ELUCIDATING THE EFFECTS OF MUTATION AND EVOLUTIONARY DIVERGENCE UPON PROTEIN STRUCTURE QUANTITATIVE STABILITY/FLEXIBILITY RELATIONSHIPS

by

Deeptak Verma

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2012

Approved by:

_____
Dr. Dennis R. Livesay

_____
Dr. Donald J. Jacobs

_____
Dr. Jun-tao Guo

_____
Dr. Joanna Krueger

ABSTRACT

DEEPTAK VERMA. Elucidating the effects of mutation and evolutionary divergence upon protein structure quantitative stability/flexibility relationships.
(Under the direction of DR. DENNIS R. LIVESAY and DR. DONALD J. JACOBS)

The importance of flexibility and stability on protein function has been recognized for over five decades. A protein must be flexible enough to mediate a reaction pathway, yet rigid enough to achieve high fidelity in molecular recognition. To understand these relationships, the main focus of our research has been a comparative investigation of proteins' dynamics and thermodynamics across both "depth" and "breadth". Specifically, we compare stability and flexibility properties across a set of human c-type lysozyme point mutations (depth), as well as across a set of functionally related β-lactamase protein orthologs (breadth). To accomplish these tasks we employ a Distance Constraint Model (DCM), which provides a robust statistical mechanical description of proteins and the relationships therein. The DCM is based on network rigidity that provides mechanical mechanism for enthalpy-entropy compensation, from which Quantitative Stability/Flexibility Relationships (QSFR) can be calculated. Our results suggest that DCM can be used for predicting stability of proteins with an average percent error of 4.3%. Deciphering changes in flexibility, DCM results suggest that the influence of mutations can lead to frequent, large and long-range effects in protein dynamics. Our breadth analyses indicate that QSFR and physiochemical property characterization of orthologs in a protein family parallel evolutionary relationship. Going further, we present protocols for clustering protein structures using their QSFR properties, thus paving way for comprehensive quantitative stability/flexibility

relationship analysis across protein families and superfamilies. To summarize, the results presented in this work provide a complete description of proteins that account for their stability, flexibility and function.

DEDICATION

For my parents and teachers who introduced me to science.

ACKNOWLEDGMENT

First and foremost, I would like to thank my advisors, Dr. Dennis R. Livesay and Dr. Donald J. Jacobs for their valuable guidance and patience over the years. I would also like to thank the members of our bio-molecular physics group, both past and present, for all the helpful conversations and support. Sincere thanks to my committee members Dr. Jun-tao Guo and Dr. Joanna Krueger for their time and valuable suggestions. I would like to appreciate UNC-Charlotte and my advisors for providing financial support. I would also like to express my appreciation to the administrative staff of ISSO, Bioinformatics department and URC for their friendly help. Last but not the least, I would like to thank my family and friends for their unflagging belief in me.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

DCM	Distance Constraint Model
mDCM	minimal Distance Constraint Model
QSFR	Quantitative stability/flexibility relationships
H-bond	Hydrogen bond
DHA	Dihedral angle
FI	Flexibility index
CC	Cooperativity correlation
DOF	Degrees of freedom
DSC	Differential scanning calorimetry
PDB	Protein data bank
HL	Human lysozyme
HEWL	Hen-egg white lysozyme
BL	β-Lactamase
EM	Expectation maximization
RMSD	Root mean square deviation

CHAPTER 1: INTRODUCTION

1.1 Importance of protein flexibility and stability

Proteins are large, complex, three-dimensional macromolecular structures consisting of many covalent bonds and noncovalent interactions that govern its stability and function [1-3]. Its functional specificity requires conformational flexibility and thermodynamic stability, i.e., a protein must be flexible enough to mediate a reaction pathway, yet rigid enough to achieve high fidelity in molecular recognition. Thereby, the importance of protein flexibility and its relationship with stability on protein function has been recognized for over fifty years [4-6]. Complex computer algorithms are required to understand the detailed biophysical and biochemical properties that elucidate structure/function relationships. Developing fast and robust biophysical models to accurately predict flexibility and stability under given thermodynamic and solvent conditions has also been a long-term challenge [7, 8]. This study employs the use of a novel biophysical model, called the Distance Constraint Model (DCM) [9, 10], which provides robust statistical mechanical description of protein both and the relationships therein. The DCM is based on network rigidity that provides mechanical mechanism for enthalpy-entropy compensation, from which Quantitative Stability/Flexibility Relationships (QSFR) can be calculated.

1.2 Distance Constraint Model

A complete description of protein stability should ideally account for a wide variety

of chemical interactions, including: covalent bonding (i.e., bond stretching, angle bending and torsional effects), nonbonded interactions (i.e., long and short range ionic interactions, hydrogen bonds, dipole interactions and van der Waals contacts), solvation, etc. Most of which are affected by solvent pH, ionic strength and other co-solute concentrations. Standard simulation methods, such as molecular dynamics, attempt to describe most of the above terms, but rarely all. The "rules" constraining the simulation are based on energetic potentials; however, free energies are the primary protein stability metric of interest. Free energies are derived from a simulation *post priori* using the method of thermodynamic integration [11]. The primary advantage of simulation methods is that they are nearly chemically and physically complete. However, simulation is extremely computationally intensive, making it prohibitive for large-scale analyses.

In response to the immense computational cost of molecular simulations, the DCM was developed in 2001 from conception to optimally balance computational efficiency with prediction accuracy by uniquely integrating mechanical and thermodynamic viewpoints of macromolecular structure, and has been improvised since. The DCM is based on a free energy functional that decomposes the total free energy into constituent parts related to specific types of interactions. While total enthalpies can be calculated from the sum of the individual components, adding entropies over all components generally will overestimate conformational entropy [12, 13]. However, the utility of a free energy decomposition is restored using DCM [14], where conformational entropy is additive over independent degrees of freedom (DOF) that are robustly identified by the model. Herein, structure is recast as a topological network of distance constraints. Each constraint within the topological framework (network) is associated with a component

enthalpy and entropy value. The conformational part of the free energy of a given framework, $G(f)$, is reconstituted from the free energy decomposition that defines the types of interactions modeled as distance constraints. The free energy is calculated from the total enthalpy and entropy of a framework by:

$$G_{cnf}(f) = \sum_{t}^{N_{int}} h_t N_t(f) - RT \sum_{t}^{N_{int}} \sigma_t I_t(f)$$

where, N_{int} is the number of different types of modeled interactions, h_t is the enthalpy of interaction t, N_t is the number of times interaction type t occurs within framework f, σ_t is the pure entropy of a single distance constraint used to model interaction type t, R is the ideal gas constant, and I_t is the number of number of independent constraints of type t, where I_t is always less than or equal to N_t. Note: for ease of interpretation and consistency, enthalpy parameters are depicted using Roman characters, whereas entropies are depicted by Greek symbols.

The salient feature within the conformational free energy calculation of a framework is that total entropy is summed over a set of independent constraints, which are determined using efficient network rigidity graph algorithms [15, 16]. Starting from $3N_a$ DOF (N_a is the number of atoms), each constraint within flexible portions of the network removes one degree of freedom. However, when an interaction is added to an already rigid substructure of the network, no further reduction in entropy occurs because all available DOF within that region have already been consumed. Since assignment of which constraint is independent or redundant is not unique, the expression for conformational entropy is also not unique. This approach provides an upper bound estimate to conformational entropy regardless which set of independent constraints are considered. However, by adding constraints as dictated by an entropy spectrum [17] that

preferentially orders them from smallest to largest entropy, a rigorous lowest upper bound is obtained. Note that a given chemical interaction can be modeled by more than one constraint. For example, covalent bonds are modeled as five constraints, H-bond as three and a torsion force is modeled as one (Figure 1.1) [9].

If thermal fluctuations did not occur, the free energy of a given protein would simply be based on the above calculation using the native state structure, but this is of course not the case. While covalent bonding is appropriately described by a large set of quenched constraints that are present within each microstate of the ensemble, fluctuating constraints

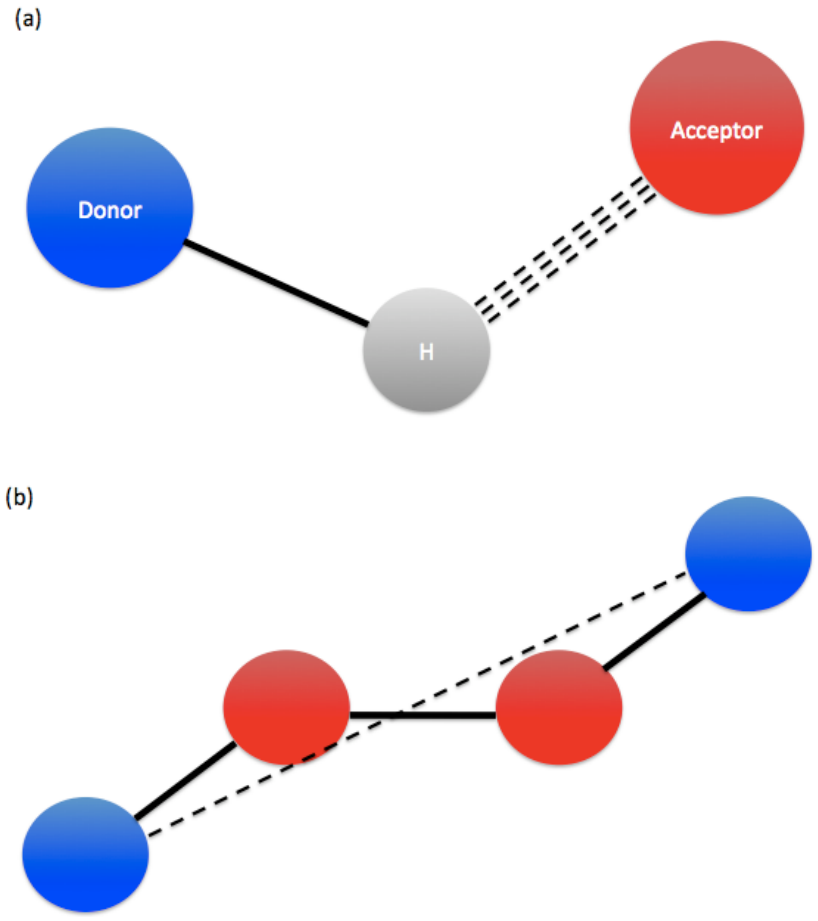

Figure 1.1: Two types of interactions are modeled in DCM: (a) Hydrogen bonds and (b) Torsion forces. Hydrogen bonds are modeled as three constraints whereas torsion force is modeled as one.

account for the forming and breaking of weak interactions that are critical to properly describe equilibrium behavior. Herein, we consider a 'minimal' set of fluctuating interaction types, specifically N_{int} = 2 types are considered: hydrogen bonds and torsion angle forces. Within this minimal DCM (mDCM), all possible hydrogen bonds (H-bonds) are defined by the native structure. H-bond enthalpies, h_{hb}^{pot}, are calculated from the native structure using the empirical potential from Dahiyat *et al.* [18]. Salt bridges are modeled as a special case of H-bonds.

The entropic cost of forming an intramolecular H-bond is linearly related to h_{hb}^{pot}, whose slope is defined by the parameter γ_{max}. Solvation terms are described through H-bonds to solvent; when an intramolecular H-bond breaks, there is a compensating reduction enthalpy given by the fitting parameter u_{sol}. As a consequence, the net effect of each intramolecular H-bond is given by $h_{hb}^{net} = h_{hb}^{pot} - u_{sol}$. While its less immediately obvious, the reduction in entropy associated with torsional effects can also be modeled using distance constraints. Here, we introduce constraints across all *i* to *i*+3 atomic pairs, which includes side chain torsions. The torsions are segregated in an Ising-like manner where *native* torsions are associated with enthalpy and entropy values $\{v_{nat}, \delta_{nat}\}$ and *disordered* torsions are associated with $\{v_{dis}, \delta_{dis}\}$. Two important "minimal" aspects of the mDCM are that: (*i.*) other than h_{hb}^{pot}, all parameter values are treated phenomenologically; and (*ii.*) all parameters are treated universally regardless of residue type. The disordered dihedral angle enthalpy, v_{dis}, is our reference energy, which is defined as zero (Table 1.1).

Hydrophobic considerations are the most severe omission from our current free energy decomposition scheme. The hydrophobic effect is a bulk colligative property

Table 1.1: Free fitting parameters used by the mDCM. The net intramolecular H-bond enthalpy is calculated as $h_{hb}^{net} = h_{hb}^{pot} - u_{sol}$, where h_{hb}^{pot} is calculated from an empirical potential.

Interaction	Parameter	Treatment	Description
H-bonds	h_{hb}^{pot}	Empirical potential	Intramolecular H-bond enthalpy
	γ_{max}	Constant	H-bond pure entropy is linearly related to h_{hb}^{pot} whose slope is controlled by γ_{max}
	u_{sol}	Fitting	H-bond to solvent enthalpy[1]
Native torsion	δ_{nat}	Fitting	Native torsion angle pure entropy
	v_{nat}	Fitting	Native torsion angle enthalpy
Disordered torsion	δ_{dis}	Constant	Disordered torsion angle pure entropy
	v_{dis}	Constant	Disordered torsion angle enthalpy

related to an increase in the number of accessible DOF upon segregation of polar and nonpolar solvents. As such, they do not directly map to a set of distance constraints. For example, evem molecular dynamics cannot directly model the hydrophobic effect because it is not explicitly included within the simulation "rules" defined by molecular mechanical force fields. Therein, the hydrophobic effect only emerges after thermodynamic integration of the trajectory phase space.

Within the mDCM, hydrophobic interactions are indirectly included by two phenomenological terms that connect to order parameters describing the number of constraints within the system. This approach works well, and is tied to the observation that hydrophobic contacts track H-bond formation [19], meaning that our phenomenological H-bond parameters implicitly account for hydrophobic interactions. Thus, as we have discussed previously [20], the u_{sol} and v_{nat} parameters implicitly account for the hydrophobic effect.

Even with such a simple model, an exact calculation of the partition function for proteins is impossible due to an astronomical number of possible frameworks. As such, a

heterogeneous mean field approach (Figure 1.2) has been developed to make the calculation tractable [9, 10]. Combining all the contributions described above, we arrive at the following free energy functional:

$$G(N_{hb}, N_{nat}) = U_{hb}(N_{hb}) - N_{hb}u_{sol} + N_{nat}v_{nat} - RTS_{cnf}(N_{hb}, N_{nat} \mid \delta_{nat}, \delta_{dis}, \gamma_{max}) - RTS_{mix}(N_{hb}, N_{nat})$$

This functional has five adjustable parameters (depending on solvent conditions and protein fold). However, from our previous work, γ_{max} and δ_{dis} have been fixed and are treated as transferable parameters, leaving only $\{\delta_{nat}, v_{nat}, u_{sol}\}$ as free parameters within the mDCM (Table 1.1). We have found that the three-free parameter mDCM provides a high degree of accuracy and robustness in predicting protein stability [10]. Typically the parameterization is been obtained by finding the appropriate parameter values to

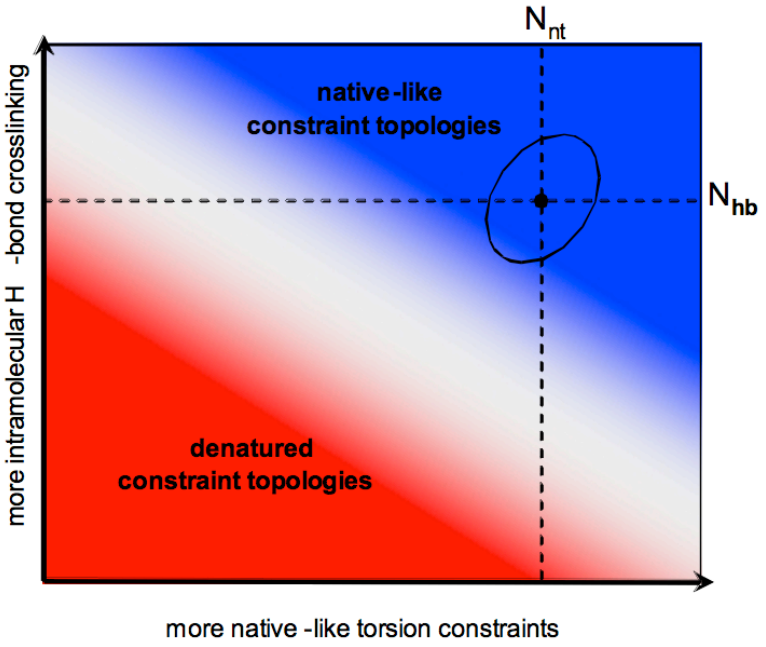

Figure 1.2: The free energy of each macrostate is calculated using a hybrid mean-field approximation by Monte Carlo sampling. N_{nt} is the number of native torsion interactions and N_{hb} is the number of hydrogen bonds. The protein explores the red area under disordered constraint topology and blue under the native constraint topology. White area indicates high-energy barrier or the transition from disordered to native constraint topology. Figure as published in [10].

reproduce experimental C_p curves from differential scanning calorimetry (DSC) using simulated annealing. These mDCM parameters are physically meaningful with ranges that are remarkably tight over a diverse set of proteins. This modeling approach has been found to provide accurate description of both thermodynamics and intrinsic flexibility in proteins in many applications [9, 10, 20-23]. Once parameterized for a given protein, the mDCM can be used for calculating quantitative stability/flexibility relationships (QSFR), which covers a broad domain of investigating protein thermodynamics, dynamics and functional relationships.

1.3 Dissertation objective and layout

The main objective of this dissertation has been to perform comparative QSFR investigations across many different proteins. To define our objectives clearly we have categorized this dissertation into two parts, "depth" and "breadth" analysis (Figure 1.3).

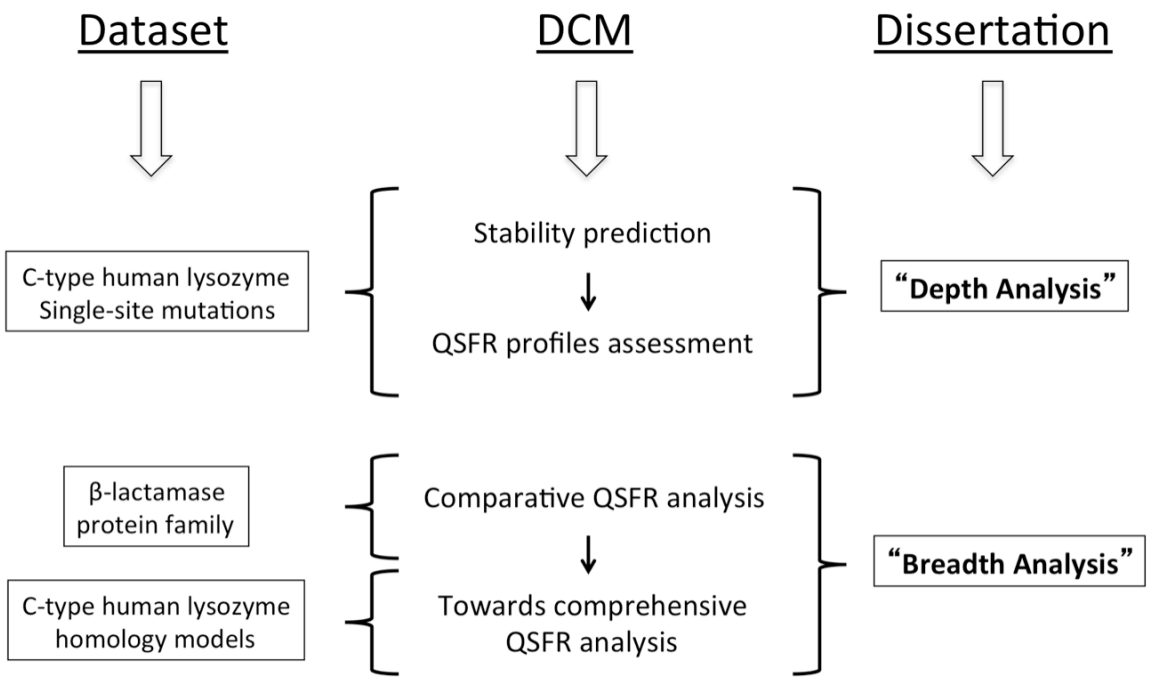

Figure 1.3: Workflow elucidating dissertation breakdown broadly classified as "depth" and "breadth" analysis.

We have compared QSFR properties across a set of single site point mutation proteins (depth), as well as across a set of functionally related orthologs (breadth). In the former, we have attempted to answer the following questions: (a) How does mutation affect stability and flexibility profiles? (b) Are the stability and flexibility changes upon mutation local or global? (c) Can a mutation improve protein stability without compromising functional efficiency? In the breadth analysis, we have analyzed QSFR and physiochemical properties across a protein family. Herein, we have tried to answer the following questions: (a) Is QSFR conservation an evolutionary driving force? (b) Do functional sites have conserved QSFR signatures? (c) Do physiochemical mechanisms change under different environmental conditions? Going further, a benchmark study has been performed that could help us expand the scope of comparative QSFR analysis towards protein families and superfamilies using homology models. Taken in its entirety, this dissertation research aims at drastically expanding our understanding of protein flexibility, stability and their relationships. A brief overview of the analyses performed is provided below.

1.3.1 Depth analysis

To predict relative stability of protein mutants: Most of our insight, regarding protein structure and stability relationship, comes from laborious experimental analyses that perturb protein structures via site-directed mutations. To save time and money, efficient computational models are desired to speed up such analyses. The main objective of this work is to employ the DCM in predicting melting temperatures (or stability) of c-type lysozyme mutants as accurately as possible. The idea employed here is to use contextual learning, i.e., using additional DCM best-fit parameters to boost statistics and achieve

better accuracy in melting temperature prediction. As published, the DCM predicts melting temperatures of proteins with an accuracy of more than 95% [3]. The results presented have been published in Current Protein and Peptide Science Journal [24].

To assess Quantitative Stability/Flexibility Relationships (QSFR) profiles of lysozyme mutants: The functional importance of protein dynamics is universally accepted, making the study of dynamical similarities and differences among proteins of the same function an intriguing problem. While some metrics are likely to be conserved across family, differences are also very common. This work investigates changes in dynamics occurring upon individual point mutations. Somewhat surprisingly, the small structural perturbations caused by mutation lead to changes throughout the protein. These changes can be quite large, actually surpassing the scale for differences between ortholog pairs. Moreover, changes in flexibility frequently occur at sites far from the mutation site. These results underscore the sensitivity of protein dynamics in connection with allostery, and help explain why differences across protein families are so common. These results have been published in *PLoS Computational Biology* [25].

1.3.2 Breadth analysis

To characterize similarity/variability within QSFR and physiochemical properties across a protein family: Comparison of protein sequences and structures sharing function has become a well-established bioinformatics paradigm, leading to countless discoveries related to protein family sequence/structure/function relationships. However, sequence and structure alone provide only crude physiochemical descriptions, thus stressing the need for more sophisticated analyses. In this work, we determine how much QSFR and electrostatic properties vary across the β-lactamase enzyme family. Our results indicate

that some properties are mostly conserved across the family, whereas others vary significantly despite the fact that all share the same high-level β-lactamase activity. Despite global variance in some metrics, systematic differences are frequently observed between evolutionary outgroups, indicating that physiochemical properties are simultaneously conserved and variable. As such, these results underscore the richness within physiochemical and QSFR properties across a protein family. Manuscript of this research work has been submitted to *PLoS Computational Biology Journal*.

To develop a homology modeling protocol to robustly predict QSFR properties for a comprehensive analysis: Expanding the scope of our QSFR investigation across protein families and superfamilies requires solved protein structures. However, developing homology modeling and assessment protocols can allow us to robustly calculate QSFR properties for unknown protein structures. But since QSFR changes are sensitive to subtle structure variations, designing a robust protocol for good model selection is of utmost importance before instigating comparative QSFR analysis across protein families with unknown structures. To benchmark our methodology, we generate an ensemble of homology models and assess them for accurate QSFR property prediction. Results suggest that clustering homology models based on common structural, thermodynamic and mechanical quantities can result in precise QSFR calculations, paving way for a comprehensive QSFR analysis across hundreds of proteins in a protein family. Manuscript of this work is under preparation and results will be submitted soon to a peer-review journal.

CHAPTER 2: PREDICTING THE MELTING POINT OF C-TYPE HUMAN LYSOZYME MUTANTS

2.1 Introduction

Due to the time and cost of molecular biology and biophysical experiments, accurate computational models to predict and explain the effects of point mutations on protein stability have long been desired. Despite some progress towards this goal, it remains largely an open computational biology problem. The majority of the successes thus far have been based on machine learning approaches (i.e., decision trees [26-28], support vector machines [29-31], and artificial neural networks [32-34]). While these methods can achieve impressive prediction accuracies, their interpretive utility spans a broad range (e.g., decision trees are somewhat interpretable, whereas artificial neural networks are not). And even under the best of circumstances, all of these empirical methods lack the descriptive power of first-principles calculations based on the underlying physics and chemistry. As such, we have been seeking to develop a biophysical calculation of protein stability.

Unfortunately, computational biophysics approaches have largely failed to achieve an expectable level of accuracy. The reason for this is that protein structures are extremely complicated, being dense networks of chemical interactions that lead to protein stabilities involving small differences between large free energy values. In fact, even misplacement of a single hydrogen atom is sufficient to render a computational model wrong [35]. The primary exception to these failures is stability predictions of solvent exposed mutations.

These mutations increase (or decrease) protein stability by optimizing (or destabilizing) long-range surface electrostatics. Because these mutations occur on the protein surface, which is less densely packed than the protein core, they are less susceptible to issues related to the intimate atomic details that frequently derail computational predictions within the core. As such, biophysical models that quantify surface electrostatics effects (i.e., Tanford-Kirkwood [36-38] and Poisson-Boltzmann theories [39-42]) have successfully reproduced experimental trends across several mutation sets and have provided explanations for a large number of confounding experimental results. When these methods fail, the origins of the mutant stability change are often generically explained as resulting from conformational changes, highlighting a fundamental limitation of nearly all biophysics-based stability prediction methods.

