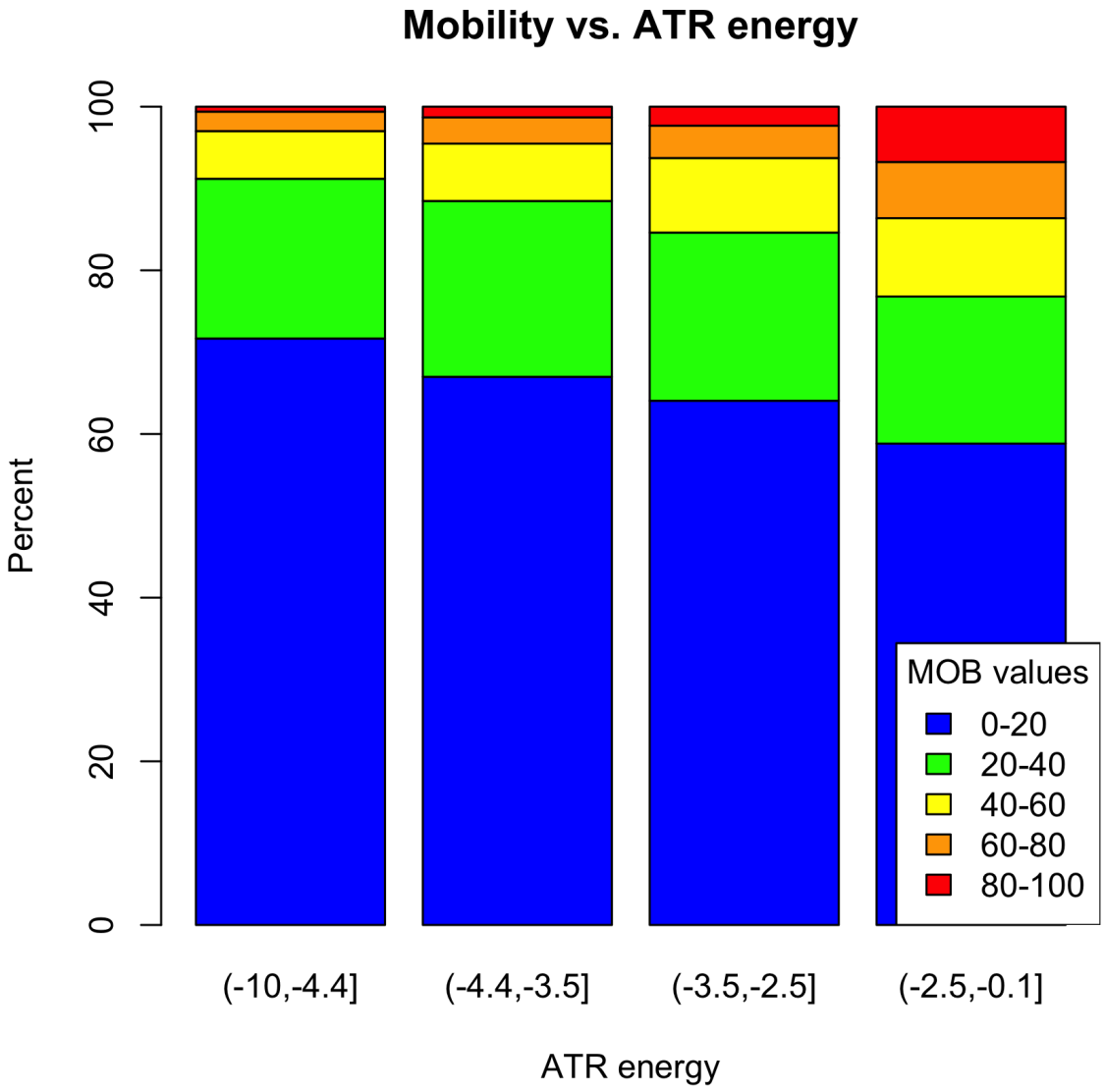


Figure A.5: Mobility score percentages for the ranges of hydrogen bond energy per residue.