To address the problem of predicting mutant stability based on conformational considerations, we test herein the ability of our distance constraint model (DCM) to reproduce stability trends within human C-type lysozyme and 14 point mutations therein. As described in introduction, DCM is a phenomenological biophysics model that requires parameterization, usually done by fitting to experimental heat capacity, C_p, curves [43, 44]. After parameterization, the DCM quantifies both the enthalpic and entropic effects of all interactions within the protein, from which a wide variety of equilibrium thermodynamic quantities (i.e., C_p, free energy, T_m, etc.) are calculated. The DCM approach is applied to 15 considered lysozyme structures each with measured heat capacity. The best-fit parameters derived from one such determination are applied to the remaining structures to assess predictive power. Over all possible permutations, this process results in an impressive average error of 4.3% (standard deviation = 3.6%) in the

prediction of the experimental T_m's, a commonly used surrogate for stability [45]. This translates to a Pearson correlation coefficient of 0.64 for predicted to experimental ΔT_m values, which is among the best values ever presented for prediction of point mutation stability focusing on conformational effects. In this approach, multiple parameter sets are found that fit the data well in addition to the above mentioned best-fit parameters associated with the lowest least squares error. In the attempt to boost statistical prediction accuracy by incorporating variability in model parameterization from additional near-optimal parameter sets and using additional parameter sets from multiple mutant structures, we find that, surprisingly, there is no statistically significant change in the predictive accuracy on average. This result is important, as it indicates that the predictions from any DCM parameterization resulting in a reasonably good fit to heat capacity is robust to transferability across mutant structures.

2.2 Methods

2.2.1 Dataset preparation and simulated annealing

Lysozyme, which is abundant in egg whites and secretions (i.e., tears, saliva, milk, etc.), is a general class of enzymes that degrade bacterial cell walls through hydrolysis of β(1,4) glycosidic linkages. Members of the lysozyme superfamily share the same α+β structural motif within their active site region. Due to ease of production and characterization, human C-type lysozyme is a common model system for protein stability investigations. C_p curves for the human C-type lysozyme and 14 lysozyme mutant proteins have been obtained from published data [46-52]. Exact PDB codes are provided in Table 2.1. Since model parameters are dependent upon solvent conditions, only site directed mutants that are spatially distinct and characterized under the same experimental

Table 2.1: Thermodynamic characteristics and best-fit parameters

Mutation	PDBID	T_m (K)	$C_{p,max}$ (kcal/[mol·K])	δ_{nat} (unitless)	v_{nat} (kcal/mol)	u_{sol} (kcal/mol)
Wild-type	1LZ1	338.7	17.5	1.16	-0.16	-1.84
K1A	1C45	336.7	13.1	1.36	-0.20	-1.71
V2A	1OUG	333.4	16.8	1.12	-0.28	-1.76
Y38F	1WQO	337.7	18.8	1.08	-0.23	-1.69
Y45F	1WQP	337.4	18.5	1.32	-0.26	-1.77
Y54F	1WQQ	337.3	17.3	1.36	-0.29	-1.90
I56T	1OUA	325.1	14.8	1.32	-0.29	-1.87
Q58G	1B7R	345.3	19.0	1.24	-0.29	-1.89
I59S	2MEG	326.2	14.4	1.32	-0.37	-1.94
Y63F	1WQR	337.6	18.5	1.24	-0.25	-1.87
P71G	1LHI	336.1	20.3	1.28	-0.32	-2.11
V74A	1OUH	337.3	18.8	1.28	-0.23	-1.77
V100A	1OUB	337.1	18.2	1.28	-0.35	-1.91
P103G	1LHJ	338.6	18.2	1.32	-0.19	-1.77
Y124F	1WQM	337.6	19.0	0.88	-0.39	-1.78
	Average	336.1	17.5	1.24	-0.27	-1.84
	Variation	1.5%	11.3%	10.5%	24.3%	5.8%

conditions are considered here. Specifically, all 15 proteins were characterized under near identical buffering conditions (pH = 2.7-2.8) and salt concentrations (0.05 M). The pH range is $2.67 \leq \text{pH} \leq 2.8$ with an average of 2.71, and with a 0.032 standard deviation.

The extreme differences correspond to about 1 K difference in T_m. In addition, all of the considered C_p curves have been produced by the same group (Yutani *et al.*), which minimizes the risk that unforeseen factors (instrument, sample preparation, protein concentration estimates, etc.) are affecting the controls in experimental data. Note that, in practice, DSC is a notoriously difficult and finicky technique to perform [53], and in general it is difficult to find a large collection of systematic data. In order to look at the intrinsic variability in the DCM, and to justify our demand on the transferability of parameterization, it is important to have all the heat capacity measurements made under identical conditions. Fortunately, the methodical work by Yutani and co-workers presented us with the opportunity to consider a diverse collection of 14 point mutations that are structurally well distributed throughout the lysozyme structure, as shown in

Figure 2.1, which also indicates their solvent accessibility. Approximately half of the mutations are exposed to solvent, which as discussed above can be well described by long-range electrostatics models. However, we purposely omit an explicit long-range electrostatics component in the presented model to assess how well the mDCM does on its own. Incorporation of long-range electrostatics is expected to further improve model accuracy. To ensure proper ionization, the H++ server [54] is used to add hydrogen atoms to the structures as expected at pH 2.7 based on calculated pK_a values. The protonated structures are subsequently energy minimized prior to the simulated annealing fitting to determine $\{\delta_{nat}, \nu_{nat}, u_{sol}\}$ as done previously [43, 44]. Prior to fitting, all curves were shifted such that $C_{p,min} = 0$.

2.2.2 Prediction of Tm values

After best-fit parameters have been determined for each mutant structure and C_p curve pair, we attempt to answer the following question: *How well does the mDCM*

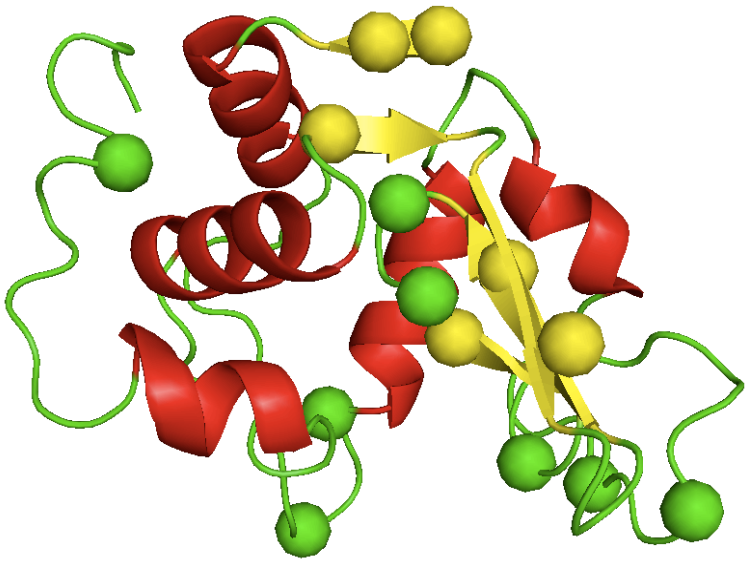

Figure 2.1: The human wild-type c-type lysozyme structure [1LZ1]. The CB atoms of the 14 mutated spatially distinct positions are highlighted. PDB molecule is color-coded by secondary structure.

reproduce T_m values on the remaining 14 structures using the best-fit parameters from the first? After considering all possible permutations, this thought experiment assesses how well, on average, the mDCM would do if only a single mutant structure and C_p curve pair were available before prediction on additional lysozymes. Meaning, we apply the best-fit parameters from mutant *i* to all of the remaining lysozyme structures. In each case, the predicted C_p curve is calculated, and its peak is used to identify the predicted T_m. We then collapse all 15 x 14 = 210 permutations into a single dataset and report average statistics.

As an extension, we also consider two scenarios where additional parameter sets are used in the prediction process. For example, we assess whether or not using the parameters from $n > 1$ 'training' lysozyme structures improves average prediction accuracy. Here, best-fit parameter sets from *n* different lysozymes are used to generate *n* different T_m predictions for a given target, which are simply averaged to give the final predicted value. However, complete enumeration of all possible permutations results in a combinatorial explosion ($>10^8$). Instead, for each target lysozyme, we have generated 100 random 14-character strings that simply list a unique identifier associated with each of the remaining lysozyme structures. For each value of $n \in \{1-14\}$, we include the first *n* lysozymes from the generated string within the 'training' set. The same 100 strings are used as we systematically consider all possible values of *n*. Over all 15 lysozyme structures, we determine the final predicted T_m value for each *n* by averaging over the 100 samples. This entire process was repeated ten times.

While only best fits have been discussed thus far, the simulated annealing procedure actually generates a large number of nearly optimal fits that are virtually

indistinguishable by visual inspection. Figure 2.2 plots kernel density functions generated using the R statistical package for the $m = 20$ best fit parameter sets for the wild-type structure. Across the parameters $\{\delta_{nat}, v_{nat}, u_{sol}\}$, the variation ranges from 2% to 24.2%. Similar results are observed for the lysozyme mutants. Related, Table 2.2 presents the percent variation in each parameter at three values of m (m = 4, 8, and 16).

Note that there is slight tendency for the variation to increase with m; however, there are several examples where the opposite occurs, highlighting the stochastic nature of finding a good parameter basin within the simulated annealing process. We assess whether or not using $m > 1$ of these near optimal fits improves average prediction accuracy. Meaning, for a given 'training' lysozyme mutant structure, we apply the m best parameters sets derived from it to each of the remaining structures, resulting in m different T_m predictions for each mutant trained on. As before, the final predicted T_m is simply the average over the m predictions for that structure. We consider each value of $m \in \{1-20\}$. Putting everything together, we consider m different parameter sets for

Table 2.2: Percent variation within the m best parameter sets.

		u_{sol} (kcal/mol)			v_{nat} (kcal/mol)			δ_{nat} (unitless)		
Mutation	**PDBID**	***m*=4**	***m*=8**	***m*=16**	***m*=4**	***m*=8**	***m*=16**	***m*=4**	***m*=8**	***m*=16**
Wild-type	1LZ1	1.9	1.3	1.8	20.9	17.3	23.2	12.7	10.4	14.1
Q58G	1B7R	0.5	0.6	1.4	5.3	5.5	9.9	5.6	6.3	10.4
K1A	1C45	2.1	1.5	4.1	11.3	9.2	12.0	1.7	1.5	10.9
P71G	1LHI	1.3	1.4	1.9	3.8	3.9	6.0	6.8	6.6	7.9
P103G	1LHJ	0.7	1.0	1.3	5.3	6.0	11.6	0.00	2.0	4.9
I56T	1OUA	2.1	2.8	5.0	9.7	8.4	10.4	12.9	13.8	15.3
V100A	1OUB	0.9	2.9	3.1	2.7	6.5	8.4	6.2	12.5	15.7
V2A	1OUG	0.3	2.2	2.7	2.8	3.7	9.2	2.1	11.9	13.7
V74A	1OUH	0.7	1.5	1.6	2.2	4.7	7.6	2.6	6.6	9.1
Y124F	1WQM	3.1	3.2	3.5	11.9	12.9	10.4	18.6	17.3	13.5
Y38F	1WQO	2.9	2.2	3.1	7.7	8.8	19.6	14.3	10.8	16.0
Y45F	1WQP	0.3	1.5	1.5	4.9	4.8	6.8	5.5	7.7	8.5
Y54F	1WQQ	1.6	1.3	1.3	4.6	3.8	8.5	3.1	5.7	11.5
Y63F	1WQR	0.8	0.8	1.2	4.0	4.8	5.2	1.6	2.4	3.7
I59S	2MEG	1.6	2.6	3.6	5.0	4.7	7.8	4.4	8.9	9.9

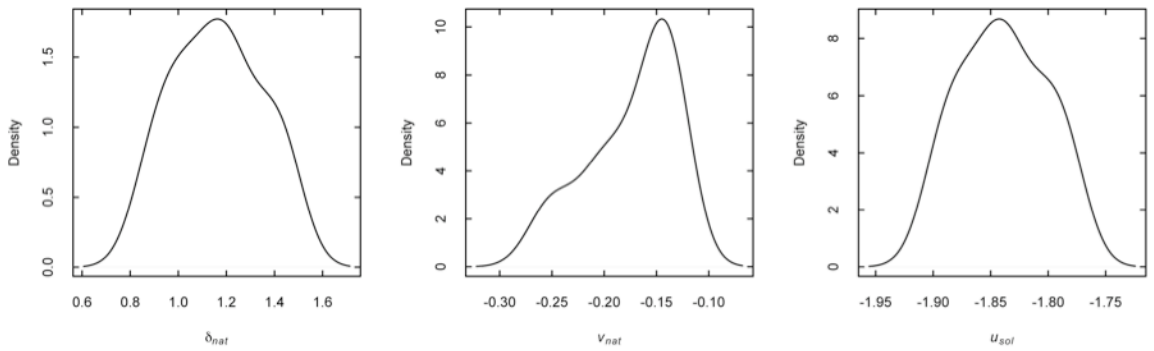

Figure 2.2: Kernel density functions for each of the three model parameters generated from the $m = 20$ best wild-type lysozyme parameter sets. The percent variation for each parameter $\{\delta_{nat}, v_{nat}, u_{sol}\}$ is 15.9, 2.0 and 24.2 percent, respectively. The density functions and percent variation values for the lysozyme mutants are similar (Table 2.2). The kernel density plots were generated using the R statistical package.

each of the n proteins used to train on. From these two defined order parameters, we collapse the average statistics onto an n x m grid to assess if increasing the number of experimental mutants to fit to and/or the number of near-optimal parameter sets improves prediction accuracy.

2.3 Results

2.3.1 Best-fit parameters and mutant stability

Best-fits for each of the C_p curves are provided in Figure 2.3, which highlights the quality of the model fits. Note that a simple solvent exposure model is actually sufficient to describe ΔC_p between low and high temperature [55]; however, such an approach cannot generate a peak within C_p, indicating that equilibrium fluctuations are not properly modeled. Within the mDCM, since we are not yet explicitly modeling solvation terms, the C_p baseline is subtracted. We analytically describe the baseline using a tanh(T) function. The rise of the tanh(T) function leads to a frequent slight overestimation of the C_p at high temperature, but this is mostly insignificant as the C_p peak is the primary region of interest. To the best of our knowledge, the mDCM remains the only free energy

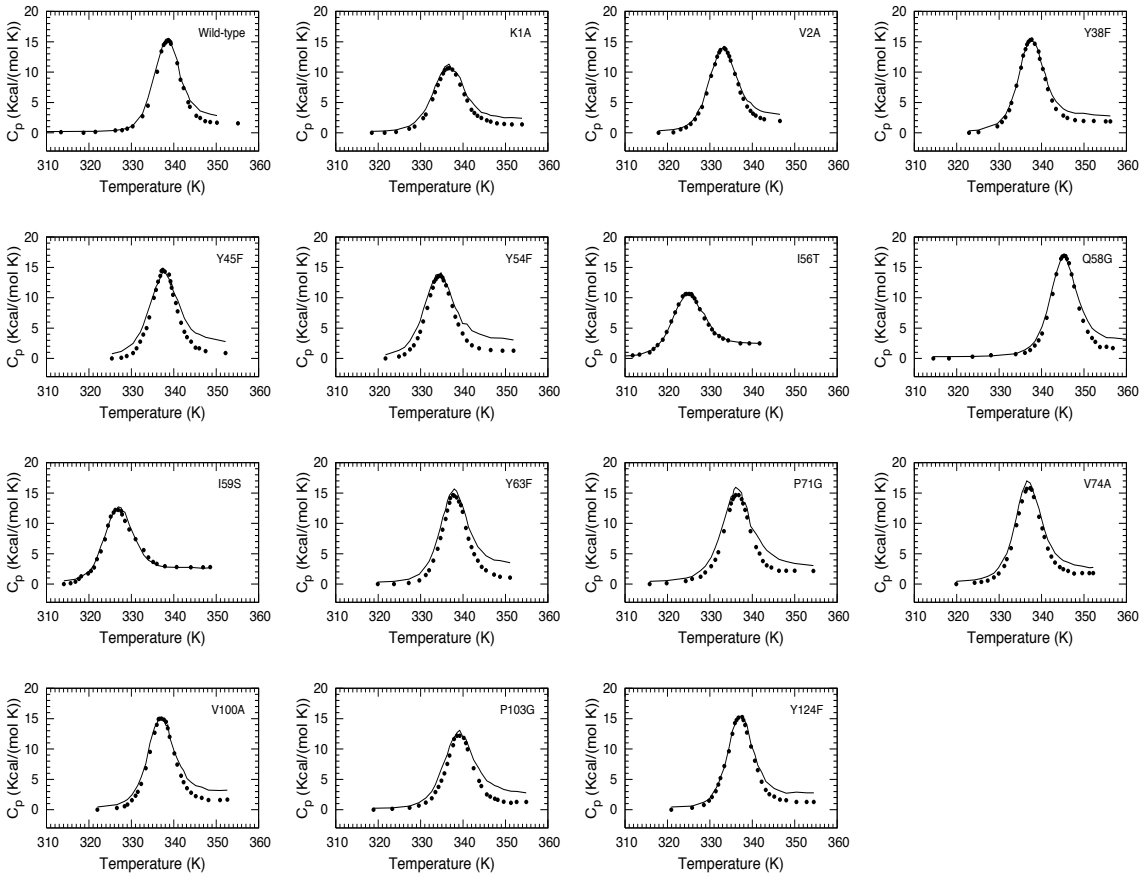

Figure 2.3: The $m = 1$ absolute best-fits to the experimental heat capacity data for the human C-type lysozyme and the 14 different point mutations considered here. Experimental data points are shown by dots, whereas the mDCM predicted curves are shown in solid line. To facilitate comparisons, the coordinate ranges in all 15 examples are equal.

decomposition scheme capable of quantitatively reproducing experimental C_p peaks. The associated best-fit parameter values are provided in Table 2.1. All parameter values are consistent with ranges established in our prior works studying globular proteins [43, 44, 56-58], and are physically meaningful. The conservation within u_{sol} is especially noteworthy, which is due to the enforced criterion that all experimental DSC solvent conditions be the same.

Not surprisingly, the vast majority of all mutations destabilize (relative to wild-type) lysozyme. In fact, only the Q58G mutation has an increased T_m, which is increased by an

astounding 6.6 K. As discussed in [52], the T_m of the P103G mutation is nearly identical to the wild-type. All of the other mutant T_m values range from 1 to ~14 K lower than that of the wild-type. While the wild-type structure has the lowest (most stabilizing) total H-bond enthalpy, there is actually only a weak correlation between total H-bond enthalpy and T_m (R = -0.57). In fact, the next most stabilizing total H-bond enthalpy is the P71G mutation, which has only the 11th largest T_m. Descriptions of the underlying H-bond networks are provided in Table 2.3. In all cases, the structures are of similar quality, as described by resolution and the observed R-value. Moreover, all of the mutant structures are from the same space group (P 2 2 2_1). Consequently, differences within the H-bond network can be reliably ascribed to conformational adjustments to relieve strain introduced by the mutation. As we have discussed previously [59], common descriptors (i.e., total H-bond enthalpy, average H-bond enthalpy, parameter values, etc.) are not good predictors of mDCM predictions across a set of closely related proteins.

Descriptors based on global topological properties of the protein fold contain much less

Table 2.3: Descriptions of the wild-type and lysozyme mutant structures. The Pearson correlation coefficient comparing the total H-bond enthalpy, number of H-bonds, and average H-bond enthalpy to the experimental T_m is, respectively, R = -0.50, -0.29 and -0.60.

Mutation	PDBID	Structure Resolution (Å)	Observed R-value	Total HB Enthalpy (kcal/mol)	Number of HBs	Average HB Enthalpy (kcal/mol)	T_m (K)
Wild-type	1LZ1	1.35	0.182	-613.9	240	-2.6	338.7
K1A	1C45	1.80	0.168	-567.8	245	-2.3	336.7
V2A	1OUG	1.80	0.173	-577.3	229	-2.5	333.4
Y38F	1WQO	1.80	0.170	-586.3	229	-2.6	337.7
Y45F	1WQP	1.80	0.174	-563.1	231	-2.4	337.4
Y54F	1WQQ	1.80	0.164	-565.9	229	-2.5	337.3
I56T	1OUA	1.80	0.148	-569.3	243	-2.3	325.1
Q58G	1B7R	1.80	0.160	-590.4	235	-2.5	345.3
I59S	2MEG	1.80	0.151	-554.3	239	-2.3	326.2
Y63F	1WQR	1.80	0.165	-589.1	239	-2.5	337.6
P71G	1LHI	1.80	0.156	-612.2	240	-2.6	336.1
V74A	1OUH	1.80	0.160	-586.1	235	-2.5	337.3
V100A	1OUB	1.80	0.160	-564.8	232	-2.4	337.1
P103G	1LHJ	1.80	0.152	-606.5	231	-2.6	338.6
Y124F	1WQM	1.80	0.164	-574.9	230	-2.5	337.6
	Average	1.77	0.163	-581.5	235.1	-2.5	336.1
	Variation	6.56%	5.740%	3.2%	2.3%	3.7%	1.5%

information than the mDCM, which is based on atomic level details affecting the network of distance constraints. Moreover, the way rigidity and flexibility propagate is non-trivial because network rigidity is a long-range mechanical interaction that results in complex emergent behavior that cannot be captured solely from local or global network characteristics.

2.3.2 Average prediction accuracy using a single parameter set

To assess how well the mDCM describes the experimental T_m values, we apply the best-fit parameters from one of the above fits serving as a transferable set of parameters, to all remaining lysozyme structures. We repeat this same process for all 15 permutations. This is the simplest scenario presented in this report, corresponding to $m = 1$ and $n = 1$. Across all 210 T_m predictions, Figure 2.4 highlights that more than 35% have errors less than 2%. The average error across all predictions is 4.3% (standard deviation = 3.6%). Clearly, the mDCM is doing a very good job at reproducing the experimental T_m values. Figure 2.5 plots the average T_m for each structure using the other 14 parameter sets. In all but three cases, the experimental T_m is within the error range defined by ± one standard deviation.

Interestingly, the two starkest exceptions (wild-type and P71G) correspond to the two structures with the most stabilizing total H-bond enthalpy, suggesting that the mDCM free energy calculation might be slightly over-dependent upon very low total H-bond energies. In practice, the primary goal of a computational model of protein stability is to assess relative stability of a mutant to the wild-type. To that end, Figure 2.6(A) provides a scatter plot of each predicted ΔT_m value ($T_{m,mut} - T_{m,wt}$) versus the experimental ΔT_m. The Pearson correlation coefficient is $R = 0.64$, which is among the

best values ever reported for biophysical models focusing solely on conformationally derived properties. Because the mDCM over predicts the wild-type T_m so drastically, all predicted ΔT_m values are negative. However, Figure 2.6(B) plots the average ΔT_m' values (defined as: $T_{m,mut}^{pred} - T_{ref}$, here $T_{ref} = T_{m,wt}^{exp}$) versus the experimental ΔT_m values, which

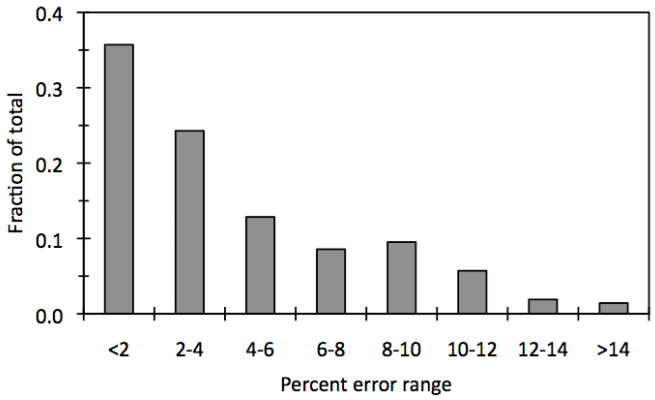

Figure 2.4: Histogram plotting the accuracy of the mDCM T_m predictions when only a single parameter set is used.

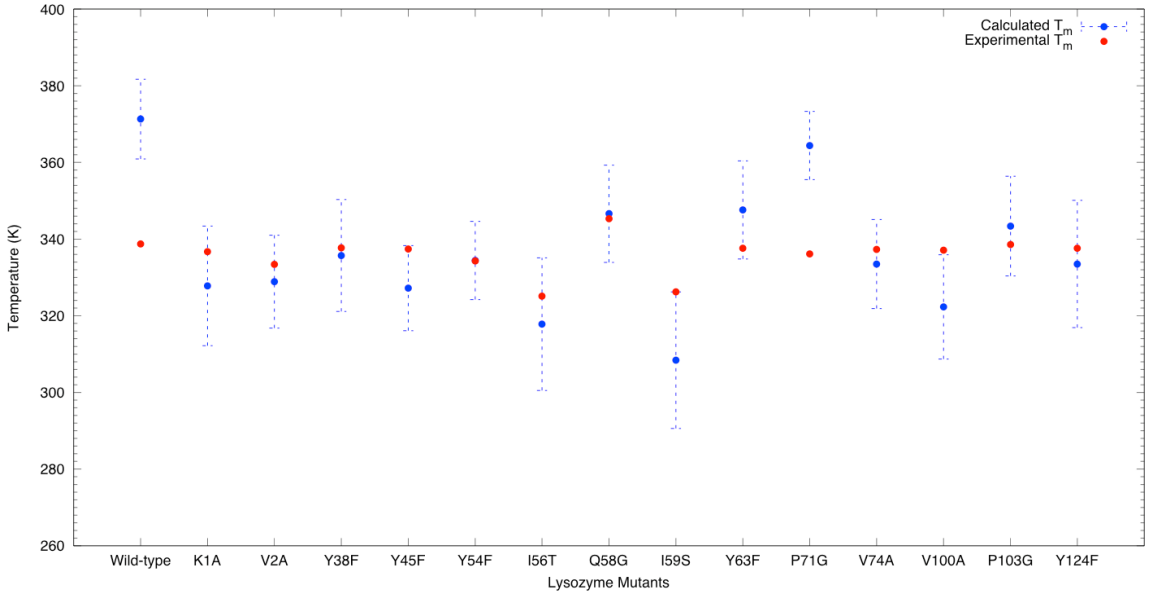

Figure 2.5: The average T_m value for each structure using each of the other 14 parameter sets is plotted (error bars equal ± one standard deviation). In all but three cases, the experimental T_m falls within the range defined by the error bars.

demonstrates that once an appropriate reference point has been established, the mDCM does a very good job of predicting stabilizing mutations to be stabilizing (quadrant 1) and destabilizing mutations to be destabilizing (quadrant 3). Only three predictions are located in an incorrect quadrant. Note that this arbitrariness in defining a reference point is generally unnecessary to resurrect a satisfactory quadrant clustering using any of the other structures as a reference point. The sole other exception is P71G, whose T_m is also over predicted by the mDCM.

2.3.3 Can accuracy be improved by additional parameterization?

Naively, it is expected that training on additional parameter sets from $n > 1$ lysozymes should improve average prediction accuracy. Similarly, it is expected that increasing parameter diversity by using m near optimal fits (up to some point before fit quality degrades) should also improve prediction accuracy. However, this is not the case here. Figure 2.7 plots multiple cross-sections from the n x m landscape. Specifically, the

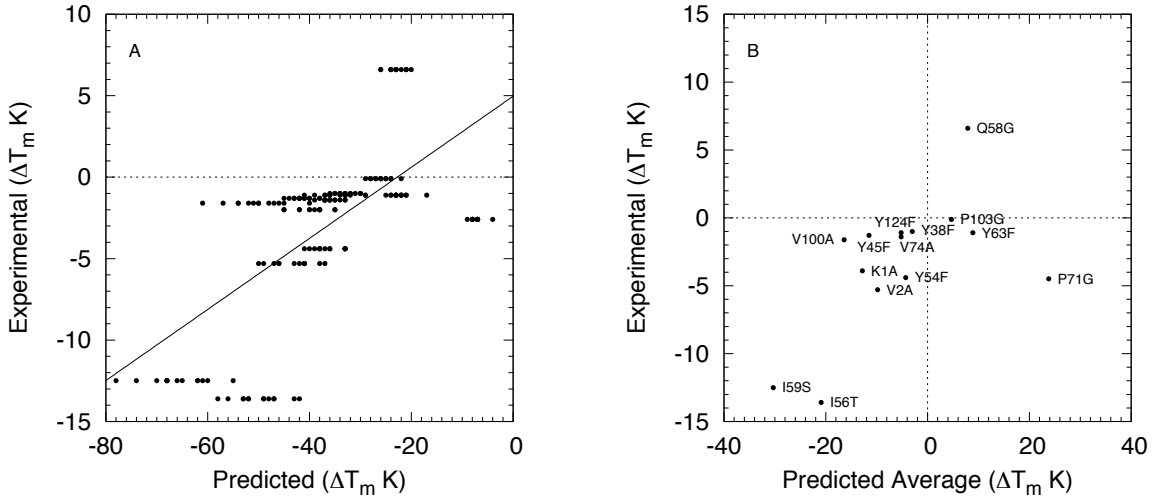

Figure 2.6: (A) The DT_m values ($T_{m,mut} - T_{m,wt}$) for each of the 14 x 13 = 182 cases is plotted against the experimental equivalent. The Pearson correlation coefficient is R = 0.64. (B) Average DT_m' values ($T_{m,mut(pred)} - T_{m,wt(exp)}$) versus the experimental DT_m values. The Pearson correlation coefficient is R = 0.60.

entire series of n is plotted for five different values of m over four different lysozyme examples.

In each case, the y-axis plots the $\langle \Delta T_m \rangle = \langle T_m^{pred} - T_m^{exp} \rangle$. Error bars indicate ± one standard deviation. To make sure that our sampling procedure is not statistically biased, we have reported average values over ten different simulations of 100 samples each. The average behavior is shown to be largely consistent across all ten simulations, and in each case the average ΔT_m values are well within the error bars of the other nine simulations.

Unexpectedly, no accuracy trends with increasing n or m are observed, meaning increased parameter diversity does not improve average prediction accuracy. Rather, for a given target, any particular parameter set gives a similar accuracy to any other set, indicating that the quality of the prediction is almost entirely dependent on the target structure itself. For example, Y45F is among the best-predicted structures, resulting in $\langle \Delta T_m^{\mathrm{Y45F}} \rangle \approx 0 \pm 10\mathrm{K}$. Conversely, as discussed above, the P71G mutation is particularly problematic, resulting in $\langle \Delta T_m^{\mathrm{P71G}} \rangle \approx 28 \pm 10\mathrm{K}$. Note that while a difference of 28K might seem large at first glance, it is in fact only an 8.3% percent error. The mutations V2A and V100A are shown as intermediate examples.

2.4 Discussion

The initial objective of this work was to improve the predictive value of the mDCM in a scenario that employs contextual learning. The idea was that a set of best-fit parameters for the mDCM, based on an experimentally determined structure and heat capacity measurement, could be used to predict the relative stability of protein mutants. As more experiments are performed, additional best-fit parameterizations could be determined based on the new systems, thereby boosting statistics, and, as such, better

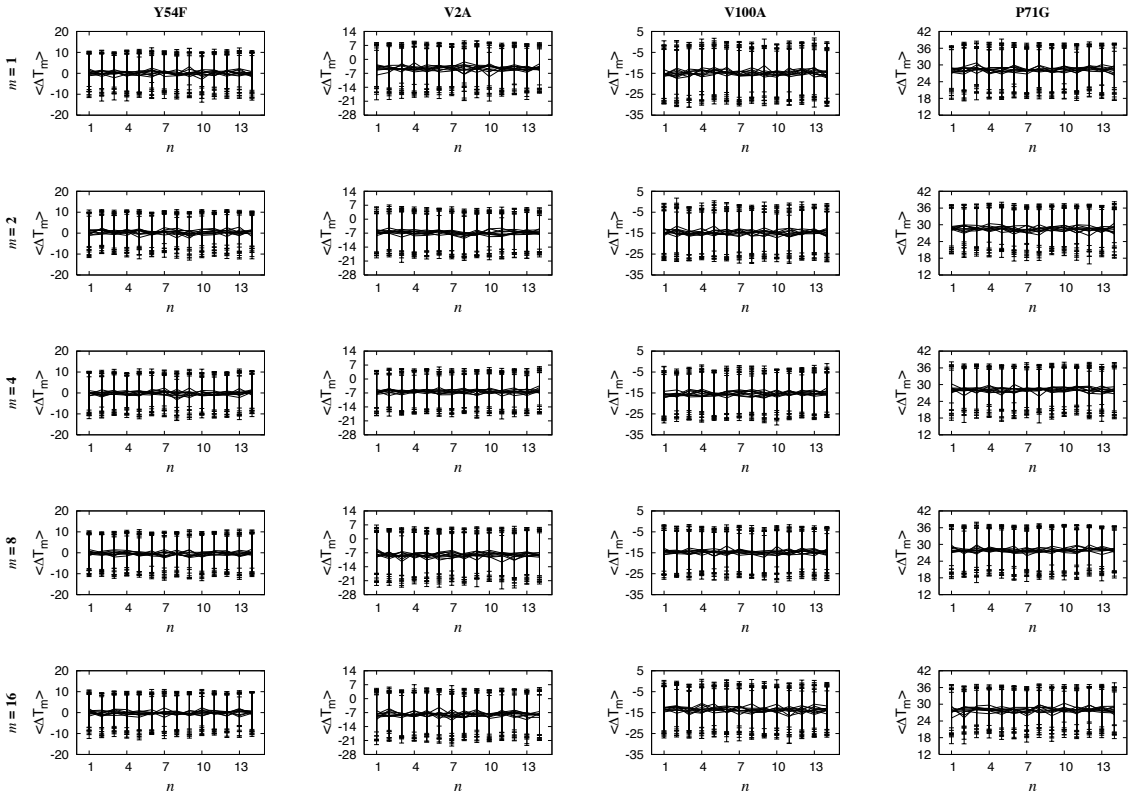

Figure 2.7: Cross-sections of the n x m landscape for four different lysozyme mutant examples (columns). In each case, average prediction accuracy is reported over all values of n for a given value of m. Five different values of m (rows) are shown. In all cases, our results surprisingly demonstrate that increasing the amount of parameter diversity does not improve the average difference between the experimental and predicted T_m values. The results from ten different simulations are shown superimposed on each other to show that our results are robust. Across the ten simulations, the average behavior is largely conserved, and in each case the average DT_m values are within the error bars of the other nine.

accuracy would occur as more experimental data is obtained. While it was initially surprising that increasing the amount of parameter diversity does not improve prediction accuracy in a statistically significant way, this result can be viewed in two ways. First, there may exist additional features of a protein that we can incorporate to filter out better parameter sets for a given structure. Second, these results reveal the saturation of accuracy inherent within the mDCM.

With the former view, perhaps the original objective of a context learning approach

can be recovered by using a more sophisticated statistical analysis. In this work, the considered model parameters provide a range of predicted T_m values, but all parameters sets are treated with equal weighting. Meaning, the collective statistics from the 'good' and 'bad' sets for a given structure cancel out in the average statistics, leading to ± standard deviation ~ 20K. If we could, somehow, only apply the parameter sets best suited to a particular structure, then the average prediction accuracy will naturally improve as n increases. To that end, it would be necessary to develop a classifier that identifies a good parameter set for a given structure to allow for a knowledge-based weighted average. For example, along these lines we considered estimating a target value for the u_{sol} parameter describing the average enthalpy for H-bonding to solvent as a function of global intramolecular properties of H-bonds. However, virtually no correlation between descriptors of global network properties and model parameters (in addition to model predictions as mentioned above) are found. Therefore, boosting the predictive accuracy using a classification scheme is likely to produce only marginal gains on prediction accuracy of a model that is intrinsically oversimplified. The expected marginal gain leads us to view these results in terms of a physical interpretation.

Based on the simplicity of the mDCM, where the H-bond network is featured so prominently in its free energy functional [43, 44], it should be surprising that the mDCM predictions yield such a high degree of accuracy and transferability of parameters. However, the mDCM has consistently proven to be a very robust and reliable model [43, 44, 56-58], presumably because of the encoded information that lies within the H-bond network. Of course, the importance of H-bonds has been part of a long-standing paradigm in protein biophysics [60]. This work benchmarks the best accuracy level that

can be hoped for using the mDCM since we are working with essentially an ideal system for its application. Yet, we are not suggesting the mDCM is the end of the story. On the contrary, the minimal DCM does not explicitly model other essential mechanisms such as hydrophobic and long-range electrostatic interactions. As demonstrated by Guerois *et al.* [61], inclusion of essential mechanisms will improve model accuracy. Similarly, we expect a considerable gain in accuracy will come forth when we employ a more complete free energy decomposition scheme.

2.5 Conclusion

As a part of this dissertation's "depth" analysis, we establish that the mDCM is a viable approach to predict the relative stability of protein mutants. Even using a single parameter set from some previously fit example, the average error of the method when applied to an unknown example is very good (average percent error = 4.3%), and it does a reasonably good job of reproducing experimental trends (R = 0.64), which is definitely good enough to be of practical value to experimentalists when making decisions about which mutations to invest time and funds for characterization. The results also point to the intrinsic limits of such a simplified model, and points to the need to develop a more complete free energy decomposition scheme.

CHAPTER 3: CHANGES IN LYSOZYME FLEXIBILITY UPON MUTATION ARE FREQUENT, LARGE AND LONG-RANGED

3.1 Introduction

Protein dynamics are intimately related to functional mechanisms [62], and changes therein can lead to observable phenotypes and disease [63]. These changes can be subtle. For example, a change in the amplitude of dynamical signatures upon ligation can lead to observable allosteric differences, even in the absence of global conformational changes [64]. While comparative assessment of structure and function is a long-standing paradigm within proteins (e.g. [65-67]), comparisons of dynamics across orthologous proteins are rare because experiments are labor intensive and costly. In spite of these difficulties, the importance of such comparisons has resulted in a small number of experimental assessments [68, 69].

Computational methods are promising alternatives to characterize and compare protein dynamics across many proteins [70-75]. In addition to being much less costly than experimental interrogations, computational methods are generally able to characterize protein backbone and sidechain dynamics in more detail than experimental means (depending upon the level of coarse-graining). Nevertheless, the computational expense associated with traditional simulations methods continues to make comprehensive analyses impractical [76]. To circumvent the cost of simulation, we have employed Distance Constraint Model (DCM) [77, 78] that provides quantified stability/flexibility relationships (QSFR) [79, 80], which is a high dimensional

description of protein thermodynamics, dynamics and their interrelationships. Specifically, we have employed a minimal DCM (mDCM) that considers hydrogen bonds (H-bonds) and native torsion forces as fluctuating interactions (*details provided in introduction*).

As a second part of the "depth" analysis and using the same human c-type lysozyme as our model system, we now try to establish how much a single-site mutation affects protein flexibility. As discussed earlier, these 14 different point mutants have been characterized under a narrow window of experimental conditions [81]. Surprisingly, we find that changes in flexibility upon mutation are very common. In fact, the number of positions with significant changes in flexibility characteristics is similar to the number of positions without change. Additionally, these changes can occur over relatively long distances, meaning they are frequently allosteric in nature. Changes that lead to increased backbone flexibility are slightly more common than changes that lead to increased rigidity. This asymmetry primarily occurs because many mutations lead to increased flexibility within lysozyme's β-subdomain. This result is noteworthy because several investigations have concluded that amyloid forming mutations lead to local unfolding in this region [82-86], which is the site of amyloid nucleation.

3.2 Methods

3.2.1 Dataset and structure preparation

In this work, we analyze X-ray crystal structures of 7 wild-type human c-type lysozymes and 14 spatially and chemically distinct point mutants. Each structure has been solved to high resolution (average = 1.8 Å), and all R-values are less than or equal to 0.19. PDBID's and all relevant structural information are provided in Table 3.1.

There are ~15 wild-type human lysozyme structures within the PDB. However, the series of cryogenic structures by Joti et al. [87] have extremely atypical properties, so we do not consider them here. In addition, the 1REZ [88] structure with a bound carbohydrate ligand also resulted in flexibility properties that were completely distinct from the remaining wild-type structures (and mutants for that matter). As such, it was also excluded, leaving the 7 considered structures.

There are many more lysozyme point mutant structures present in the PDB than the 14 considered here; however, this dataset has been carefully selected so that the C_p characterizations have been done under nearly identical experimental conditions [89-95].

Table 3.1: Structural and thermodynamic characterization of the dataset. Note that all structures come from the same P $2_1\,2_1\,2_1$ space group. In the fifth column, the a-carbon RMSD of each structure is compared to the 2NWD wild-type structure after minimization, which is the structure closest to the centroid of the wild-type set. Maximum C_p value in units of kcal/(mol·K). In all cases, δ_{nat} is equal to 1.24.

Protein	PDBID	Resol. (Å)	R-value	RMSD (Å)	T_m (K)	Max C_p	Total # of HB	u_{sol}	v_{nat}
WT	1JWR	1.4	0.18	0.7	339	15.6	244	-2.13	-0.31
WT	1LZ1	1.4	0.18	0.6	339	17.5	240	-1.85	-0.14
WT	1LZR	1.5	0.14	0.5	339	15.5	250	-1.86	-0.21
WT	1LZS	1.6	0.17	0.7	339	16.3	244	-2.35	-0.37
WT	1REX	1.5	0.19	0.8	339	15.5	234	-2.00	-0.24
WT	1REY	1.7	0.17	0.8	339	15.1	229	-1.89	-0.12
WT	2NWD	1.0	0.13	--	339	15.5	238	-1.78	-0.19
	Average	1.4	0.17	0.68	339.0	15.9	239.8	-1.98	-0.23
	Variation	15.4%	13.4%	17.1%	0.0%	5.1%	2.9%	10.1%	39.7%
K1A	1C45	1.8	0.17	0.9	337	13.1	245	-1.66	-0.18
V2A	1OUG	1.8	0.17	0.8	333	16.8	229	-1.78	-0.26
Y38F	1WQO	1.8	0.17	0.8	338	18.8	229	-1.72	-0.20
Y45F	1WQP	1.8	0.17	0.8	337	18.5	231	-1.79	-0.28
Y54F	1WQQ	1.8	0.16	0.8	337	17.3	229	-1.86	-0.29
I56T	1OUA	1.8	0.15	0.8	325	14.8	243	-1.84	-0.28
Q58G	1B7R	1.8	0.16	0.7	345	19.0	235	-1.90	-0.30
I59S	2MEG	1.8	0.15	0.8	326	14.4	239	-1.96	-0.40
Y63F	1WQR	1.8	0.17	0.7	338	18.5	239	-1.86	-0.24
P71G	1LHI	1.8	0.16	0.8	336	20.3	240	-2.10	-0.33
V74A	1OUH	1.8	0.16	1.0	337	18.8	235	-1.76	-0.23
V100A	1OUB	1.8	0.16	0.7	337	18.2	232	-1.91	-0.36
P103G	1LHJ	1.8	0.15	0.8	339	18.2	231	-1.73	-0.18
Y124F	1WQM	1.8	0.16	0.8	338	19.0	230	-1.92	-0.32
	Average	1.8	0.16	0.80	335.9	17.6	234.8	-1.84	-0.28
	Variation	0.0%	4.8%	9.8%	1.5%	11.8%	2.4%	6.2%	23.9%

As reported in stability prediction analysis in chapter 2, they have all been experimentally characterized using differential scanning calorimetry (DSC) under similar buffer conditions (pH = 2.7 to 2.8) and salt concentration (0.05 M). If this were not the case, model parameters would also reflect differences within the solvent conditions, thus obfuscating our direct comparisons. Moreover, full C_p curves must also be available in the literature for us to fit to. Finally, the C_p curves were generated by the same research group, which is important because DSC is a finicky technique that has systematic errors depending on differences in protocol and instrument. At the time of publishing this work, the 14 mutants studied here are the only ones that satisfy all of these criteria.

In all cases, hydrogen atoms are added using H++ server to ensure proper ionization [96] at the pH of the DSC experiments. The electrostatic parameters used are 0.05 M salinity and external/internal dielectrics of 80 and 6, respectively. Subsequently, the all-atom structures are minimized using the Molecular Operating Environment software using the Amber force field [97], which are then input into the mDCM.

3.2.2 Model parameterization

The mDCM is parameterized by finding values of $\{u_{sol}, v_{nat}, \delta_{nat}\}$ that best reproduce the experimental C_p data using the same simulated annealing protocol previously employed [81]. Across the dataset, the resultant best-fit parameters are very similar. Nevertheless, we checked how the observed sensitivity is dependent on model parameterization. That is, a change in model parameters might change the nature of the FI and CC results, and potentially change the conclusions. To explore this concern, we first applied individual 3-parmater fits, and then fit the C_p data using 2-free parameters per mutant while keeping the entropic parameter δ_{nat} fixed across the dataset (Table 3.1).

Note that we used a similar strategy in prior works since the value of δ_{nat} is related to protein fold [77, 78, 80, 98]. Encouragingly, the C_p curves are again accurately reproduced, and the FI and CC values are both quantitatively consistent with the 3-parameter model. Furthermore, quantitatively similar FI and CC results are also obtained using a constant $\{u_{sol}, v_{nat}, \delta_{nat}\}$ parameter set taken as the average over the 3-parameter best-fits (results not shown). For simplicity, the data presented throughout the report is solely based on the 2-parameter model, keeping in mind that the similar quantitative results arise from the other two-parameter sets.

The parameter differences observed in Table 3.1 phenomenologically reflect physical differences between the mutants that are not explicitly considered by the model. For example, as we have demonstrated previously [80], parameter variation is expected to account for differences in hydrophobic interactions. The extent of parameter variation observed here is relatively small, generally within the variation expected for multiple equally good fits. Moreover, while thermodynamic quantities (i.e., T_m) are somewhat sensitive to parameterization and input structure resolution, we have consistently demonstrated that mechanical FI and CC quantities are quite robust to parameter differences [80, 81, 98]. As such, the parameter differences have negligible affect on the presented results.

3.2.3 Flexibility index and cooperativity correlation

The flexibility index (FI) and cooperativity correlation (CC) are ensemble-averaged quantities over the native basin in the free energy landscape at the melting temperature. For a given macrostate, a sample constraint network is constructed using the probabilities for individual constraints to be present as described previously [17]. When no native

torsions are present and no H-bonds are present, all the rotatable-bonds in the network are labeled from 1 to N. As constraints are added to the network, some of these bonds will become part of rigid regions. Then, for a given constraint network, a rigidity analysis is performed, and each *a priori* rotatable bond is identified as being: (*i.*) flexible because it is part of an under-constrained region, (*ii.*) locked because it is part of an isostatically rigid region, or (*iii.*) locked because it is within an over-constrained region. These three types of regions define clusters within the protein. No other possibility can occur [99], and all rotatable bonds are assigned to 1, and only 1, cluster. If the cluster is over-constrained, this means there are more constraints in the region than is necessary to make it rigid. If the cluster is isostatic, then the region is rigid, but there are just enough constraints to make it rigid. If there are not enough constraints within a certain region, it will be flexible.

Each bond is assigned a flexibility index, f_i, that is defined based on a single constraint network as follows. If the bond in question is part of an isostatically rigid region, $f_i = 0$. If the bond in question is part of a flexible region, the number of rotatable bonds within that flexible region is counted, and is denoted as H. The number of independent disordered torsions within that same flexible region is counted, and is denoted as A. To represent the density of independent DOF within the flexible region, the value $f_i = A/H$ is assigned to all bonds within this cluster. Finally, if the bond in question is found to be in an over-constrained region, the total number of *a priori* rotatable bonds are counted, and denoted as D. Furthermore, the total number of redundant constraints within that region are counted, and is denoted as B. The value $f_i = B/D$ represents the density of redundant constraints within this over-constrained region, and it is assigned to

all the bonds within this cluster. Once this counting is complete for every cluster, every *a priori* rotatable bond in the protein will have a flexibility index assigned to it. To distinguish between densities of DOF versus redundant constraints, the f_i values corresponding to flexible regions are positive, whereas the above f_i values in over-constrained regions are multiplied by -1. We focus our analysis herein on just the backbone *a priori* rotatable bonds that comprise the f and y angles of all residues (except proline, for which there is just a ψ angle).

In the final stages of the process, we typically average over 1000 or more realizations to obtain averaged mechanical properties for a given macrostate, (j,k). Then, for the i-th *a priori* rotatable bond, we have $FI(i \mid j,k) = \overline{f}_i(j,k)$, where the bar is used to indicate an arithmetic mean over all samples randomly generated by Monte Carlo sampling subjected to the given macrostate (j,k). The reported FI for the i-th *a priori* rotatable bond is given as: $\langle FI(i) \rangle = \sum_{j,k} \overline{f}_i(j,k)\, p(j,k)$.

We employ a similar procedure to calculate the average value of CC. The main difference is that CC represents a pair correlation so the end result is a symmetric square matrix rather than a one-dimensional array. The variable $c_{m,n}$ is equal to f_m if the m-th and n-th *a priori* rotatable bonds are simultaneously found to be in the same flexible, isostaticaly rigid or over-constrained region. This is because the same value is assigned to all *a priori* rotatable bonds within a given cluster type. The correlation becomes apparent whenever two distinct types of clusters are identified. For example, if the m-th and n-th rotatable bonds are both found to be in rigid clusters, but these clusters are distinct, then $c_{m,n}$ is equal to 0. In general, $c_{m,n} = 0$ if the m-th and n-th *a priori* rotatable bonds belong to distinct clusters (whether of the same type or not). Thus, it should be noticed that no

distinction is made between two *a priori* rotatable bonds being simultaneously found in the same isostatic rigid cluster versus in two different rigid clusters. It turns out that the relative frequency of two bonds being in an isostatic rigid region is very low. The distinction for why $c_{m,n} = 0$ was initially a concern, and different measures have been considered. However, it was found that the reported average CC plots provide ample information regarding how flexibility and rigidity propagate through a protein [77-80, 98, 100, 101]. We prefer to use the CC plot based on the density information as described here because it directly connects to the FI. In the next stages of the calculations, $CC(m,n \mid j,k) = \bar{c}_{m,n}(j,k)$ is the conditional average for a given macrostate, and the reported CC is given as $\langle CC(m,n) \rangle = \sum_{j,k} \bar{c}_{m,n}(j,k)\, p(j,k)$. Using this procedure, CC plots identify all pairwise residue-to-residue couplings across the structure (Figure 3.1c). Consequently, correlated motions associated with a high density of DOF show up in red, while a high density of redundant constraints show up in blue. Regions that are marginally mechanically stable or simply uncoupled show up as white.

3.2.4 Accessing changes in flexibility

Perhaps the most critical aspect of the presented work is determination of what constitutes a change in flexibility and what does not. That is, what degree of precision is present with the mDCM flexibility measures? This point is particularly important in this work because, using normal structure comparison metrics, the mutant dataset considered here is very similar to the wild-type structure. To address this point, we establish a baseline of ambient flexibility changes across a set of 7 wild-type structures [88, 102-105], such that differences within the background profile arise from subtle differences in the wild-type X-ray structures (Table 3.1). The baseline flexibility profile for each

residue position for each residue FI value or pixel for CC is calculated as the average value over the set ±1 σ, where the standard deviation, σ, is respectively calculated over each data set at the corresponding residue or pixel. Then, any mutant flexibility metric within one standard deviation is considered "no change." A value falling in the range between one and two standard deviations away from the mean defines "moderate" changes, whereas "large" changes are defined as greater than 2 standard deviations from the mean. As discussed above, Figure 3.1a plots FI versus residue number for the wild-type baseline profile. The difference data presented in Figures 3.3 and 3.4 has been discretized into bins based on the above σ ranges. However, difference data in Figures 3.2 and 3.5 retain quantitative relative differences by setting the response in the change of flexibility to zero when it is within the noise level, and only allowing the signals to show up. In ΔFI_n and ΔCC_n the data is normalized in the following way:

$$x_n = \begin{cases} \min\left(\dfrac{x-\sigma}{2\sigma}, 1\right) & \text{if } x > \sigma \\ 0 & \text{if } |x| \le \sigma \\ \max\left(\dfrac{x+\sigma}{2\sigma}, -1\right) & \text{if } x < -\sigma \end{cases}$$

The outcome of the above equation is that all values within the background profile are colored white, whereas continuous color schemes are used for the moderate change bins. The min() and max() functions are employed to threshold the coloring such that all "large" changes are colored the same maximum shade of red or blue. Further, because the values are normalized by context dependent standard deviations they in essence provide a degree of statistical significance for the observed change. That is, a change could be quantitatively large in raw values, but appear weak if the background variability was large. On the other hand, for extremely small standard deviations, the change will appear

disproportionally large. However, this concern is largely unfounded as the per-pixel standard deviations in both ΔFI_n and ΔCC_n are relatively uniform (data not shown). In fact, plotting the raw differences actually makes changes appear roughly twice as frequent as we observe with the normalized scores, which would only strengthen the main conclusions of this chapter. In other words, the normalized plots filter out response main conclusions of this chapter. In other words, the normalized plots filter out response that does not have a signal large enough to distinguish against the background noise.

3.3 Results

3.3.1 Intrinsic flexibility of wild-type lysozyme

Previously, we used the mDCM to predict mutant melting temperatures with an average error of 4.3% [81]. Going further, the primary goal of this investigation is to critically evaluate the consequences of single point mutations on lysozyme flexibility. However, before doing so, we must first quantify wild-type lysozyme's intrinsic flexibility characteristics to be used as our reference point.

We define an average flexibility profile using a set of 7 different human wild-type lysozyme structures. Therein, differences in flexibility solely arise from differences in the X-ray crystal structures. Moreover, the variability across the dataset establishes a baseline precision for the calculated properties. Values within ±1 standard deviation (±1 σ) from the mean of the wild-type set are taken to be within background noise, and are thus deemed equivalent. Figure 3.1a plots the flexibility index (FI), which is an mDCM output that characterizes local flexibility. Positive values quantify flexible regions, whereas negative values quantify rigidity. Additionally, the variability within FI across the 7 wild-type structures is also shown. Figure 3.1b maps the average flexibility profile to structure

(blue = rigid, whereas red = flexible). In general, helices are mostly rigid, whereas spanning loop regions are mostly flexible. The β-subdomain is marginally rigid, with some interspersed flexibility. The β-subdomain is attached to the core via a known hinge region that is identified by the mDCM [78]. The flexible hinge region and lysozyme's

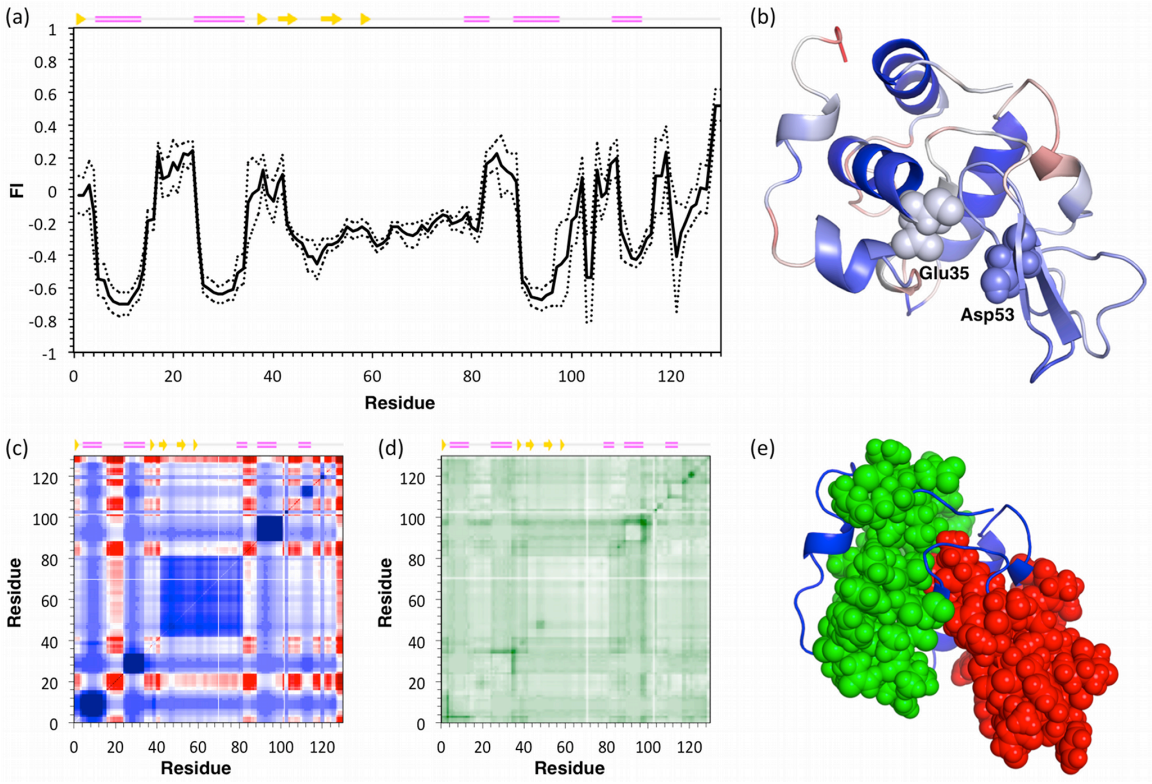

Figure 3.1: Intrinsic flexibility characteristics for lysozyme are shown. (a) The average flexibility index (FI) across a set of seven wild-type lysozyme structures is plotted versus residue number (solid line). The dashed lines indicate ±1 σ, which defines the noise range within the quantity. (b) Lysozyme is color-coded according to average FI values in panel (a), where red regions indicate flexibility (FI > 0) and blue indicates rigidity (FI < 0). (c) The cooperativity correlation profile of 2NWD identifies all pairwise mechanical couplings. Red indicates residue pairs within the same correlated motion, whereas blue indicates residues within the same rigid cluster. White indicates no mechanical coupling. Panel (d) shows the relative per pixel standard deviation across the wild-type set where darker color represents a greater value. There are two large rigid clusters identified in panel (c), which are highlighted in panel (e). The first (green) is defined by helices α1, α2, α4 and α5, whereas the second (red) corresponds to the β-subdomain. The active site is located at the cluster interface, and the hinge motion indicated in panel (b) allows the enzyme to close around its substrate.

two catalytic residues are also highlighted. Most of the other flexible regions correspond to loops connecting secondary structure elements.

A higher order description of protein dynamics is provided by CC, which characterizes correlated motions and co-rigidity or pairwise residue-to-residue mechanical couplings. Figure 3.1c plots the CC for the 2NWD structure, which is the closest to the geometric center of the wild-type set. Blue coloring identifies co-rigid residue pairs (meaning residue pairs with high probability of occurring within the same rigid cluster), whereas red coloring identifies flexibly correlated pairs (residue pairs within a correlated motion). Mechanically decoupled regions are colored white. The per-pixel variation across the wild-type set is plotted in Figure 3.1d. Within Figure 3.1c, two prominent rigid clusters can be identified. The first is composed of helices α1, α2, α4 and α5, whereas the second spans the β-subdomain region (Figure 3.1e). The active site and accompanying hinge motion corresponds to the cluster interface, which allows the enzyme to close around its carbohydrate substrate.

3.3.2 Changes in backbone flexibility upon mutation

The primary goal of this report is to investigate changes in lysozyme dynamics upon mutation. To that end, we analyze changes in FI and CC that occur upon mutation. The profiles defined above establish when a change in flexibility is significantly above background noise. Figure 3.2a plots the normalized change in FI (ΔFI_n) for each mutant where red indicates increased flexibility, and blue indicates increased rigidity. Some common responses are identified regardless of the details of the mutation. Interestingly, flexibility increases frequently occur within the β-subdomain regardless of mutation position, while an increase in rigidity within the β-subdomain almost never occurs.

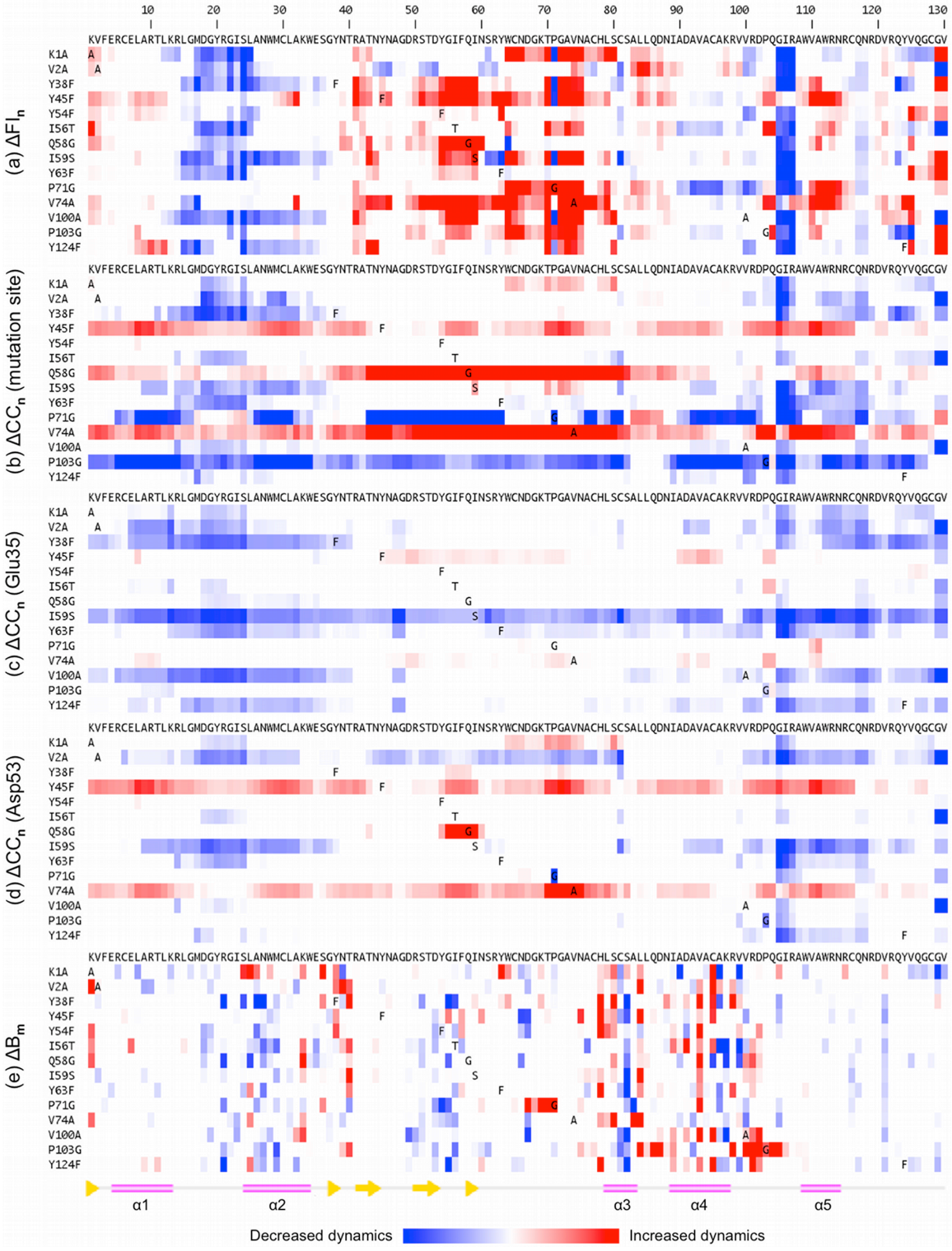

Figure 3.2: Comparison of backbone flexibility and changes across the dataset are shown. Panel (a) plots changes in ΔFI_n for each mutant relative to the wild-type structure. In the same manner, changes in cooperativity correlation (CC) with respect to the mutation site, Glu35 and Asp53 are respectively plotted in panels (b), (c) and (d). Panel (e) plots changes in the median normalized B-factors across the dataset.

Changes in the α-subdomain are slightly less frequent with the most common responses having increased rigidity within the α1/α2 loop and a 3-residue segment of the α4/α5 loop.

Despite the above trends, many site-specific differences are obvious. Binning the ΔFI_n values across a collapsed dataset of all 14 mutants underscores this point. Figure 3.3a indicates that the dynamics are appreciably changed in 48.0% of the residues upon mutation. Interestingly, the percentage of residues with increased flexibility (28.0%) is

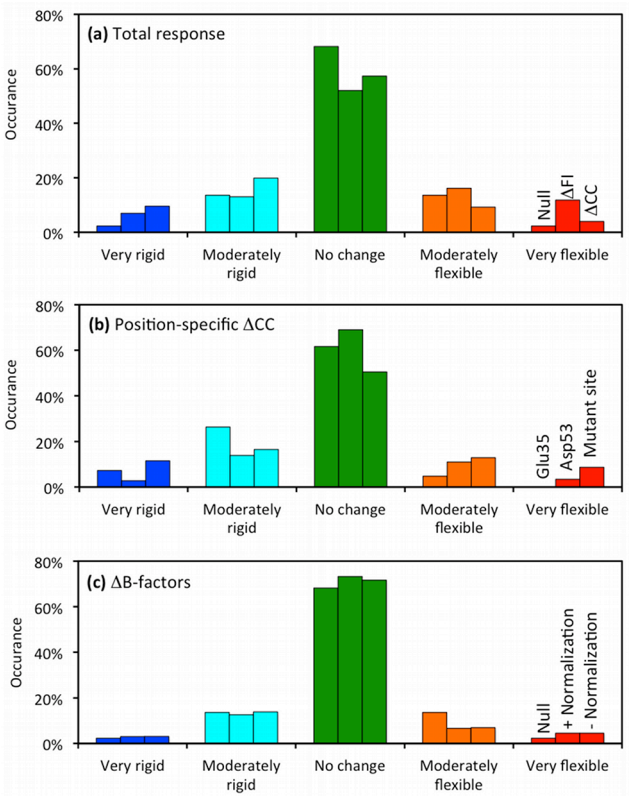

Figure 3.3: Flexibility response histograms are shown. Across a collapsed dataset constructed from all 14 mutant structures, each residue is binned based on changes to QSFR properties. The bins are color-coded by: green = no change, cyan and blue = moderate and large increases in rigidity, and orange and red = moderate and large increases in flexibility. In each panel, the bin order is conserved and indicated at the right. Panel (a) plots the null expectation histogram (highlighted with diagonal hashing) alongside the overall changes in flexibility index and cooperativity correlation. Panel (b) plots changes in cooperativity correlation with respect to specific residues: the mutation site, Glu35 and Asp53. Finally, panel (c) re-plots the null expectation alongside changes in B-factors (with and without median normalization).

slightly more than the percentage with increased rigidity (20.0%). This result makes intuitive sense because all but one of the mutants decreases structural stability. We segregate moderate flexibility changes from large changes using a cutoff of ±2 σ. Percentages of large increases in flexibility are slightly more than large increases in rigidity (11.8 vs. 7.0%). Based on the ±1 σ definition of the "no change" background profile, the null expectation is that 68.2% of the positions should have "no change." Further, moderate changes within 1 to 2 standard deviations, and large changes greater than 2 standard deviations, have null expectations of 13.6% and 2.3%, respectively. Figure 3.3a clearly indicates that we observe more changes in FI than this random expectation. Using the chi-square statistic, the differences within the observed and random expected histograms are strongly significant (Table 3.2). That is, changes in flexibility upon mutation are more common than the background variation across the set of wild-type structures.

Using the same coarse-grained color scheme as Figure 3.3a, the first column in Figure 3.4 color-codes the mutant lysozyme structures by ΔFI_n values. In each, the structures are shown in nearly identical orientations, and the mutated residue, Glu35 and Asp53 are rendered in spacefill view to orient the viewer. In addition to highlighting the frequency of changes in flexibility or rigidity upon mutation, this figure emphasizes that changes can be quite long-ranged. For example, the I59S mutation, which occurs within the β-subdomain portion of the active site cleft, affects the most distant portions of the structure. Even more pronounced is the P71G mutation. The mutation site is located on the outmost reach of the β-subdomain, yet it causes helix α4 at the hinge and the α4/α5 loop within the main core of the protein to significantly rigidify. Concurrently, the β-subdomain and

helix α5 become much more flexible.

3.3.3 Changes in cooperativity correlation upon mutation

Going further, Figure 3.5 shows the normalized changes in cooperativity correlation (ΔCC_n) upon mutation, which reveals a much more rich and interesting set of changes in flexibility. Again, we characterize the degree of change with respect to the mean wild-type CC values using the same standard deviation ranges as above. Across all mutants, an increased correlated flexibility is observed in 42.7% of the CC values. Interestingly, the bias towards increased correlated flexibility observed in ΔFI_n is not present. Rather, ΔCC_n results are skewed in the opposite direction (Figure 3.3a). Specifically, increased rigidity correlation is observed in 29.5% of the ΔCC_n values, whereas only 13.2% have increased flexibility correlation. This asymmetry stresses the physical distinction between the two metrics. While the ΔFI_n results describe changes in backbone flexibility within a localized region, ΔCC_n identifies changes in pairwise mechanical couplings that uncover cooperative effects. The results from our dataset indicate that most of the increases in

Table 3.2: Statistical significance of the observed histograms. Bin sizes within the expected histograms are defined from the variation across the set of wild-type structures: large changes > ±2 σ, moderate changes are ±1-2 σ, and no change is between ±1 σ, from which background bin probabilities are calculated. The chi-square statistic is use to compare the expected and observed histograms, and the reported p-values quantify the probability that the histograms are equivalent. In all cases, the histograms are determined to be statistically distinct from the null expectation.

Flexibility metric	**p-value**
ΔFI_n	9.1E-212
ΔCC_n (all positions)	0.00
ΔCC_n (Glu35 only)	1.6E-122
ΔCC_n (Asp53 only)	2.7E-4
ΔCC_n (mutation site)	1.8E-237
ΔB_m	2.6E-24
ΔB_r	2.4E-22

backbone flexibility are localized, frequently within the b-subdomain (with Y45F and V74A being the primary exceptions). Put otherwise, the local increases in flexibility identified by ΔFI_n are largely decoupled from other motions, which is why ΔCC_n does not show a large increase in correlated flexibility. Conversely, the increased rigidity correlation across the dataset indicates that most of the increases in backbone rigidity are frequently coupled to other rigid regions throughout structure.

As with ΔFI_n, the differences in ΔCC and the null expectation are strongly significant (Table 3.2). Another key deviation from the ΔFI_n results is the high variability across the set of mutants. For example, the Y54F mutant has little overall affect on the set of mechanical couplings within lysozyme. Conversely, the same mutation at position 45 leads to a large increase in flexibility correlation, whereas the Y→F mutations at positions 38, 63, and 124 slightly increase co-rigidity. A similar juxtaposition occurs within the V→A mutations. V74A drastically increases correlated flexibility; however, V2A has the opposite affect by drastically increasing correlated rigidity. While these cases represent nearly homogenous changes in CC, most of the remaining mutants have a mix of both increased correlated flexibility and correlated rigidity. Taken together, the large and diverse mutant-specific changes within the ΔCC_n results underscore the high sensitivity of the metric, which we have discussed previously [80, 100, 101].

It is technically difficult to exhaustively compare all changes because the two-dimensional nature of the data precludes linear descriptions along the lysozyme sequence. As such, we extract for further analysis strips of ΔCC_n values from the full plot for a single residue point of reference. Here, we examine ΔCC_n with respect to the mutation site and the two catalytic residues. These results are reported alongside the ΔFI values

just discussed in Figures 3.2-3.4, which underscores the richness within ΔCC_n. For example, changes in CC with respect to Glu35 are common. Moreover, they can be quite large and frequently propagate over long distances. The same is true for ΔCC_n with respect to the mutation site. On the other hand, changes with respect to Asp53 are somewhat suppressed, yet still statistically significant. These cases emphasize that the extent and location of changes within the mechanical couplings is dependent upon the reference point. Similar types of differences are observed when examining ΔCC_n from other points of reference.

3.3.4 Flexibility is distinct from mobility

Protein dynamics can be quantified in many ways. Therefore, it is important to distinguish flexibility from mobility. From rigidity theory, flexibility indicates that a network is deformable, but it need not be mobile. For example, a stationary pivot of a swinging pendulum is highly flexible, but not mobile. On the other hand, the end of the pendulum can simultaneously be rigid and highly mobile [106]. Because of this physical distinction, it is useful to benchmark how mobility changes upon mutation. To that end, we compare changes in α-carbon atomic displacement parameters (B-factors) of each mutant structure to the wild-type profile. However, before doing so, it should be stressed that caution must be employed when analyzing B-factors in terms of mobility because protein crystals are not homogeneous. That is, protein structure B-factors reflect both temporal (i.e., mobility) and spatial disorder across the crystal lattice. B-factors are quantitatively affected by occupancies. Occupancies less than one can be an indication of disorder, but lead to improved R-factors [107]. As such, even when multiple structures have the same space group, direct comparisons of B-factors reflect substantially more

than just differences in mobility. Thus, using B-factors to reflect mobility is only truly accurate when all other error sources have been removed. To help mitigate some of these caveats, we normalize B-factors using the median-based method of Smith et al. [108].

After normalization of the α-carbon B-factors within each structure, we calculate the wild-type background profile in the same way as above. Subsequently, normalized B-factors from each mutant structure are compared to the normalized wild-type profile using the same σ ranges as above in order to classify no change, compared to moderate and strong changes. Surprisingly, the histogram of median normalized B-factor changes (ΔB_m) (Figure 3.3c) is substantially different from the flexibility changes. Specifically, there are fewer changes in B-factors than one would expect based on the wild-type profile. This suppression of changes is statistically significant (Table 3.2). Moreover, there is no correlation between the ΔFI_n quantities and ΔB_m values (results not shown), underscoring the differences between flexibility and mobility. Despite the cautionary note above about B-factor comparisons, we also compare the raw B-factor changes (ΔB_r) to determine if normalization is biasing the results. Figure 3.3c also shows that there are no appreciable differences between the ΔB_m and ΔB_r histograms. For completeness, the ΔB_m values are reported alongside the ΔFI_n and ΔCC_n results in Figures 3.2-3.4. No correlation is found using raw data as well.

3.3.5 Structural considerations of flexibility changes

Table 3.3 counts the number of residue responses that occur for a given solvent accessibility and distance separation (mutation α-carbon to response α-carbon) range. The collapsed dataset of all residues is stratified by solvent accessibility for both the response (top) and mutation (bottom) sites. In each case, exposed, moderate, and buried

Table 3.3: Residue response statistics. Each cell counts the number of residue responses (ΔFI) that correspond to a given solvent accessibility range (or structural element) for a given distance to the mutation site. The collapsed dataset of all residues is stratified by response residue solvent accessibility in the top half of the table, whereas the collapsed dataset is stratified by mutation site solvent accessibility in the bottom half. The ratio value in the last column is the number of residues with altered flexibility divided by the number of residues with no change.

	Large rigidity increase	**Moderate rigidity increase**	**No change**	**Moderate flexibility increase**	**Large flexibility increase**	**Ratio**
Distance from response site = 0 to 8 Å						
Buried	3	7	26	7	17	1.31
Moderate	1	5	30	10	12	0.93
Exposed	1	3	21	8	12	1.14
Union	5	15	77	25	41	1.12
Distance from response site = 8 to 16 Å						
Buried	24	38	130	28	31	0.93
Moderate	9	31	130	35	27	0.79
Exposed	9	9	81	36	18	0.89
Union	42	78	341	99	76	0.87
Distance from response site ≥ 16 Å						
Buried	25	41	155	53	23	0.92
Moderate	23	49	174	49	29	0.86
Exposed	32	54	200	68	46	1.00
Union	80	144	529	170	98	0.93
Structural characterization of response site						
Helix	95	170	553	117	45	0.77
Strand	1	4	51	30	26	1.20
Coil	31	63	343	147	144	1.12
α-Subdomain	115	220	693	177	69	0.84
β-Subdomain	12	17	254	117	146	1.15
Mutant residue is buried						
0-8 Å	4	9	33	9	16	1.15
8-16 Å	32	52	145	41	25	1.03
≥ 16 Å	28	37	153	41	25	0.86
Union	64	98	331	91	66	0.96
Mutant residue moderately exposed						
0-8 Å	0	4	23	14	17	1.52
8-16 Å	5	18	108	39	40	0.94
≥ 16 Å	22	58	175	78	49	1.18
Union	27	80	306	131	106	1.12
Mutant residue is exposed						
0-8 Å	1	2	21	2	8	0.62
8-16 Å	5	8	88	19	11	0.49
≥ 16 Å	30	49	201	51	24	0.77
Union	36	59	310	72	43	0.68
Structural characterization of mutant residue						
Helix	25	40	133	33	29	0.96
Strand	28	45	180	84	53	1.17
Coil	74	152	634	177	133	0.85
α-Subdomain	65	120	408	108	79	0.91
β-Subdomain	62	117	539	186	136	0.93

respectively corresponds to the top, middle, and bottom thirds of all relative solvent accessibilities, which maintains similar observations in each stratum for the response and mutation sites. The ΔFI_n bins again correspond to those in Figure 3.3. Interestingly, in both cases solvent accessibility has little effect on the response rate. In all cases but one, the ratio of changes to no change is approximately one. That is, a change in flexibility is generally as frequent as no significant change. Note that we focus on the ratio of changes because this normalizes out the size discrepancies --- the strata corresponding to larger distances will naturally have bigger counts simply because there are fewer residues close to the mutation compared to farther away. The one noticeable exception to this general trend is when the mutant residue is solvent exposed, for which there is a significant decrease in flexibility changes. This relative lack of effectiveness in causing a change in flexibility makes intuitive sense because solvent exposed positions are naively expected to be more tolerant to mutation due to reduced steric constraints. Table 3.3 additionally provides statistics comparing structural features of the response and mutation sites. First, the dataset is stratified by secondary structure.

As discussed above, there is a slight reduction in the relative response rate for α-helical positions. Conversely, there is slight increase in the β-strand positions, which is strongly skewed towards increases in flexibility. Table 3.3 also provides statistics for the α- and β-subdomains, which parallels the secondary structure results. That is, the β-subdomain is highly susceptible to increased flexibility upon mutation. Conversely, a mix of changes in the α-subdomain commonly occurs, albeit at a rate slightly lower than no change. Interestingly, the ratios are more similar (~1) across secondary structure and subdomain boundaries when focusing on the mutation site, with coil residues being the

sole exception. Mutation of coil residues tends to have a decrease in the relative response rate, which simply reflects the same observation above for mutation of solvent exposed residues. The ratios for ΔCC_n are qualitatively similar, albeit slightly less across the entire dataset. The average ratio for ΔCC_n is ~ 0.7, meaning a lack of change in CC is more common than a change. Nevertheless, changes in CC that have been observed as general trends in prior work [80, 100, 101] are observed here as evident in most cases within Figure 3.5, where drastic changes usually appear within a small number of strips. However, there are certain cases (i.e., V2A, Y45F, and V74A) where virtually the whole CC plot is affected.

3.4 Discussion

3.4.1 Changes in flexibility upon mutation are common and large

Previous works have investigated how familial divergence affects protein dynamics and, as a consequence, allostery. Using DCM, initial work along these lines compared a mesophilic and thermophilic RNase H pair [80], which reproduced experimental conclusions regarding the balance between molecular flexibility and thermodynamic stability [109-112]. Subsequently, comparisons were expanded to 4 bacterial periplasmic binding homologs [100] and 9 oxidized thioredoxin structures [101]. Taken together, these collective results suggest an intriguing mix of conservation and variability within stability and flexibility. Pairwise mechanical couplings that provide a higher order description of flexibility and rigidity are generally sensitive to small differences. The latter result highlight how small structural variations are amplified into global differences as mechanical couplings propagate through the network.

In other studies, DCM has been employed to link mechanical and thermodynamic

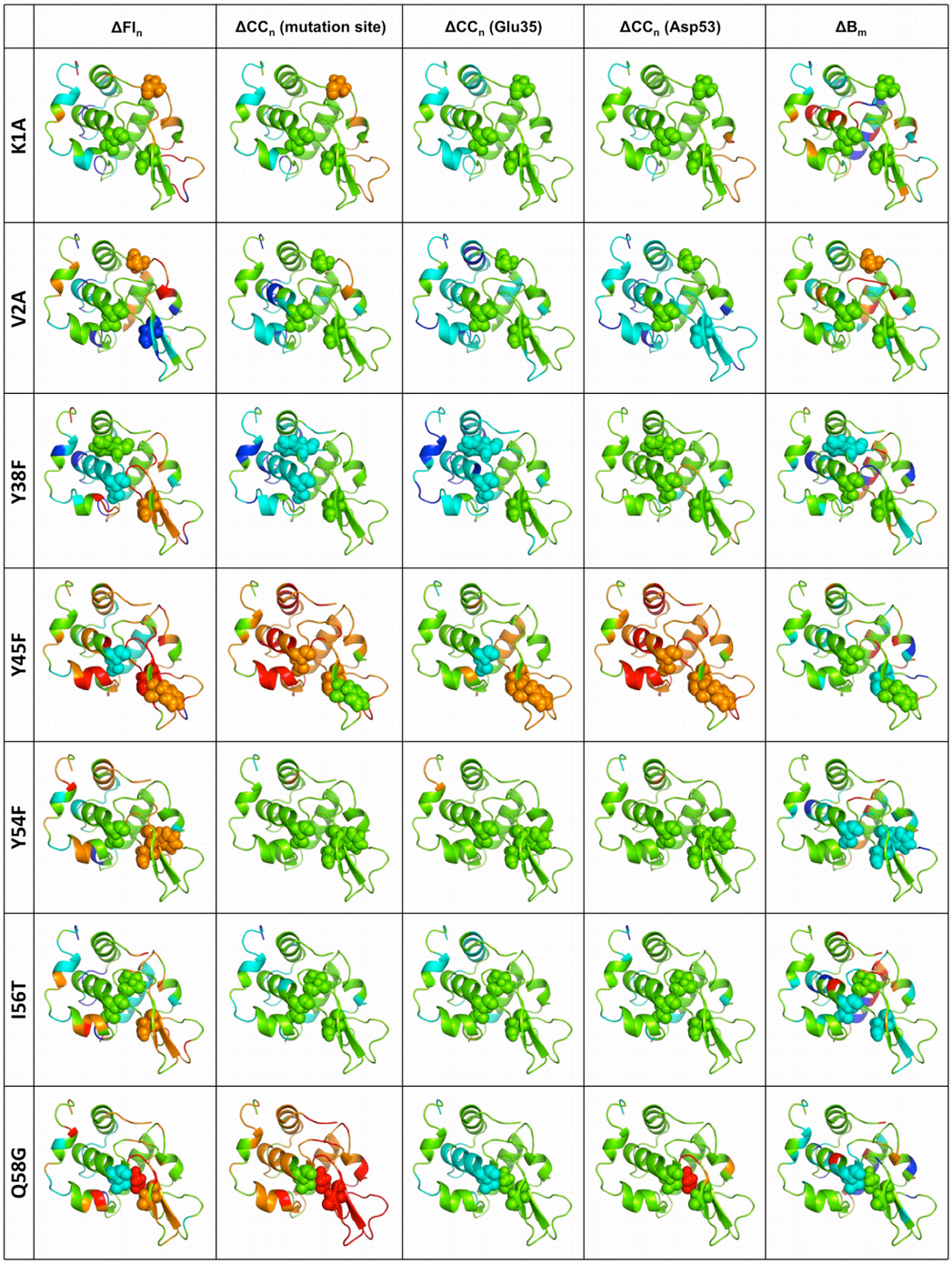

Figure 3.4: The affects of mutation on protein flexibility are mapped to structure. The five columns correspond to DFI_n, DCC_n with respect to Glu35, DCC_n with respect to Asp53, DCC_n with respect to the mutation site, and DB_{norm}. In all cases, the histogram bins in Figure 3.3 define the coloring schemes. The orientation of each protein is nearly identical across the figure. In each structure the catalytic pair (Glu35 and Asp53) and the mutated residue is rendered in spacefill. Importantly, this figure emphasizes the long-range nature of the response.

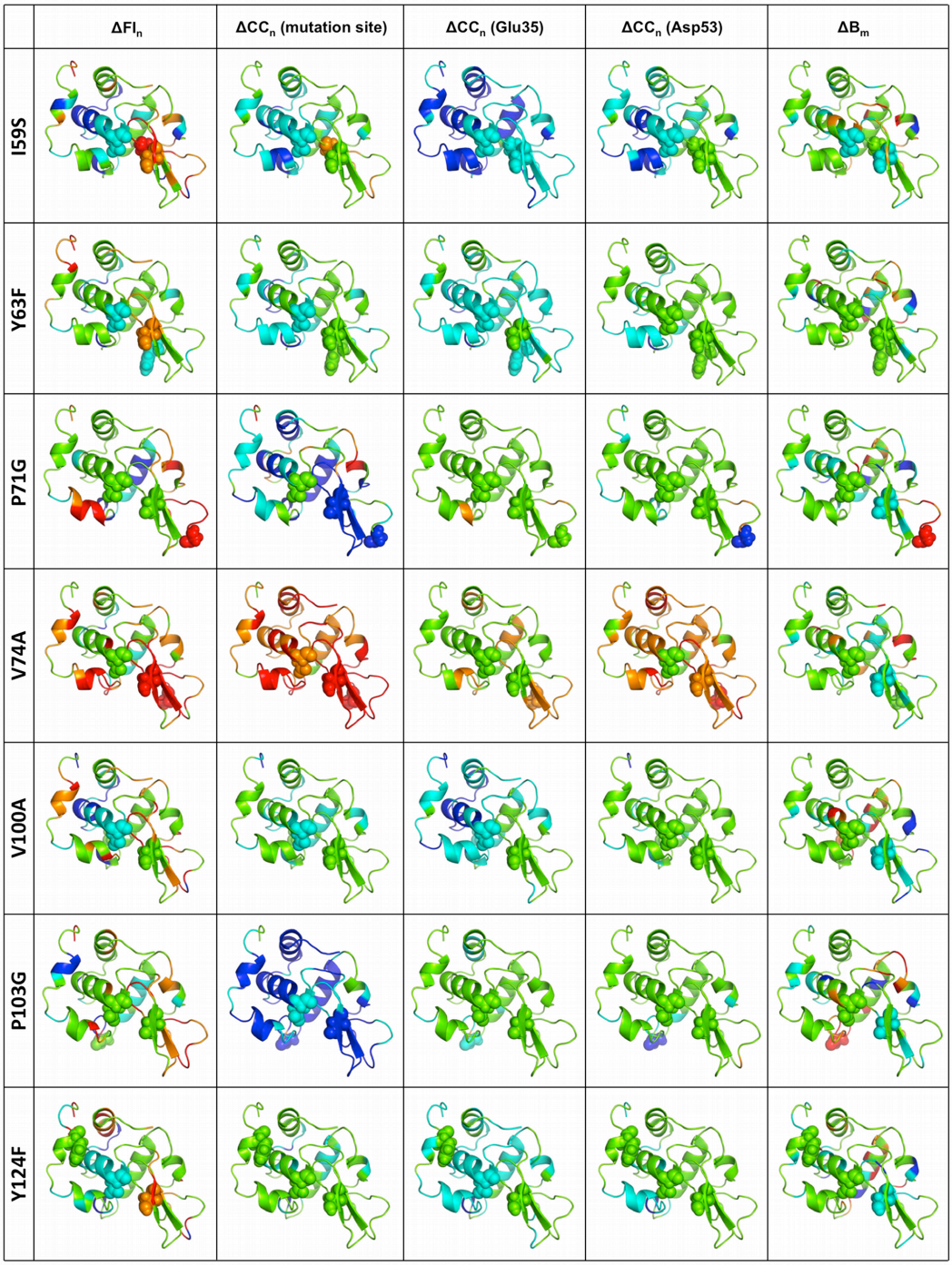

Figure 3.4: (continued)

response to allostery, where a perturbation method is used to identify putative allosteric sites [98]. Therein, small numbers of constraints have been introduced to mimic the effect of ligand binding, from which new QSFR properties are calculated using the same

structure. Large changes in QSFR metrics indicate an allosteric response. Application of this method to 3 CheY orthologs indicates that the most conserved response occurs within the β4/α4 loop, which is known to be important to propagation of the CheY phosphorylation signal [113, 114], yet residue-level response is quite variable, leading to the conclusion that allosteric response is both variable and conserved across the CheY family. The variability in ΔCC observed above further demonstrates diversity and sensitivity of allosteric response, which is consistent with observed variations within allosteric response across protein families [115].

The ubiquity of differences observed across sets of orthologous proteins, which is consistent with myriad experimental results, leads one to wonder about the origins of the familial divergence. That is, how many mutations are needed to observe significant differences in protein dynamics? In spite of the rather small structure differences, it is common for changes in flexibility to occur throughout structure, including at locations remote from the mutation site. As indicated by the histograms in Figure 3.3a, changes in both flexibility metrics are common. Specifically, while no change is the most frequent response, 42-48% of the residues undergo an appreciable change upon mutation. These distributions are obtained by sampling a collapsed dataset composed of all residues for each protein in the dataset (or as a variant to the method, across the entire protein except for a local window centered on the mutation site). This means that it is not the case that one particular mutant will make virtually no change, whereas another will make a large change. Rather, a typical mutant includes many sites with increased flexibility and increased rigidity throughout the protein. Exactly where the changes occur has a great variance in general, but the statistical expectation of having compensation between one

part of the protein increasing in rigidity while another part of the protein increases in flexibility seems very consistent across our dataset. The percentage of positions leading to increased backbone flexibility (27.9%) is slightly greater than the percentage increasing rigidity (20.0%). In summary, *changes in backbone flexibility upon mutation are common*, where local changes across the protein are typically composed of comparable amounts of an increase and decrease in flexibility distributed throughout the protein. Essentially, the protein is maintaining a global level of marginal mechanical stability within the native state at the melting temperature of the mutant. Changes in CC are also common; however, the differences between increased flexibility and increased rigidity are more asymmetrical. As discussed above, it is found that flexibility increases upon mutation tend to be localized, whereas increases in rigidity are likely to be coupled to remote structural sites. This result is not a matter of simple statistical chance that as more regions become rigid, the tendencies of these regions to coalesce into larger rigid regions increase. Rather, the increase in co-rigidity is counter-intuitive based on this reasoning, since there is an overall decrease in rigidity across the protein upon most mutations. This simultaneous effect suggests sparse and ramified rigid pathways are carved out by the mutations, which is critical to maintain marginal mechanical stability within the protein at its melting temperature. Here, critical means that further degradation of this pathway is likely to lead to unfolding as rigidity in the protein is lost [116].

To further support the conclusion that changes in flexibility upon mutation are common, we also assess the flexibility differences between human wild-type and hen egg white lysozyme (HEWL). Figure 3.6a compares changes in HEWL backbone flexibility (relative to human wild-type) to the mutant changes summarized above. Surprisingly, the

number of differences between the two orthologs is generally slightly less than observed within the mutant dataset. While, on average, 48.0% of the mutant positions have a change in FI, only 41.1% of the HEWL positions changes. Although there is relative decrease in number of flexibility differences, the number of changes that do occur is statistically significant ($p = 2.0E-7$). Moreover, the scale of the ΔFI values for HEWL falls within the variation across the human mutant dataset despite the fact that the pairwise sequence identity is only 61%. That is, even with a significantly reduced sequence identity, there are no wholesale differences in flexibility. Put otherwise, the changes in backbone flexibility within the mutant structures are clearly large since they are on the same scale as the much more divergent HEWL ortholog. Similarly, the HEWL ΔCC_n results (Figure 3.6b) are also easily within the mutant dataset range established in Figure 3.5.

It is worth noting that our dataset composition is inherently biased towards rigidity. That is, the studied mutations are all amendable to crystallography, which eliminates many possible mutations that destabilize the structure so much that it is too flexible to form a crystal lattice. As such, our conclusions regarding the frequency and extent of flexibility changes would be even greater if it were feasible for us to study all possible mutations because extreme increases in flexibility upon mutation are actually mutations because extreme increases in flexibility upon mutation are actually underrepresented in our dataset.

3.4.2 Changes in flexibility can be long ranged

We have segregated responses into moderate and large changes (Figure 3.3). As expected, moderate changes are the most common, but large changes in FI and CC also

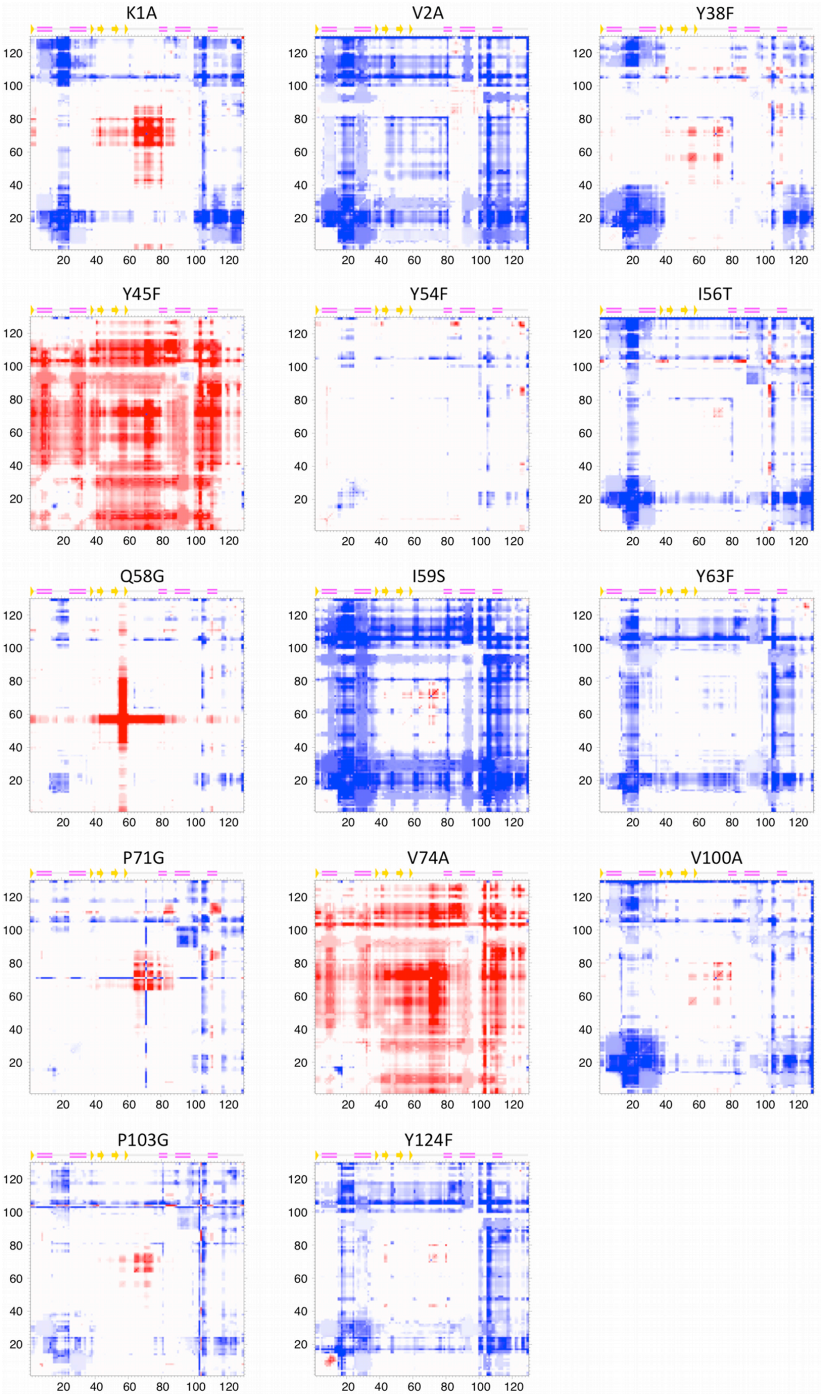

Figure 3.5: Cooperativity correlation (CC) difference plots show the differences in pairwise mechanical couplings between each mutant structure and the wild-type reference. Red indicates increased correlated flexibility within the mutant structure, whereas blue indicates increased correlated rigidity. Juxtaposed to the ΔFI results that show significant uniformity within their response, the ΔCC_n values are highly variable across the set of mutants.

occur frequently (respectively, 18.3 and 13.5% of the time). While the definition distinguishing between moderate and large is somewhat arbitrary, the ubiquity of large changes is clearly shown in Figure 3.2. Moreover, large changes in backbone flexibility can occur anywhere in structure, but some clustering is evident. Specifically, large increases in rigidity are more likely to occur within the α1/α2 and α4/α5 loops, whereas large increases in flexibility tend to occur within the β-subdomain. Conversely, there is little clustering of CC response. These two opposing observations further emphasize our previous results that FI is strongly related to overall structural topology, whereas CC is highly sensitive to small differences within the H-bond network [100].

The visual survey of the first column in Figure 3.4 shows that changes in flexibility are rarely localized around the mutation site, but rather generally propagates over long distances. This observation is confirmed by the counts in Table 3.3. However, skewness in raw counts can be expected by the increased number of sites that are present in the strata corresponding to larger distances. Interestingly, the ratio of changes to no change for short, medium and long distances are all nearly equal to one (with the two exceptions explained above in the results section). The similarity in the ratios is somewhat surprising because the naïve expectation is that short-range changes would be much greater than long-ranged due to dampening effects. As such, these results indicate that *changes in protein flexibility upon mutation can be long-ranged.* Upon further statistical analysis, it is found that the ratios are not regimentally affected by solvent accessibility of the mutation or response site. In addition, the distance between the mutation-response pair has no systematic affect, meaning that neither structural distance nor solvent accessibility has a large biasing affect on the results. The sole exception being that mutations at

solvent exposed positions is less likely to lead to changes in flexibility. Note that there are insufficient data to perform a statistically significant two-dimensional stratification that considers both response residue and mutant accessibilities.

3.4.3 Relating computational and experimental observations

Our results collectively indicate that point mutants cause a rich and diverse set of flexibility changes throughout structure. Generally, changes in both flexibility and rigidity within the protein upon mutation occur concurrently to maintain marginal mechanical stability at the new melting temperature. Many changes are localized, but significant portions propagate over surprisingly long distances. While we cannot make a direct quantitative comparison to experimental results because the observed response properties are fundamentally distinct, changes in NMR order parameters show similar response richness. For example, many reports have used N-H S^2 order parameters to demonstrate that changes in backbone dynamics can be quite large upon mutation (e.g., see [117-121]), yet the magnitude of the changes are generally within the scale wild-type order parameter distributions [122]. The observed changes in backbone flexibility are qualitatively equivalent (Figure 3.6). Moreover, localized increases in dynamics have been observed despite globally similar average structures [123] and stabilities [124] between the wild-type and mutant proteins. Particularly noteworthy are experimental results that mirror the complexity that we uncover on lysozyme on two additional small model-system proteins. First, concurrent increases in dynamics and rigidity have been demonstrated in the V54A *Eglin c* mutant [125], which epitomizes the changes in lysozyme flexibility within in Figures 3.2 and 3.4. Second, long-ranged changes in dynamics have been observed within the F22L and A20V mutants of protein L [126],

which is again shown for changes in lysozyme flexibility in Table 3.3.

Methyl sidechain S^2 order parameters characterize ps-ns timescales, whereas backbone S^2 order parameters characterize slower motions. While the DCM does not model dynamical timescales per se, experimental investigations that probe both further underscore the complexity and long-range nature of changes in protein dynamics upon mutation. For example, Igumenova et al. demonstrated that calmodulin backbone dynamics are largely unchanged upon mutation [127]. However, sidechain motions are significantly altered by the D58N mutation in the Ca^+-binding loop, which are spread over long distances. Interestingly, the pseudosymmetric D95N mutation has no appreciable affect on sidechain dynamics. Similarly, Clarkson and Lee characterized two valine-to-alanine eglin c mutants [128]. Large dynamical changes were observed as much as 13 Å from the mutation site. The V54A actually causes a network of residues to increase in rigidity despite the fact that the mutation is thermodynamically destabilizing. Changes in the V14A mutant, which is also buried in the core of the protein, were much less. This diversity of response led the authors to conclude, "*...dynamical responses will be context-dependent,*" which is epitomized by our lysozyme dataset. That is, the affects of mutation are quite varied and highly dependent upon the local details of the perturbation, which propagate in complex and unexpected ways.

The Dobson lab has characterized dynamical changes in lysozyme, with a special focus on mutant amyloidogenicity. In particular, changes in I56T and D67H were studied using hydrogen/deuterium exchange NMR and mass spectrometry [129]. (Note that the I56T mutation is included within our dataset.) They showed that b-subdomain dynamics in the D67H mutant are changed extensively, whereas changes occur much less in the

I56T mutant. This result broadly agrees with our results, which indicate that I56T dynamics are changed much less than mutants with the biggest responses (e.g., Y45F, I59S, V74A, and V100A). Taken together, our conclusions are therefore in line with many experimental characterizations of changes in protein dynamics upon mutation.

3.4.4 Amyloid formation and the β-subdomain

Based on our previous investigations, we believe the above results could be generalized to most globular proteins. In addition, our results also reveal an interesting effect specific to lysozyme. That is, a large number of mutations, regardless of location or type, cause increased flexibility within the β-subdomain, which in many cases can be thought of as local unfolding. This point is noteworthy for two reasons. First, this result again highlights the long-range nature of dynamical changes because many of the mutations occur outside of the β-subdomain. Second, several experimental reports have suggested that mutations leading to amyloid in lysozymes and the related α-lactalbumins are due to structural changes, which may include local unfolding, in the β-subdomain [82-86]. As such, the partially unfolded β-subdomain may serve as a nucleation site for amyloid growth. Of course, our results do not address this issue, but they do parallel the earlier experimental conclusions. For example, ΔFI clearly indicate that the amyloidogenic I56T mutation has increased flexibility within β-subdomain (Figure 3.4). Similarly, our results indicate that several other mutants display at least as much flexibility therein, including K1A, Y38F, Y45F, Q58G, I59S, P71G, V74A, V100A and P103G. As such, it is tantalizing to consider that they might also be amyloidogenic. We have searched the literature and, to the best of our knowledge, these mutants have not been characterized. We therefore present them as blind predictions, and hope that others

will consider characterizing their amyloidogenicity.

3.4.5 Relating the observed changes to protein family evolution

Across the dataset, changes in protein flexibility upon mutation are common, large and can be long-ranged. That is, the stark variation in dynamics observed across protein families unexpectedly occurs early in the divergence process through a combination of flexibility increases and decreases. However, the observed changes seldom significantly alter global flexibility. The relative similarity in positive and negative ΔFI_n values suggest that as divergence occurs, marginal mechanical stability is generally maintained because only incremental overall changes will be typically encountered by any given mutation. In other words, a single mutation will typically not overwhelmingly rigidify the protein nor overwhelmingly increase flexibility. Rather, structure subtly rearranges in response to the mutation to maximize enthalpy-entropy compensation. That is, a global increase in rigidity creates a large reduction of conformational entropy that is unfavorable, and a global increase in flexibility creates a large loss in enthalpy (weakened native contacts) that is unfavorable. Thus, the native state ensemble of the protein seeks to find the lowest free energy that typically requires a balance between flexible and rigid structural regions, suggesting that a mixture of rigidity and flexibility is typical at physiological conditions.

These results suggest that global increases in rigidity or flexibility upon mutation are rare because the local responses are derived in the noise (random fluctuations) around overall being neutral, with only a slight advantage towards increased flexibility in this case. The implication of the above is that successive mutations during the evolutionary process are generally necessary to substantially alter global flexibility characteristics.

Viewed from a dynamics point of view (excluding selection in maintaining function), the process is a random walk capable of nudging the protein towards global increases in rigidity or flexibility. However, conservation of function is likely to select against systematic drift that leads to large differences in flexibility with respect to the function and stability of the wild-type protein. In that vein, the suppressed flexibility differences observed in HEWL actually suggest that additional compensating mutations can reestablish desired dynamical properties. For example, the similarity between human wild-type and HEWL β-subdomain flexibility is very persuasive given how susceptible this region appears to be to increased flexibility within the point mutants (Figure 3.6).

The changes observed in Figure 3.5 indicate that a single mutation is sufficient to significantly alter global CC properties, where the accumulative effect of a few mutations should be sufficient to go beyond the range of differences we have observed across protein families. As successive mutations appear, conservation of function again provides the selection bias for proteins to maintain globally similar dynamics while evolving to varying stability characteristics. This scenario explains the considerable diversity in detailed dynamical changes occurring from a single point mutation, while general statistical characteristics remain robust.

3.5 Conclusion

In this report we demonstrate that changes in human c-type lysozyme flexibility upon mutation are frequent, large, and can be long-ranged. Depending upon metric tracked, residue-specific flexibility is changed 42-48% of the time across the dataset. The mutation-induced structural perturbations propagate over long distances. In fact, the average distance between the mutant and affected residue is 17-20 Å. While direct

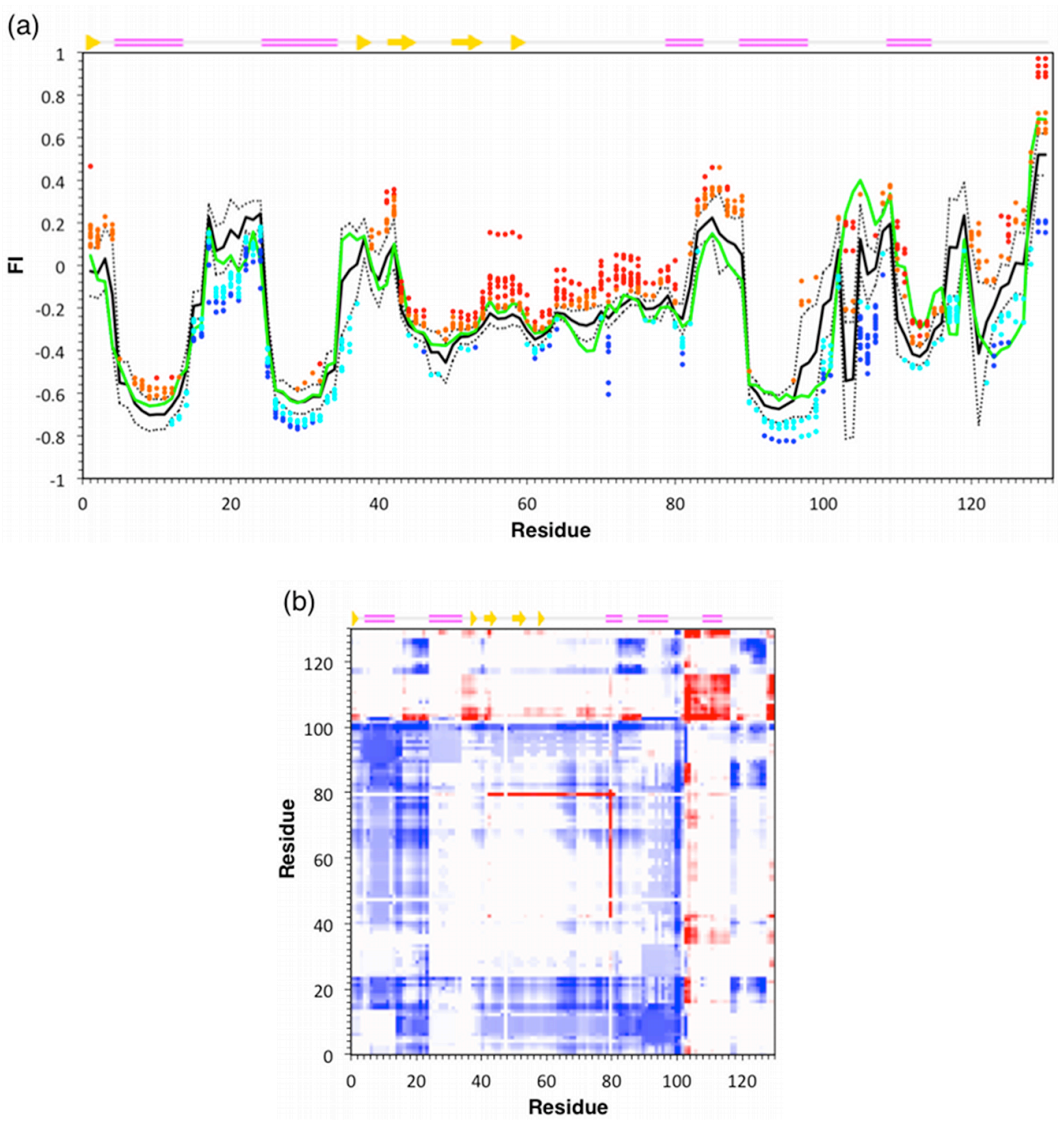

Figure 3.6: Mutational affects on flexibility. (a) Lysozyme backbone dynamics are characterized by a flexibility index (FI). Positive FI values measure flexibility, whereas negative values measure rigidity. The structure is isostatically (marginally) rigid when FI = 0. The black solid line indicates the average human wild-type lysozyme profile, whereas the dashed lines indicate ±1 σ. The mutant sites that moderately score beyond the background are indicated using the same coloring scheme as Figure 3.3. The green solid line indicates hen egg white lysozyme backbone flexibility (HEWL), which is generally more similar to the wild-type profile than the human mutants. (b) The difference between human wild-type lysozyme and HEWL cooperativity correlation is shown. The coloring scheme is the same as in Figure 3.5.

quantitative comparisons to experiment are impossible due to different physical response

characteristics studied and lack of experimental characterizations on most of the dataset, the frequency, scale and complexity that we find in flexibility changes are principally consistent with multiple NMR characterizations of mutant dynamics in a variety of proteins, including lysozyme. Intriguingly, we have shown that changes in flexibility upon single site mutation are generally larger than differences between hen egg white lysozyme (HEWL) ortholog to the human wild-type. In particular, most mutants lead to increased β-subdomain flexibility; however, β-subdomain flexibility within the human and HEWL ortholog remains conserved. Based on a random selection of mutations, this result is highly improbable because the human and HEWL lysozymes only have 61% sequence identity. As such, we hypothesize that evolutionary compensating mutations in HEWL have reestablished desired properties. Going further with these important evolutionary observations we will try to elucidate the conservation and variation in backbone flexibility and other QSFR properties across protein families in the next chapter.

CHAPTER 4: VARIATIONS WITHIN CLASS-A β-LACTAMASE QSFR AND PHYSIOCHEMICAL PROPERTIES REFLECT EVOLUTIONARY, BUT NOT FUNCTIONAL, PATTERNS

4.1 Introduction

The bulk of our knowledge concerning protein family evolution has come from comparative analyses of the large body of sequence and structure data produced over the last five decades. While this data has been invaluable to our current understanding, sequence and static structural descriptions provide only a narrow glimpse into functional mechanisms. Consequently, there has been a growing interest to include structural and functional details into molecular-evolutionary analyses [130-132]. In our previous chapters we discussed stability and flexibility changes in human lysozyme proteins due to point mutations. An important observation regarding conservation of backbone flexibility between human and HEWL lysozymes clearly indicate evolutionary relationships. But, for a complete understanding of these relationships, both conservation and variation must be characterized across a protein family. Since conservation is the ultimate evolutionary driving force [133], protein orthologs tend to be significantly more similar in function than paralogs, and this functional similarity holds true with increasing sequence divergence as well [134].

In this evolutionary characterization study or "breadth analysis", we select β-lactamase (BL) enzyme family that provides an excellent mix of preserved and adaptable biophysical properties requiring evolutionary/functional relation interpretation. On the

functional aspect, BL enzymes have a chemically diverse set of substrates. Moreover, many BL enzymes can act on the same substrate despite being from evolutionarily distinct out-groups, leading to questions related about the presence (or absence) of conserved mechanistic strategies. In this report we seek to determine if conserved biophysical properties underlie the functional differences across the BL enzyme family.

Antibiotic resistance continues to outpace our ability to bring new antibiotic drugs to market [135], leading to substantive fears about our continued ability to combat bacterial infections that are currently relatively benign. Central to this growing global health concern is the bacterial enzyme β-lactamase (BL), which is produced by some bacteria [136]. BL confers resistance to penicillin and related antibiotics by hydrolyzing their conserved 4-atom β-lactam moiety, thus destroying their antibiotic activity [137]. Bacteria of all species depend on a cross-linked peptidoglycan layer, which preserves cell shape and rigidity. This peptidoglycan layer is primarily composed of alternating β(1,4)-linked monosaccharides, specifically N-acetylglucosamine and N-acetylmuramic acid. The latter is modified by a pentapeptide that always ends with two D-alanine residues. Cross-linking of peptidoglycan units is catalyzed outside the cytoplasmic membrane by cell wall transpeptidase enzymes. In this cross-linking process, a peptide bond is formed between penultimate D-alanine on one chain and pimelic acid (in Gram-negative) or L-lysine (in Gram-positive) residue on the other. The terminal D-alanine is cleaved off after the linkage is formed with the penultimate residue. β-lactam antibiotics effectively inhibit bacterial transpeptidases, consequently they are often called penicillin binding proteins (PBP). By inhibiting cell wall synthesis, the bacteria become highly susceptible to cell lysis.

In response, bacteria have evolved BL enzymes to defend themselves against β-lactam antibiotics. BL has, in fact, evolved from the functional domain of PBP through the acquisition of the new hydrolase activity [138]. The BL enzyme family is broad and is characterized by varying degrees of antibiotic resistance activity. In fact, extended spectrum β-lactamases (ESBL) also confer resistance to cephalosporins, which had previously eluded BL hydrolysis [139, 140]. ESBLs are evolved from traditional BL genes, generally through mutations within the active site [141, 142], thus highlighting the critical importance of subtle differences within members of the BL family.

To date, more than 470 BL enzymes have been identified and are typically classified into 4 classes (A to D) based on sequence similarity [143]. Bush *et al.* developed a classification scheme for BL proteins based on their functional characteristics [144]. Protein structures belonging to classes A, C and D have similar folds and all have a mechanism that involves a catalytic serine residue, whereas class B enzymes are zinc metalloenzymes that have a distinct fold. In this work we characterize the most clinically relevant class-A family.

Previous works have highlighted how conservation of electrostatic properties can mediate conserved function across a protein family, withstanding sequence and structural variability during evolutionary processes [145, 146]. In this report, we elucidate the variation and conservation of electrostatic and QSFR properties in class A BL family. Our dataset includes twelve BL enzyme structures, each originating from different bacterial species. Here, we show that – as expected – conservation of various electrostatic and dynamical properties is a common notion used by protein families to maintain function. However, we also observe a striking number of differences; however, these

differences do not correlate with antibiotic resistance patterns. Rather, the biophysical variations are explained by evolutionary relationships, suggesting that convergent resistance specificities utilize distinct biophysical mechanisms.

4.3 Results and discussion

4.3.1 Conservation and variation in residue pK_a values

Due to their clinical significance, serine-based class-A β-lactamase proteins are one of the most widely characterized enzyme families. The catalytic mechanism involves acylation of residue Ser-70 at the active site. However, identification of the general base that activates this serine residue has always been a subject of controversy. As such, two distinct residues have been proposed. While one hypothesis suggests that this role is played by the conserved Glu-166 [147-150], the other proposes Lys-73 [151-153]. In support of the first hypothesis, crystallographic data and MD studies [150] have suggested the presence of a conserved bridging water molecule that might act as a relay molecule for the transfer of proton between Ser-70 and Glu-166. Based on other experimental studies involving Glu-166 mutation [153, 154], the second hypothesis proposes an unsymmetrical mechanism involving two different general bases, Lys-73 and Glu-166 that carry out acylation and deacylation respectively. Swaren *et. al.* [155] have argued that substrate binding raises the pK_a of Lys-73, which contributes to lowering of energy barrier for Ser-70, highlighting the importance of Lys-73 in proton transfer. Conversely, kinetic studies of several Glu-166 mutant enzymes [156] have displayed decreased rates of acylation and deacylation, emphasizing that Glu-166 is more important. Due to this absence of Glu-166 negative charge in mutant proteins, the Lys-73 side chain exhibits a lower pK_a shift, acting as an alternate general base in hydrolyzing β-lactam

ring [157]. Going by either hypothesis, the experimental studies described above have convincingly brought out the importance of both Lys-73 and Glu-166.

The protein sequence contains information about its structure, which in turn dictates its function. This highlights the basic principle of protein evolution, i.e., conservation of function. Several other residues have also been identified in BL that are catalytically important: Ser-70 being the primary catalytic residue; Lys-73, Glu-166, Ser-130, Lys-234 as secondary catalytic residues. Finally, Asn-136, Arg-164, Asp-179 are other important residues that maintain the active site structure (Figure 4.1). All of these residues are in spatial vicinity of Ser-70 and affect substrate recognition and catalysis. Detailed sequence and structural comparison across the class A family has identified similar structural and functional elements that span over active site residues mentioned above [158-161]. These conserved elements are SXXK, SDN, EXXLN and KTG.

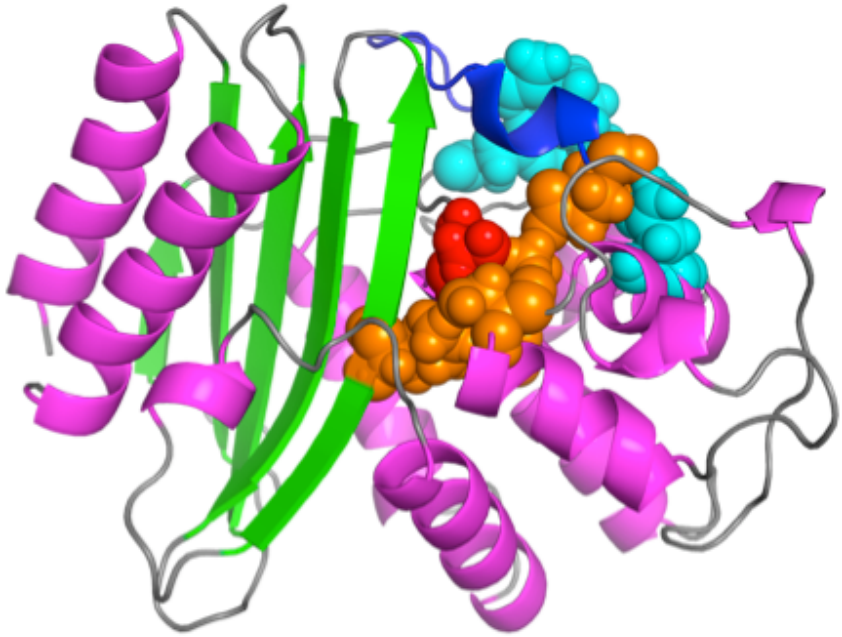

Figure 4.1: Structure of a Class A β-lactamase enzyme. The active site is located at the domain interface. The catalytic residue Ser-70 is shown in red. Other catalytic residues are shown in orange, whereas the Ω-loop is shown in blue at the top. Residues that maintain the structural integrity are shown in cyan.

Conservation of important electrostatic properties is also a commonly employed mechanism that leads to conserved function. Figure 4.2a shows calculated residue pK_a shifts (shifted away from their model values) across twelve BL proteins. Interestingly,

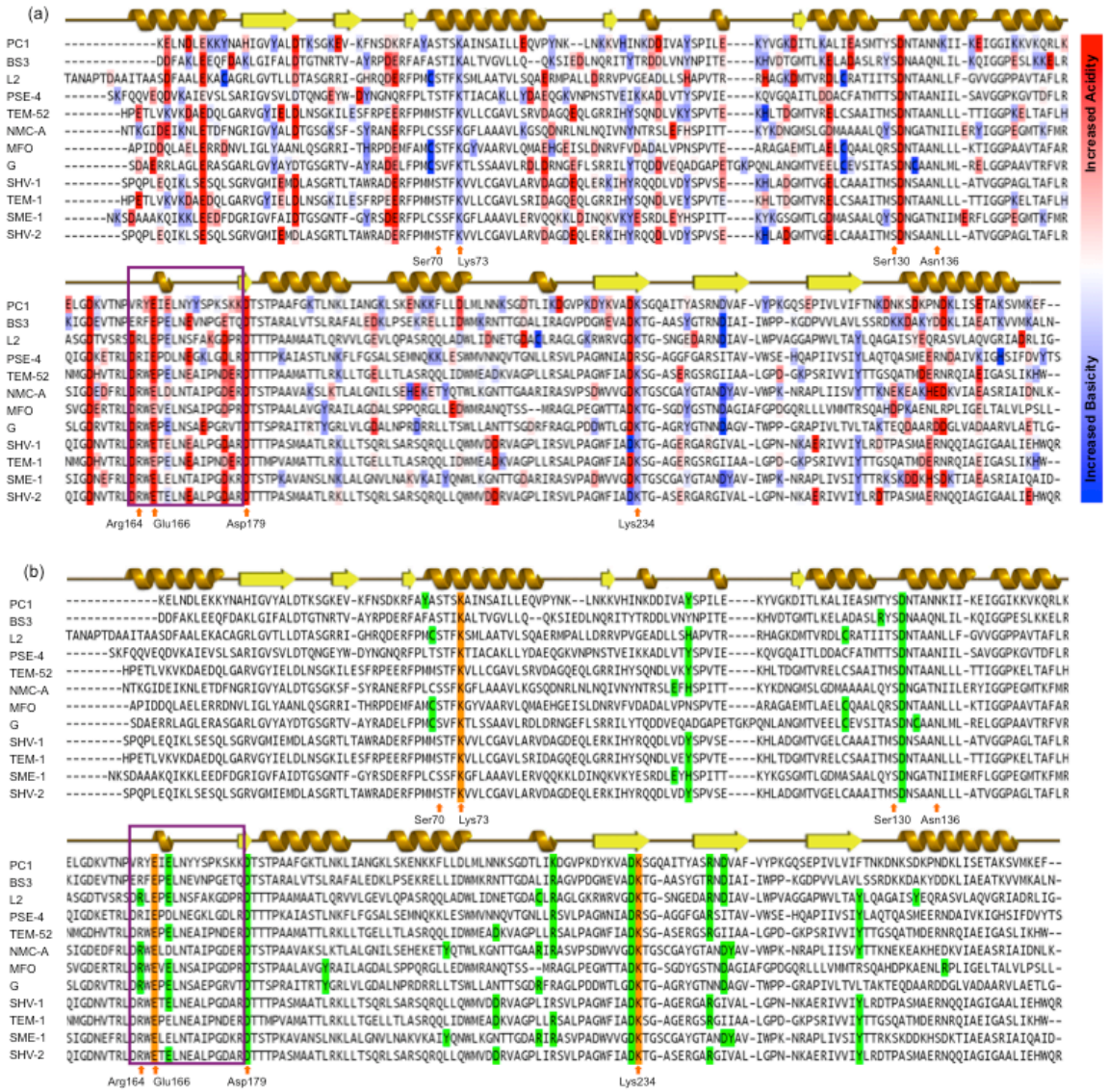

Figure 4.2: Electrostatic properties of β-lactamase family. (a) Multiple sequence alignment of twelve β-lactamases color-coded by shifts in residue pK_a values from model values. Residues colored red express increase acidity, whereas residues colored blue show increase in basicity. (b) Residues colored green exhibit strong electrostatic interactions with catalytic residues colored orange (> |±1| kcal/mol). The identified residues are also highlighted in the β-lactamase structure provided in Figure 4.1c. The Ω-loop region is indicated by the purple box. A cartoon representation of secondary structure is displayed on top of each alignment, while active sites are displayed below.

these pK_a shifts are mostly conserved, emphasizing a common mechanistic strategy. We further investigate the site-site interactions of residues that have strong electrostatic interactions (more than 1 kcal/mol) with the secondary catalytic residues Lys-73, Glu-166 and Lys-234 (Figure 4.3). Remarkably, all conserved electrostatic sites overlap with the four conserved element regions, highlighting the strength of the active site electrostatic forces. All pairwise active site interaction energies are listed in Table 4.1.

More interestingly, all these sites have a conserved pK_a shift. Asp-131, Glu-166, Asp-179 and Asp-233 display strong acidic character, whereas Lys-73 and Lys-234 exhibit weak conserved basic behavior. Lys-73, which acts as proton extractor from Ser-70, needs to be deprotonated for acylation. As such, there is a cationic electrostatic microenvironment surrounding Lys-73, which is created by nearby basic residues Lys-234 and Arg-244 [153]. When Arg-244 is missing (which is the case in the NMC-A,

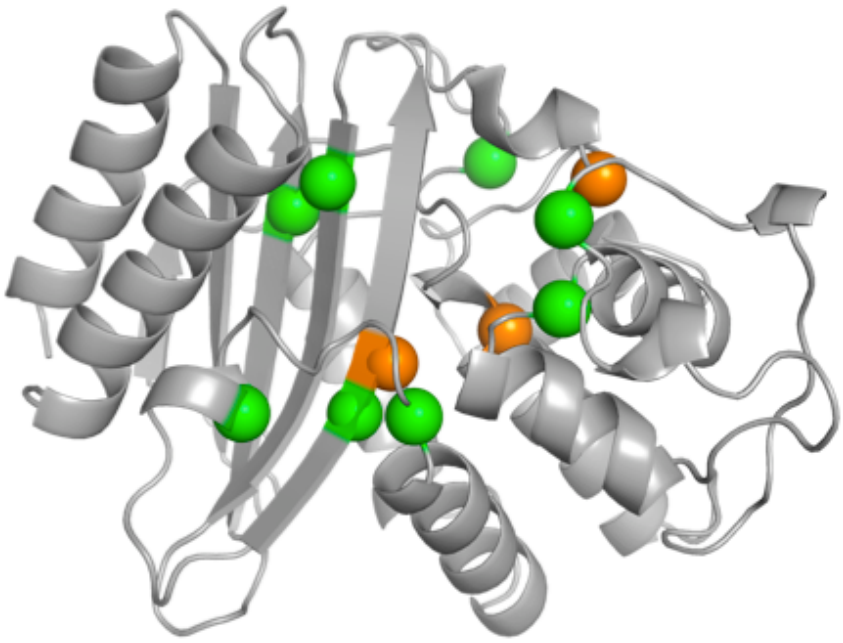

Figure 4.3: Conserved electrostatic networks (cf. Figure 4.6b) are mapped to a BL structure. Green colored spheres represent α-carbons of residues interacting strongly with catalytic residues, which are highlighted in orange.

Table 4.1: Summary of the active site electrostatic network. With the exception of the last column, values in the upper triangle provide the minimal distance (in Å) between atom pairs in the two residue side chains, whereas values in the lower triangle quantify the pairwise electrostatic potential energy (expressed in kcal/mol). Values in the last column provide the minimal distance (in Å) between atom pairs of the electrostatic network residues and the catalytic Ser-70. The reported values are the average across the dataset, and the coefficient of variation is shown in the parentheticals (expressed as a percent).

	Lys-73	**Asp-131**	**Glu-166**	**Asp-179**	**Asp-233**	**Lys-234**	**Ser-70**
Lys-73	--	6.0 (8)	2.8 (10)	10.9 (5)	11.2 (5)	4.1 (8)	3.1 (13)
Asp-131	-3.4 (18)	--	8.1 (5)	16.7 (3)	14.6 (5)	8.3 (5)	8.7 (7)
Glu-166	-7.6 (15)	1.7 (12)	--	8.1 (6)	15.0 (4)	7.7 (7)	3.4 (28)
Asp-179	-1.0 (12)	0.3 (9)	1.7 (10)	--	19.3 (3)	13.9 (5)	10.2 (5)
Asp-233	-0.9 (10)	0.4 (11)	0.5 (8)	0.3 (11)	--	5.3 (4)	11.3 (8)
Lys-234	4.0 (13)	-1.5 (10)	-1.7 (10)	-0.6 (14)	-3.3 (13)	--	4.7 (10)

MFO and G orthologs), this role is acquired by Arg-164 as shown in our active site electrostatic networks plot (Figure 4.2b and 4.3).

Another important feature of BL proteins is the Ω-loop (comprising of residues 163-178) that is involved in substrate recognition. Additionally, the Ω-loop comprises Glu-166, which is critical for deacylation activity. Our results reveal a strongly conserved acidic behavior within Glu-166, which activates a water molecule in the vicinity to attack carbonyl carbon of the acyl-enzyme. This ensures a back-delivery of the abstracted proton to Ser-70 γ-O atom, leading to enzyme regeneration [150].

4.3.2 Conservation and variation in electrostatic potential maps

The above results highlight the importance of conserved local electrostatic properties. Conversely, unlike the archetype example of copper, zinc superoxide dismutase [145], Figures 4.4 demonstrates that global electrostatic potential maps can be quite varied across the whole family. Yet potential maps are conserved within evolutionary outgroups. For example, the closely related and clinically relevant orthologs TEM-1, TEM-52, SHV-1 and SHV-2 have similar electrostatic potential maps. Similarly, other outgroups

conserve visual electrostatic features.

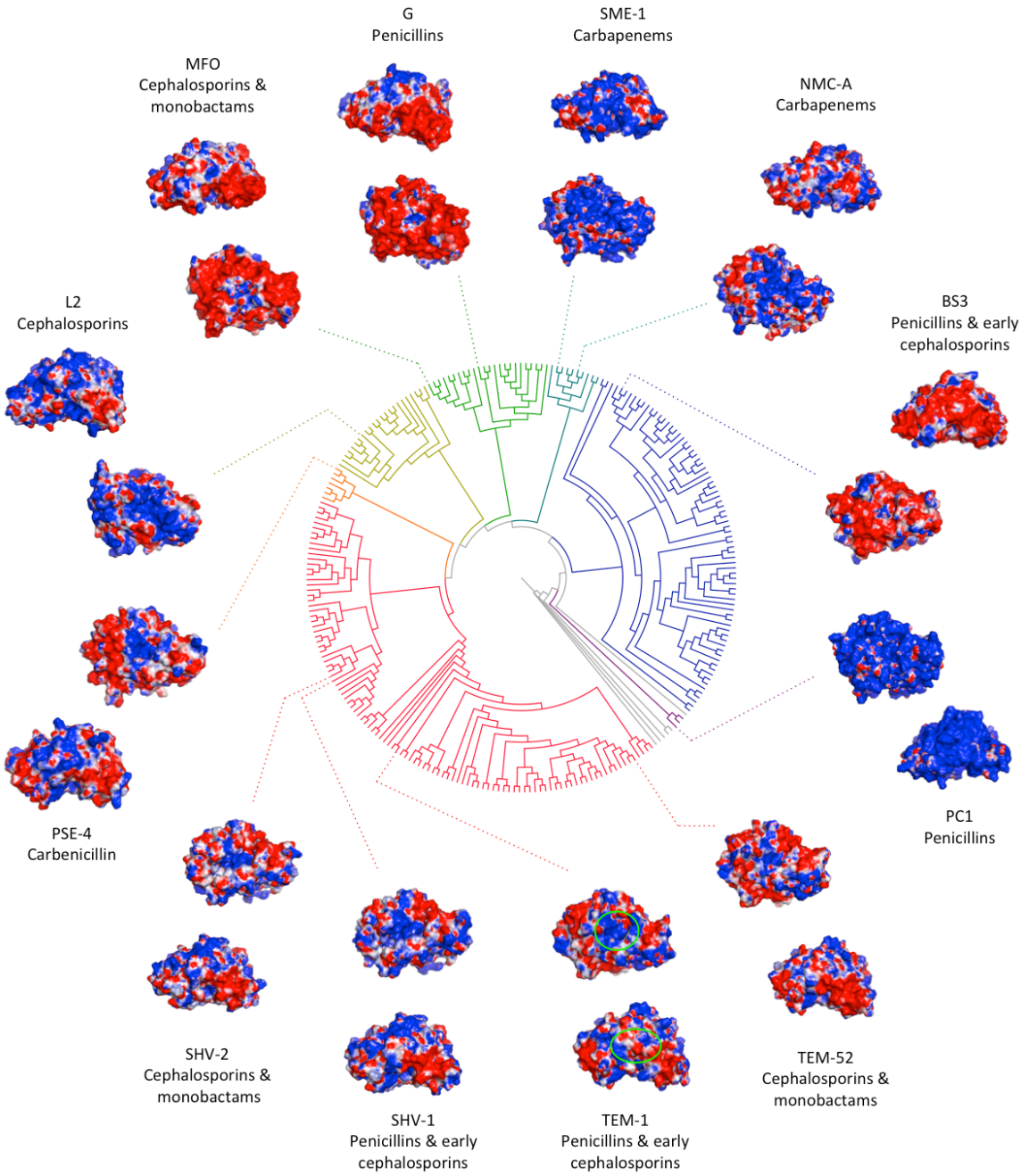

Figure 4.4: Sequence and electrostatic property relationships. The class-A β-lactamase family phylogeny is shown, which differentiates into 7 subgroups using a constant cut-level. Outgroups are represented by a unique color for better visualization. The structures closest to the phylogeny are oriented to highlight the active site region, which is indicated in green in the TEM-1 ortholog. Structures in the outer ring have been rotated in the y-direction by approximately 90 degrees, which highlights the Ω-loop region, also indicated in green.

Differences within the electrostatic potential maps are not unexpected owing to the sequence and structural variability within the dataset. Pairwise sequence identities range from 27% to 98%, which propagates to α-carbon root mean square differences up to 2.6 Å. Moreover, the net charge of these twelve enzymes ranges from -6 to +15 (Table 4.2). This large structural variation with distinct electrostatic properties raises the question, "*How does nature maintain the common functionality of enzymes?*" Key sequence/structure motifs provide an insight into the description of the underlying conservation.

Table 4.2: Electrostatic and H-bond characterization. Coefficient of variation = standard deviation / mean * 100.

Enzyme	SXXK Charge	EXXLN Charge	Ω-loop Charge	Total Charge	Total HB Energy	Number of HB	Avg. HB Energy
G	1	-2	-2	-3	-1631.3	559	-2.9
TEM-1	1	-2	-4	-5	-1609.6	513	-3.1
NMC-A	1	-2	-3	1	-1588.3	520	-3.1
SME-1	1	-2	-1	7	-1728.4	550	-3.1
PSE-4	1	-2	-2	-4	-1548.0	529	-2.9
TEM-52	1	-2	-4	1	-1590.9	557	-2.9
L2	1	-2	-1	2	-1534.6	528	-2.9
SHV-2	1	-2	-3	1	-1633.2	539	-3.0
SHV-1	1	-2	-3	1	-1570.7	532	-3.0
MFO	1	-2	-2	-6	-1503.0	507	-3.0
PC1	1	-2	2	15	-1559.3	495	-3.2
BS3	1	-2	-4	-5	-1669.1	522	-3.2
Average	1	-2	-2.3	0.4	-1597.2	529.3	-3.0
CV	0.0%	0.0%	76.1%	1430.7%	3.9%	3.7%	3.8%

Sequence conserved regions SDN and KTG carry a conserved charge of -1 and +1 across all twelve BL enzymes. Interestingly, the other two key regions with mutable sites, SXXK and EXXLN, have conserved charge of +1 and -2 respectively. EXXLN lies within the 16-residue Ω-loop (XRXEXXLNXXXXXXXX) that maintains an overall negative charge (except PC1) ranging from -2 to -4. The conserved electrostatic

properties of key regions range from simple local conservation of charge to complex evolutionary origins of BLs. Conservation of charges at mutable motifs and Ω-loop are achieved through concerted mutations. When there is a charge changing mutation at these important electrostatic regions, there is a charge compensating mutation elsewhere.

An important observation in electrostatic potential maps is the presence of a conserved negative electric field spanning over α-helix H3, H4, H6 and Ω-loop (Figure 4.5). As discussed above, this Ω-loop region is rich in electrostatic interactions that maintain the structural integrity and hence the enzyme activity. It has been suggested that the conservation of charge and electrostatic interactions at the secondary structure level is important for electrostatic steering purpose and delicately maintaining the energetic stable active site region. Similar conclusions are drawn in [145, 146], where authors have highlighted the importance of conservation of electrostatic properties within enzyme

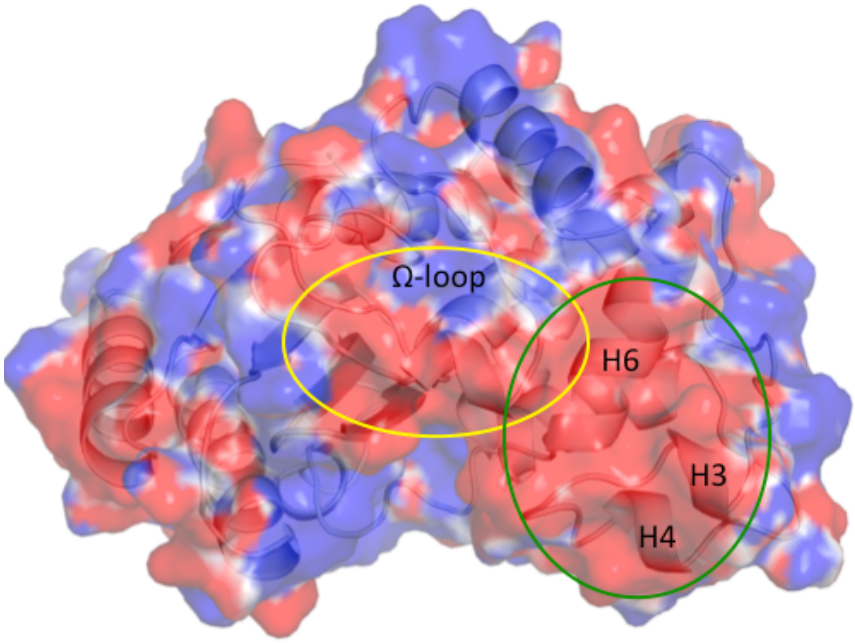

Figure 4.5: Electrostatic potential values of ±1 kcal/mol are mapped to the protein surface. Red indicates negative potential, while blue indicates positive potential. The structure is oriented to display a conserved negative potential region at the interface of the Ω-loop and helices H3, H4 and H6.

families. Through steering forces, the substrate is directed towards the active site of the enzyme by the balanced electric field around the enzyme. Electrostatic potential maps of BLs in Figure 4.4 compare electrostatic profiles to derive functional similarities and evolutionary relationships in protein family. No evolutionary grouping can be established based upon substrate type. This is not surprising because BL proteins are under heightened evolutionary pressures from continued antibiotic overuse, thus leading to extended spectrum antibiotic resistance. Experiments have suggested that a BL protein can expand its substrate spectrum with just one amino acid substitution [162]. In all, the above results demonstrate that evolution maintains a subtle balance in sequence/structure/function relationships, which are complex and difficult to decipher. Nevertheless, the electrostatic analyses bring out the key residues that overcome this evolutionary pressure and preserve the overall electrostatic environment of the protein family.

4.3.3 Conservation and variation in flexibility/rigidity

As discussed in chapters 2 and 3, we have employed minimal Distance Constraint Model (mDCM) to establish the utility of comparative stability/flexibility relationship analyses. Along these lines, mDCM has been used to compare mesophilic and thermophilic RNase H pair [163], periplasmic binding homologs [164] and oxidized thioredoxin protein structures [165, 166]. In chapters 2 and 3, we have demonstrated that mDCM can predict stability changes (with 4.3% average error) [167] and characterize residue dynamical changes upon single site mutations in human lysozymes [168]. The results from above analyses support the same theme that backbone flexibility remains conserved across protein families, however, the pairwise mechanical couplings that give

insight of higher order descriptions of flexibility and rigidity are sensitive to small differences.

Only a small number of class A BL proteins have been studied by NMR and molecular dynamics simulation. As such, little is known about variation and conservation of dynamical properties across the BL protein family. In this part of the analyses, we have tried to quantify the consequences of evolution on BL protein dynamics. Figure 4.6a displays the multiple sequence alignment of twelve BL proteins color-coded by backbone flexibility index (FI). Residues colored blue are rigid, whereas the ones colored red are flexible. Figure 4.7 shows protein structures that follow similar color-coding described above. Figure 4.6b quantifies the average FI across the complete dataset displaying average FI curve with +/- 1 standard deviation. Positive FI values reflect the amount of excess degrees of freedom in flexible regions, and negative values reflect the amount of excess constraints in rigid regions. These results highlight two significant points. First, the BL enzymes have a rigid backbone, and second, this backbone rigidity is conserved across the whole family. Normally, our calculations do not predict structures to be so rigid, but this is consistent with NMR S2 order parameter descriptions [169]. The extent of rigidity is also visible at the N and C termini of BS3, TEM-1, SME-1 and SHV-2. The flexibility/rigidity results of BL proteins presented in Figure 4.6a are rank ordered based on increasing average rigidity characteristics. All BLs exhibit extended spectrum antibiotic hydrolyzing activity, except for PC1 and G; PC1 being the least rigid. Across the alignment, the secondary structure elements appear rigid, whereas intervening loops are flexible (except the Ω-loop). Three flexible regions have been identified as shown in Figure 4.6c: flexible region 1 at helix H3, flexible region 2 between H9 and H10 and

finally flexible region 3 at H11. While helix H10 is rigid, it is sandwiched between two flexible regions, meaning it could also have high mobility because the rigid body can "swing" from the flexible hinge in the same way a pendulum swings at a flexible pivot. We point this out because molecular dynamic studies have shown increased mobility in helix H10 upon substrate binding [170].

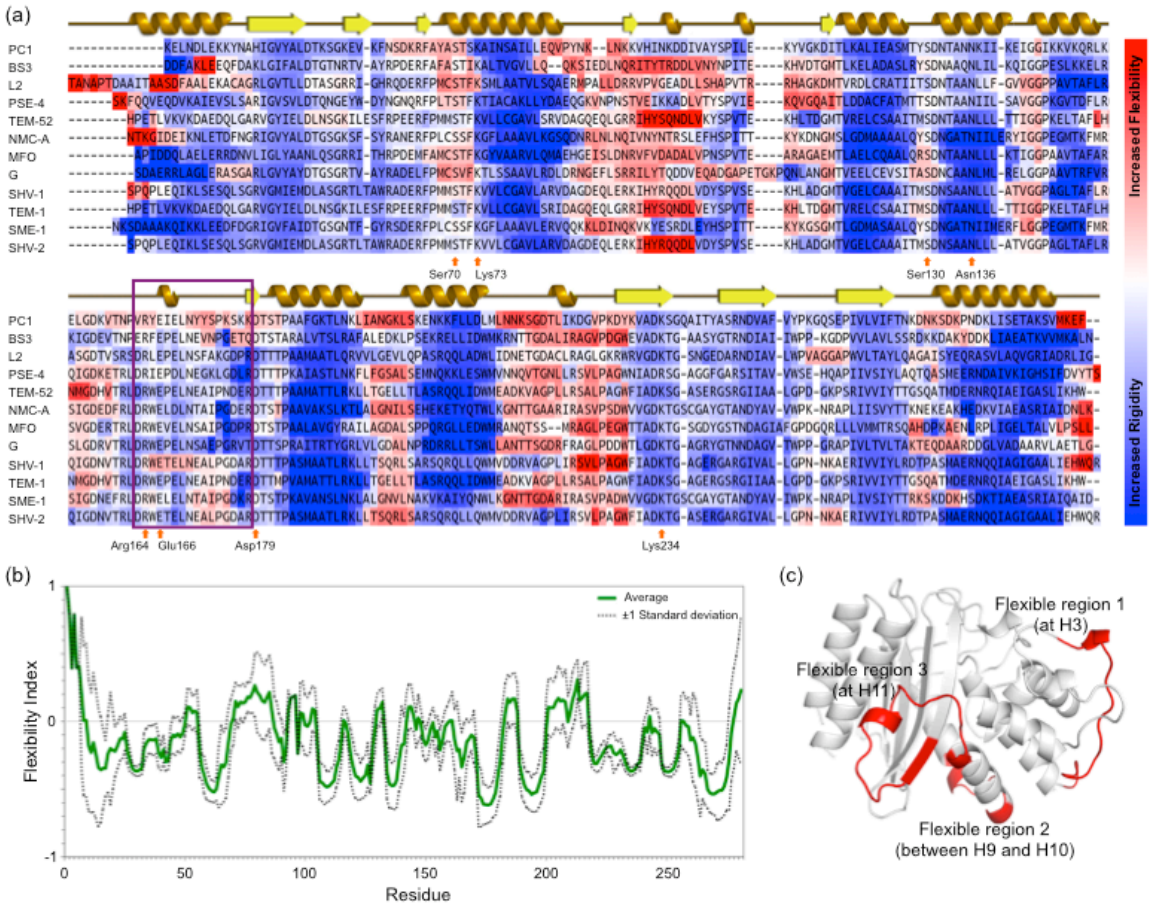

Figure 4.6: The backbone of β-lactamases is rigid and conserved. (a) The flexibility for each structure is mapped onto the multiple sequence alignment of the class-A β-lactamase family. The backbone of residues colored red is flexible, whereas blue indicates rigidity. The spectrum bar illustrates the extent of flexibility and rigidity, which ranges from -1 to +1. (b) Flexibility index values averaged across the family are shown in in green, whereas the dashed lines highlight fluctuations (as defined by ±1 standard deviation). (c) Visual observation of backbone flexibility identifies three main flexibility regions that are mapped on to the structure. These flexible loops might have an important role in protein functionality.

Mobility within the Ω-loop is thought to be important for substrate recognition and

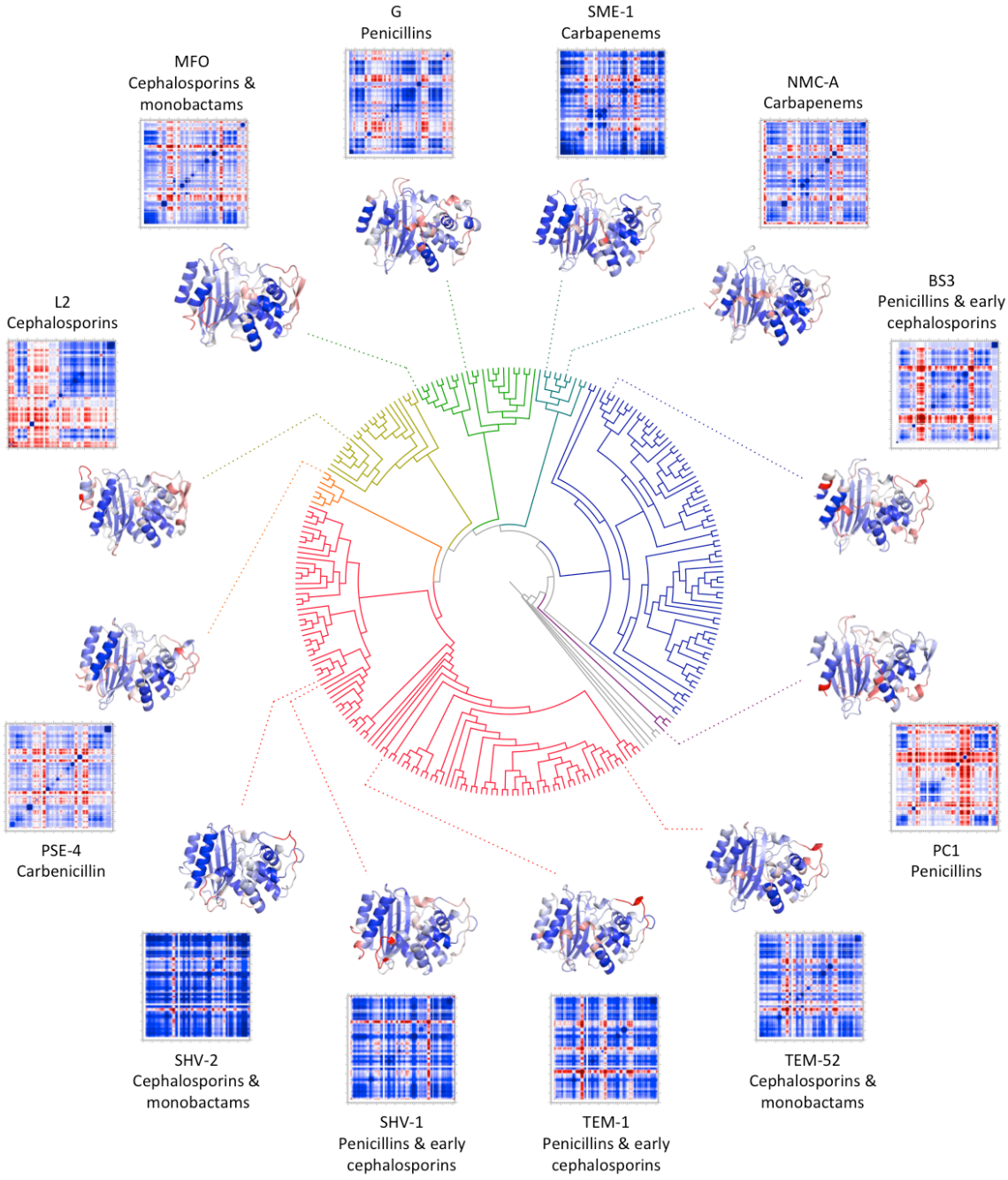

Figure 4.7: The phylogenetic tree along with the corresponding protein structures and cooperativity correlation plots. Sequence and structure dynamics are evolutionary related as evident from cooperativity correlation clustering. Structures are color coded by backbone flexibility index, which illustrates that all BL family members are primarily rigid with some punctuating flexible loops. Conversely, pairwise allosteric couplings are overall varied, yet typically conserved within evolutionary outgroups.

catalysis. Dynamic simulations performed in the past have suggested that the Ω-loop is rigid with order parameters comparable to other secondary structure elements [171]. The authors also illustrate the importance of flexibility at the tip of the Ω-loop, which is important for the opening and closing motion. Interestingly, mDCM results indicate that the Ω-loop is consistently isostatic, that is, marginally rigid (Figure 4.8). As discussed above, the Ω-loop also includes a key catalytic residue Glu-166 that performs the deacylation step. Furthermore, deletion of the Ω-loop makes the protein deacylation deficient resulting in the formation of stable acyl-enzyme complexes [172]. The marginally rigid Ω-loop suggests its catalytic importance where rigidity is important for reproducibility in substrate binding, yet also allowing for motion that might be functionally required. The Ω-loop region spans over three out of eight catalytic residues. Except Asn-136, all catalytic residues exhibit similar isostatic nature even though they occur throughout the BL sequence.

In stark contrast to the global variability observed across the BL dataset, the marginal rigidity and electrostatic properties of the active site region remain evolutionary conserved. In most cases, small increases in new activities can be directly attributed to only a handful of active site mutations that sterically allow new substrates to bind [137]; yet active site rigidity is maintained. In fact, this active site rigidity was recently utilized to develop new BL inhibitors using a fragment based drug design strategy [173]. These results support the view that steric and electrostatic complementarity between active site and different antibiotics are primarily responsible for BL resistance activities [174].

In addition to backbone flexibility, we also analyzed cooperativity correlation (CC) metric that describe pairwise mechanical couplings. As illustrated in Figure 4.7 CC

between a pair of residues in their native state can be rigidly correlated (colored blue), flexibly correlated (colored red), or uncorrelated (colored white). Taken together, the full CC plot can thus be considered a snapshot of all allosteric couplings within structure. In a previous investigation of periplasmic binding proteins, the variability within the cooperativity correlation was explained by differences within the H-bond network. Interestingly, the H-bond network of BL proteins remains conserved (discussed below), yet we observe substantial diversity and richness of CC throughout our dataset. In this way, the results presented here are much closer to the results with thioredoxin [166],

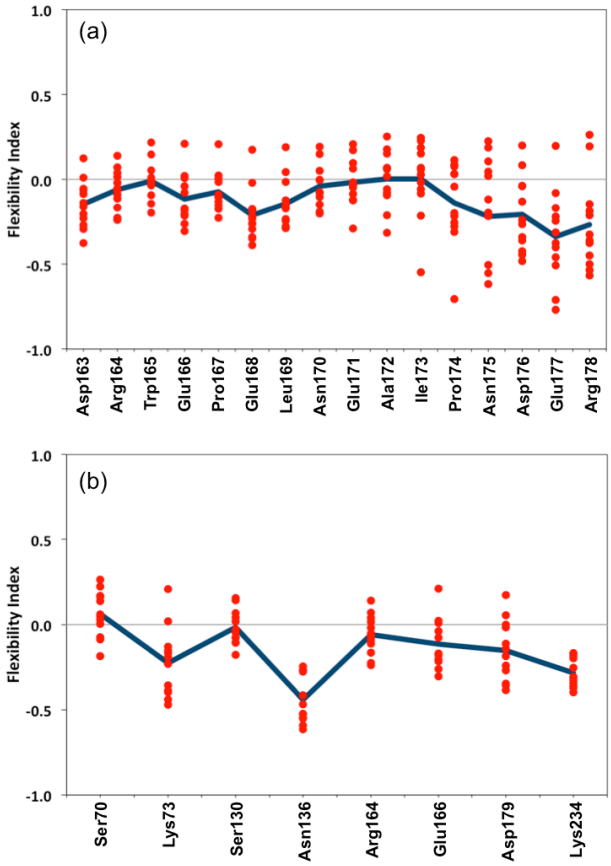

Figure 4.8: The backbone flexibility index reveals the nearly conserved isostatic nature of both (a) the Ω loop and (b) the eight active site residues. The red line indicates average values. Marginal rigidity is important for maintaining the active site structure, while also allowing for the flexibility needed for substrate recognition and catalysis.

CheY [175] and lysozyme [168] that stress the sensitivity of CC, and thus allostery, to subtle structural perturbations. To further investigate this susceptibility within BL, we again layer the physical descriptions of structure onto the BL phylogeny. As with the electrostatic potential maps, CC properties again cluster in a way that reflects local evolutionary outgroups (Figure 4.7). For example, TEM-1, TEM-52, SHV-1 and SHV-2 are largely composed of a single rigid cluster, which is consistent with earlier NMR [169] and MD [176] assessments of TEM-1 that indicated it is quite rigid. Carbapenemases SME-1 and NMC-A represent a close evolutionary pair, and thus have similar flexibility properties. Conversely, the L2 cephalosporinase, which belongs to a distinct outgroup, is atypically flexible.

An attempt was made to quantify electrostatic/rigidity relationships that could explain varying antibiotic resistance and selectivity, but was unsuccessful on repeated attempts. Nevertheless, these results clearly demonstrate how systematic differences within electrostatic properties and cooperativity correlation parallel the overall phylogeny across BL enzyme family. Further, it is interesting to note how nature preserves the active site dynamics and their electrostatics properties during evolution. Conservation of function provides the selection bias for proteins to maintain globally similar dynamics while evolving to varying substrate recognition patterns.

4.3.4 Conservation and variation in hydrogen bond network

Table 4.2 describes the global H-bond statistics showing the number of H-bonds and average total energy across the twelve BL structures. Since the mDCM is in large part based on H-bond networks, it is critical to understand how their variation can affect dynamical properties. H-bond statistics show that the number of H-bonds varies from

495 to 559, whereas the average H-bond energy ranges from -2.86 to -3.20 kcal/mol. From other studies we have noticed that the number of H-bonds can be trivially explained by the size of the protein [164]. However, due to their relative constant size, no such correlation is observed within this dataset. We also find that the above variations do not trivially predict differences within backbone FI and CC. That is, structures with more H-bonds are not necessarily more rigid than those with fewer. As we have discussed previously [166, 168], this observation again stresses that topological considerations get lost in global metrics due to nonadditive nature of the mDCM, which has a considerable effect on the output.

We employ a simple but effective approach for comparing H-bond networks by plotting H-bond density plot and H-bond contact maps to visualize essential differences (Figure 4.9). For better visualization, we have subtracted the mean H-bond from the H-bond density at per residue level (Figure 4.9a). This strategy helps accentuate important H-bonds that might be critical for stability and function. There is a rich density of H-bonds at strand β1, the Ω-loop and β9, which are conserved throughout the family. An overlapped H-bond contact map of all the twelve BL structures gives us an insight of regions with strong H-bond interaction (Figure 4.9b and 4.9c). The site labeled 1 shows strong interactions between three regions that extend over the all the key catalytic sites. Similarly, experimental studies [172] have highlighted the importance of strong interactions between (*i.*) Lys-73 and Glu-166, (*ii.*) Arg-164 and Asp-179, and (*iii.*) Asn-136 and Glu-166. The authors emphasize that removing any of these interactions can make the enzyme catalytically inefficient, while also disturbing its stability. Site 2 on the contact map displays the presence of strong interaction at the Ω-loop region, which are

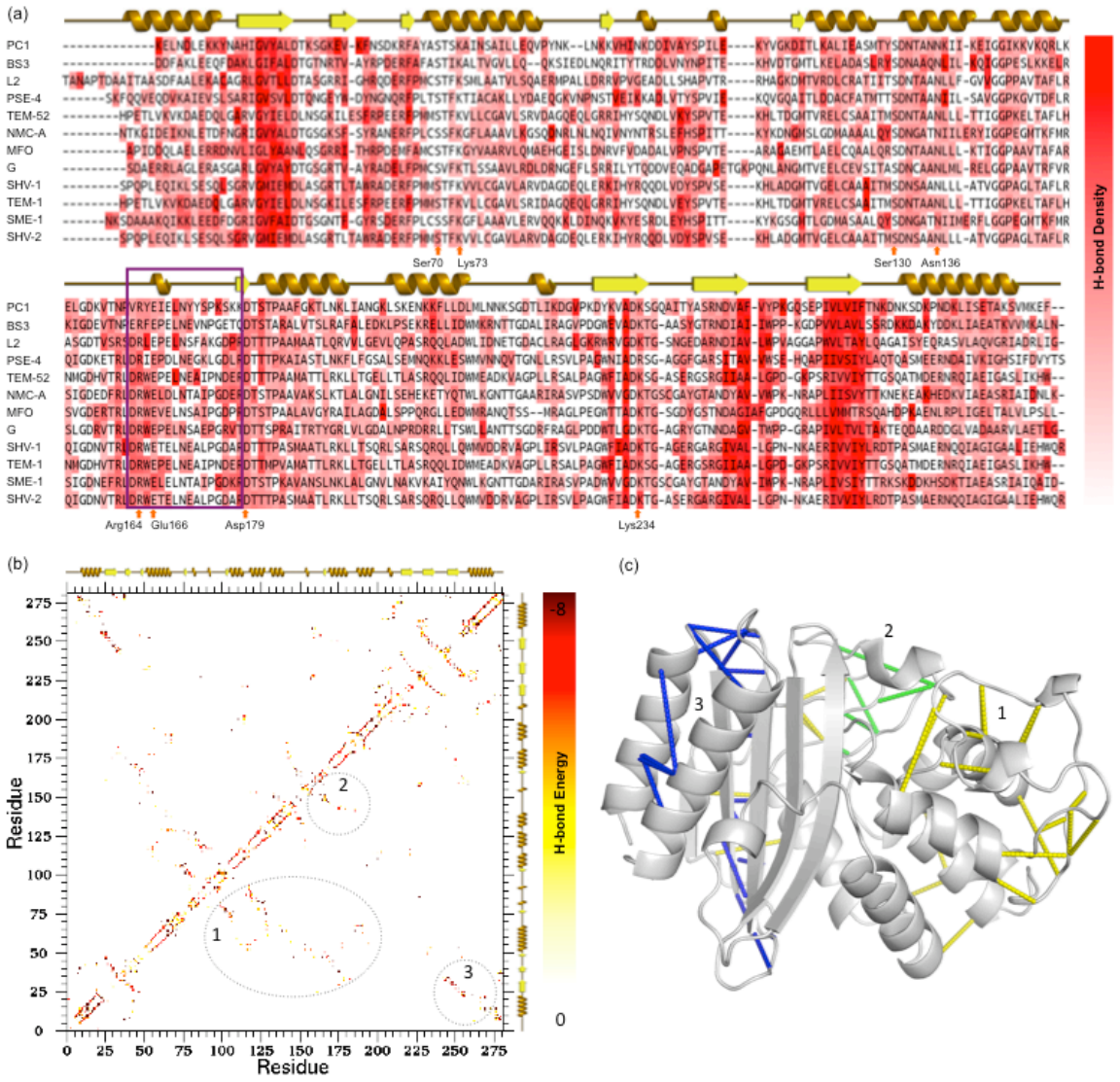

Figure 4.9: Conservation in H-bond networks. (a) H-bond density is plotted per residue, which identifies regions rich in H-bond interactions. The Ω-loop is shown inside purple box and active site locations are marked as well. (b) Overlapped H-bond contact maps reveal three important sites important for maintaining the active site structure integrity and substrate catalysis. (c) The CA-CA atoms of residues at the corresponding sites (1,2 and 3) are depicted by yellow, green and blue lines respectively. For better visualization only strong H-bond connections have been displayed on the structure

assumed to be important for maintaining functionality. Site 3 illustrates the presence of strong interactions between strands β1 and β9, which are assumed to be related to structural stability. Furthermore, strong H-bond interactions are observed at secondary

structures as expected. The conservation of H-bonds within these secondary structures is what leads to rigidity being conserved along the backbone. However, H-bond conservation within the contact map analysis does not mean that their energies are equivalent and this could lead to variable flexibility within cooperativity correlation. Conclusively, the qualitative conservation within secondary structure H-bonds lead to conservation of backbone rigidity across our dataset of twelve BL proteins.

4.2 Methods

4.2.1 Dataset preparation

In this study, twelve different class A BL structures are investigated to provide a large evolutionary cross-section for detailed analysis [178-189], while maintaining a feasible number for data and visual assessment. All structures have been solved to high resolution and R-values are lesser than or equal to 0.22. As provided in Table 4.3, three out of twelve structures exhibit penicillinase activity while the rest belong to one of the

Table 4.3: Structural and catalytic characterization of the dataset. Functional class is defined by Bush et al. [144, 177].

Organism	Enzyme Name	PDBID	Res (Å)	R-value	Function Class	Extended spectrum?	Substrate
S. albus	G	1BSG	1.9	0.15	2a	No	Penicillins
E. coli	TEM-1	1BTL	1.8	0.16	2b	Yes	Penicillins & early cephalosporins
E. cloacae	NMC-A	1BUE	1.6	0.19	2f	Yes	Carbapenems
S. marcescens	SME-1	1DY6	2.1	0.18	2f	Yes	Carbapenems
P. aeruginosa	PSE-4	1G68	2.0	0.17	2c	No	Carbenicillin
K. pneumonia	TEM-52	1HTZ	2.4	0.22	2be	Yes	Extended-spectrum cephalosporins & monobactams
S. maltophilia	L2	1N4O	1.9	0.16	2e	Yes	Extended-spectrum cephalosporins
K. pneumoniae	SHV-2	1N9B	0.9	0.13	2be	Yes	Extended-spectrum cephalosporins & monobactams
K. pneumoniae	SHV-1	1SHV	2.0	0.18	2b	Yes	Penicillins & early cephalosporins
M. foruitum	MFO	2CC1	2.1	0.17	2be	Yes	Extended-spectrum cephalosporins & monobactams
S. aureus	PC1	3BLM	2.0	0.16	2a	No	Penicillins
B. licheniformis	BS3	4BLM	2.0	0.16	2b	Yes	Penicillins & early cephalosporins

following classes: broad-spectrum, extended-spectrum, carbapenamase, cephalosporinase or carbenicillinase. Moreover, all enzymes are inhibited by clavulanic acid and their structures are remarkably similar; the pairwise α-carbon root mean square deviation (RMSD) ranges from 0.73 to 2.57 Å (Figure 4.10).

4.2.2 Model parameterization

As previously discussed, the mDCM is parameterized by finding values of (u_{sol}, v_{nat}, d_{nat}) that best reproduces the experimental C_p data using simulated annealing method (Figure 4.11). We parameterize the model using the C_p curve from *B. cereus* [190] and the evolutionarily closest structure BS3. Focusing on our group of twelve class A BL proteins with well-conserved structures of the same function, we have transferred the three adjustable parameters obtained from above to all the other members. With this fixed parameterization, we have confirmed that mDCM correctly predicts all BL orthologs to have a single peak in C_p and a two state folding/unfolding transition in free energy. Apart from these twelve BLs, an attempt was made to calculate QSFR quantities of three other proteins, but this was unsuccessful and hence not included in the analysis.

We have consistently demonstrated that while thermodynamic quantities (i.e., T_m) are somewhat sensitive to parameterization and input structure resolution, the mechanical FI and CC quantities are mostly robust to parameter differences. Nevertheless, a single parameter set across the dataset, guarantees that QSFR differences only arise from structural differences. Also, results from [163, 165-167] have demonstrated that QSFR properties are insensitive to parameterization, and have minimal influence on CC and FI values. As such, the conclusions regarding changes in QSFR properties are robust.

4.2.3 Phylogeny

For expanding BL sequential coverage, we collect approximately 1100 sequences after searching through the nonredundant protein database using BLASTP [191]. The protein sequence culling algorithm PISCES [192] is employed to filter sequences at 98% mutual sequence identity cutoff. This reduced dataset, which also includes twelve class A BL protein sequences, is further aligned by MUSCLE [193] followed by phylogenetic

	G	TEM-1	NMC-A	SME-1	PSE-4	TEM-52	L2	SHV-2	SHV-1	MFO	PC1	BS3
G		37.3	40.0	39.0	30.8	37.3	40.1	37.4	37.7	41.2	31.1	42.2
TEM-1	2.29		32.1	33.0	42.1	97.3	37.5	64.9	66.4	37.0	30.4	36.7
NMC-A	1.63	2.10		73.0	33.5	33.1	36.0	33.6	34.0	37.8	33.1	41.0
SME-1	1.62	2.15	0.82		35.7	34.2	38.2	36.6	37.0	35.5	33.9	42.6
PSE-4	1.99	1.71	1.67	1.79		42.6	34.9	43.8	44.5	32.1	34.6	29.7
TEM-52	2.32	0.73	2.17	2.28	1.73		37.9	67.2	66.4	37.4	31.5	37.1
L2	1.86	1.77	2.02	1.99	1.73	1.80		37.4	37.7	40.1	29.2	35.2
SHV-2	2.51	1.52	2.18	2.29	1.80	1.57	2.26		98.1	35.1	27.6	33.6
SHV-1	2.50	1.51	2.14	2.26	1.67	1.57	2.20	0.85		35.9	27.6	34.0
MFO	1.72	1.99	1.50	1.57	1.57	2.01	1.85	2.21	2.14		31.5	39.1
PC1	2.20	2.38	2.15	2.27	2.16	2.42	2.14	2.57	2.42	2.18		42.6
BS3	1.48	1.98	1.55	1.64	1.74	2.00	1.54	2.23	2.16	1.58	1.57	

RMSD (Å) — Sequence identity (%)

Figure 4.10: Dataset similarity. All-to-all percent sequence identity (blue) and structural RMSD (red, in units of Å) are provided to highlight (dis)similarity..

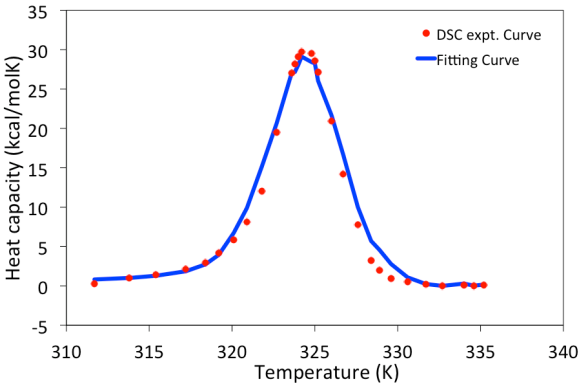

Figure 4.11: The best-fit heat capacity curve by mDCM is shown with u_{sol} = -2.61, v_{nat} = -0.32 and d_{nat} = 1.61, which are within normal ranges established by our previous studies. These three fitting parameters are required to calculate free energy of the protein accurately using Eq. 1.

tree construction using maximum-likelihood, meaning the phylogenetic tree shown in Figures 4.4 and 4.7 is purely derived from sequence information. The twelve BL protein sequences span across the evolutionary tree, which provide a robust structural coverage as well. However, we arrange these twelve BL sequences independent of the larger set, using both sequence and structural information by Protein Align tool in MOE [194], to achieve better visual comparison across our set.

4.2.4 Continuum electrostatic calculation

Additions of hydrogen atoms, residue pK_a calculations and intramolecular electrostatic interactions have been performed on energy minimized protein structures using H++ web server [195]. Hydrogen atoms were added and their positions optimized (MD based) after calculating ionization states of the titratable residues using Poisson-Boltzmann continuum electrostatics. The server uses MEAD suite of programs, and detailed information of the algorithm can be found here [195]. The salinity and pH conditions are kept consistent with the conditions used in the original DSC experiment, i.e., 0.06M salt concentrations and pH 7.0; and a solvent dielectric constant of 80 and an interior protein dielectric of 6. Residue acidity and basicity changes (Figure 4.4a) are calculated with respect to model pK_a values from [196]. The .pqr file generated from H++ containing charge and radii information is fed into APBS [197] to generate electrostatic potential data. The protein is centered on a 65 x 97 x 65 grid. The electrostatic potential maps in Figures 4.4 and 4.5 are displayed at +/- 1 kcal/mol.

4.2.5 Hydrogen bond network

H-bond density for a residue *i* is defined as:

$$Hbond_i^{Density} = \frac{\sum_{j=1}^{\#\ of\ residues} Hbond_{i,j}^{Energy}}{\sum_{j=1}^{\#\ of\ residues} Hbond_{i,j}^{Count}}$$

where, $Hbond_{i,j}^{Energy}$ is the hydrogen bond energy between residue *i* and *j*; and $Hbond_{i,j}^{Count}$ is the number of hydrogen bonds formed between residue *i* and *j*. The summation of energies divided by total number of hydrogen bonds provides hydrogen bond density at per residue level (Figure 4.9a). The hydrogen bond network contact map, shown in Figure 4.9b, is an overlapped network of all twelve BL proteins. The residue positions on the network follow multiple sequence alignment as described above. As such, identical donor and acceptor residue pair positions across the dataset are achieved for easy visual network assessment of hydrogen bond energies.

4.4 Conclusion

The BL enzyme family represents an interesting case study in protein family evolution. While conservation of function is the primary driving force in the evolution of most protein families, rampant antibiotic overuse has introduced new pressures leading to new resistance activities. The bulk of these new activities have been trivially explained by steric changes within the BL active site [198]; however, we wanted to determine if there were additional physical variations that underlie the various resistance activities. Our results indicate that conserved active site electrostatic networks and a fairly rigid backbone structure are critical to BL function (Table 4.4). Going further, we were unable to associate conserved differences with resistance activities. Rather, variations were conserved within evolutionary outgroups emphasize sequence/structure relationships. For example, PC1 and G are both characterized as penicillinases, but their global electrostatic and flexibility properties are quite different due to divergence. As such, it follows that this results underscores that multiple structure/function mechanisms can converge on the same function. Conversely, MFO and G are from the same outgroup and thus have

Table 4.4: Summary of conserved of variable physical properties across the class-A β-lactamase family

	Conserved Properties	Variable properties
Electrostatics	Per residue pK_a values	Electrostatic potential maps (yet conserved within outgroups)
	Active site electrostatic network	
	Anionic patch on electrostatic potential maps	
QSFR	Backbone flexibility properties	Cooperativity correlation (yet conserved within outgroups)
	Ω-loop is isostatic	
H-bond network	Conserved at per residue level	

similar physical properties, yet they have vastly different activities (MFO has extended spectrum activities and can be classified as a cephalosporinase that can also hydrolyze monobactams). In summary, our results uncover common physical origins that underlie the conservation of function and several physical properties that expand our understanding of the molecular basis of evolution with the class-A BL enzyme family. Furthermore, the inability of these physical differences to explain antibiotic resistance activities gives further support to the hypothesis that new activities are trivially described by active site shape and sterics.

It is seen that the above familial analysis is limited to just twelve BL protein structures that effectively explain evolutionary relationships. Filling structural gaps using sequence information from phylogenetic tree could bring out other important descriptions across protein families and superfamilies. This brings us to the following question: *Is it possible to calculate QSFR properties of proteins that have no x-ray structures and just sequences?* Chapter 5 elaborates this development further.

CHAPTER 5: TOWARDS COMPREHENSIVE ANALYSIS OF PROTEIN FAMILY QUANTITATIVE STABILITY/FLEXIBILITY RELATIONSHIPS

5.1 Introduction

For a complete comparative biophysical characterization across protein families and superfamilies, the foremost limitation is the ability to get good x-ray structures. In the past we have performed comparative QSFR analyses for bacterial periplasmic binding homologs [199], oxidized thioredoxin [200] and β-lactamase protein families and the maximum number of proteins assessed for a familial analysis was 12. A wider breadth analysis should include hundreds of proteins for a complete comparative QSFR analysis. Unfortunately, there are only 78 families out of 3900 SCOP families that have 25 or more distinct orthologs with experimentally solved structures. A large-scale QSFR analysis methods on the scale of dozens to 100+ structures would require efforts to fill-in these structure gaps using homology modeling. From our previous analysis it is known that mDCM can detect subtle variations in QSFR properties even due to single point mutations [201] [202]. Hence, the key to reproduce accurate QSFR predictions will solely depend on good homology models. In this study we have benchmarked QSFR properties of 65 human lysozyme homology models against 7 different x-ray structures of human c-type lysozymes. The idea here is to select good homology models that can reproduce QSFR properties of x-ray structures accurately. If successful, this study would be a significant advancement in building selection criteria for choosing better models for precise QSFR predictions. Once proved, this methodology would help us increase the

structural coverage for our comprehensive breadth analysis across protein families and superfamilies.

5.2 Methods

5.2.1 Preparation of homology models

In this benchmark study, 65 human lysozyme homology models are constructed from 13 different templates, i.e., 5 models from each template using MODELLER [203] using default settings. To ensure proper ionization, the H++ server [204] is used to add hydrogen atoms to the structures as expected at pH 2.7 based on calculated pK_a values. Other structural details are provided in Table 5.1.

These 13 templates, from different species, have a wide range of sequence identity with human lysozyme varying from 37.6% to 77.7%. To benchmark each homology model, we also construct QSFR profiles from 7 existing human lysozyme x-ray structures. For each QSFR profile we establish a background profile range using ±1 standard deviation from x-ray structures' average QSFR profiles at each reside position. Details of

Table 5.1: Orthologs used for constructing human lysozyme protein structure. 5 models are built from each template.

Organism	Template PDB ID	Resolution (Å)	R-value
Turkey	135L	1.30	0.189
Northern bobwhite	1DKJ	2.00	0.177
Domestic silkworm	1GD6	2.50	0.181
Chicken	1HEL	1.70	0.152
Helmeted guineafowl	1HHL	1.90	0.170
Tasar silkworm	1IIZ	2.40	0.231
House mouse	1IVM	-	-
Ring-necked pheasant	1JHL	2.40	0.214
Echidna	1JUG	1.90	0.170
Rainbow trout	1LMN	1.80	0.174
Dog	1QQY	1.85	0.178
Horse	2EQL	2.50	0.234
Japanese quail	2IHL	1.40	0.165

different QSFR metrics used in this analysis are assessed further.

5.2.2 Model parameterization

The mDCM is parameterized by finding the best set of $\{u_{sol}, v_{nat}, d_{nat}\}$ that best reproduces the heat capacity curve using simulated annealing [205, 206]. Each model and x-ray structure has been fit to the same C_p curve obtained from the DSC experiment performed by Takano *et. al*.[207]. Best-fit curves for models from *rainbow trout* are displayed in Figure 5.1. Other model structures exhibit similar heat capacity fit trends although there are differences in parameters. Interestingly, the least squares fitting error is not correlated to homology model accuracy, highlighting the importance of other structural features that contribute towards prediction of accurate thermodynamic and mechanical features. Initially, this study included 80 homology models. However, 15 homology models were not included in the analysis due to unsuccessful model parameterization. Discussions regarding changes in thermodynamical and mechanical quantities arising from parameter differences are discussed later.

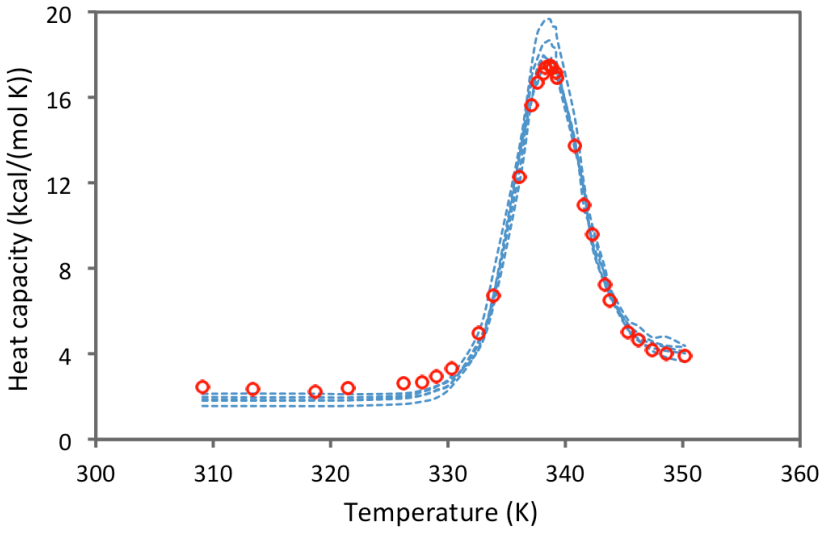

Figure 5.1: Five homology models constructed from *rainbow trout* template are fit to human lysozyme heat capacity curve. Curve represented by red-dots is obtained using DSC experiments, whereas the five dashed-blue curves represent fits.

5.3 Good homology models

The definition of a good homology model is somewhat arbitrary and difficult to describe. Moreover, there are many scoring functions that can assess the quality of homology models based on statistical potentials and physics-based energy calculations [208, 209]. The errors resulting from poor quality models are the most problematic aspect of this work. For any comprehensive QSFR analysis in the future, we would require models that are of satisfactory quality, which can accurately predict the QSFR quantities for proteins with unknown x-ray structures. In this benchmark study, along with 65 lysozyme model structures, our dataset also includes 7 original x-ray human wild-type lysozyme structures. As discussed in [202], we ask ourselves a similar question: *What constitutes QSFR metric deviation from background profile and what does not?* To address this point for benchmarking QSFR prediction, we have used the same methodology for establishing the QSFR metric profile as defined in chapter 3. Any model QSFR metric within ±1 standard deviation (i.e., ±1σ) of x-ray structures' QSFR baseline is considered to be a "good prediction", at a given residue position. A prediction value falling beyond ±1σ defines "poor prediction" for that QSFR metric.

Figure 5.2 compares percentage of residues, for each model, within ±1σ of backbone flexibility index profile of x-ray structures with various sequence and structure similarities between models and x-ray structures. These comparison studies clearly suggest that a close agreement between models and x-ray structures results in a better flexibility index prediction. On the other hand, a careful observation highlights the presence of many models that are false-positives and false-negatives. Considering a cut-off of percent residues for a good prediction at 60%, there are 8 models that result in a

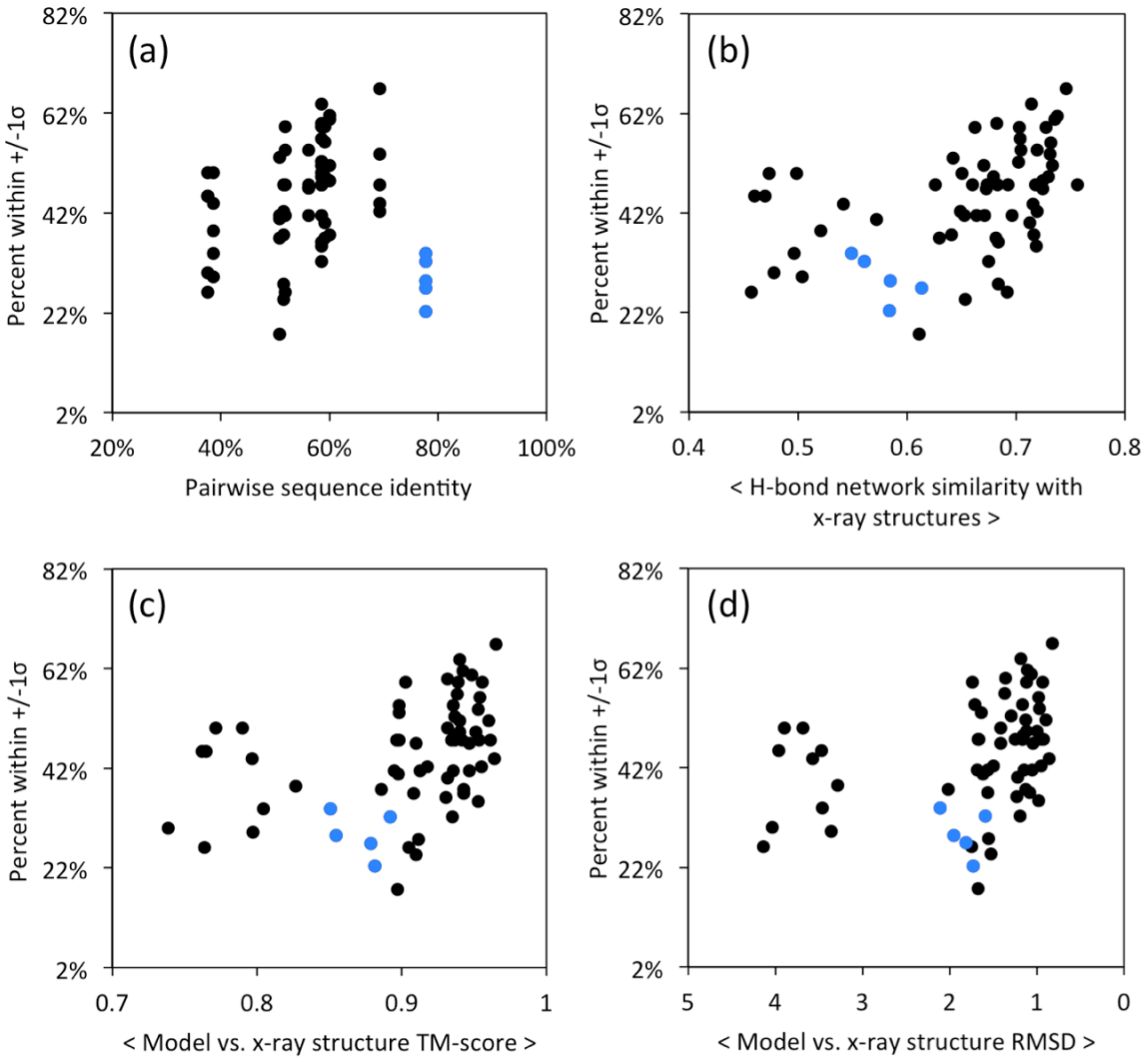

Figure 5.2: Comparison of flexibility index accuracy prediction for all 65 homology models against (a) pairwise sequence identity, (b) hydrogen bond network similarity, (c) TM-scores and (d) structure RMSD with original x-ray crystal structures averaged across all 7. Models in close agreement with x-ray structures reproduce flexibility index with higher accuracy. However, many false-negatives and false-positives are present as well. The blue data points represent NMR structures. 82% is the best-expected score possible comparing original x-ray structure with average.

true prediction. However, comparing hydrogen-bond similarity, there are as many as twice this number that lie within the false-positive region, hence posing a challenge in development of a threshold for model selection. These varying results are not unexpected since subtle changes in the designed model structures can cause drastic differences in

hydrogen bond interactions.

Structure comparisons using RMSD show similar trends. Moreover, structure deviations in the loop regions can undermine better QSFR predictions (Figure 5.2c and 5.2d). Hence, another structure comparison method called the TM-score was used which is less sensitive in the loop regions [210]. However, these scores highlight similar trends as well.

A good model may not necessarily represent the true physiochemical property of the original structure resulting in many false-negative model structures. Comparison of pairwise sequence identity (between models and structures) highlights the importance of "twilight zone" in homology modeling as well. Results show that models designed from templates having high sequence identity result in better FI prediction. The best human model prediction arises from *rainbow trout* (1LMN) resulting in highest QSFR prediction accuracy of 67%. However, the overall accuracy of 5 models from the same template has a wide range of prediction accuracy reaching as low as 42%. Another surprising observation is a poor FI prediction by models designed from *house mouse* (1IVM), the only NMR structure template in our dataset (shown by blue dots in Figures 5.2 and 5.3). The sequence identity of this template is 78% whereas the average prediction accuracy is approximately 30%.

The above preliminary comparisons of models with original x-ray structures suggest that a comprehensive QSFR analysis is definitely possible, provided we filter true-positive homology models to improve prediction correctness by boosting statistics. The upper boundary of the plots (corresponding to 82%) in Figure 5.2 defines the best result obtained by comparing each of the original x-ray structures to the average x-ray

structures' backbone profile.

Also, the results described in Figure 5.2 are benchmarked against just a single QSFR metric, i.e. flexibility index, undermining other QSFR metrics that can also help in enhancing prediction statistics, if considered. Furthermore, comparison of H-bond network and calculation of TM-scores and RMSDs require original x-ray structures. Except sequence identity, none of these comparison scores would be available for establishing thresholds due to lack of original x-ray crystal structure(s). To overcome this problem, we can use model quality assessment score called QMEAN [211, 212], which is one of the best methods for model quality assessment [213]. This scoring quantity considers secondary structure interaction potential, degree of solvent exposure and other structural quantities for structure assessment.

Figure 5.3b compares QMEAN scores with FI correlation, which provides a different

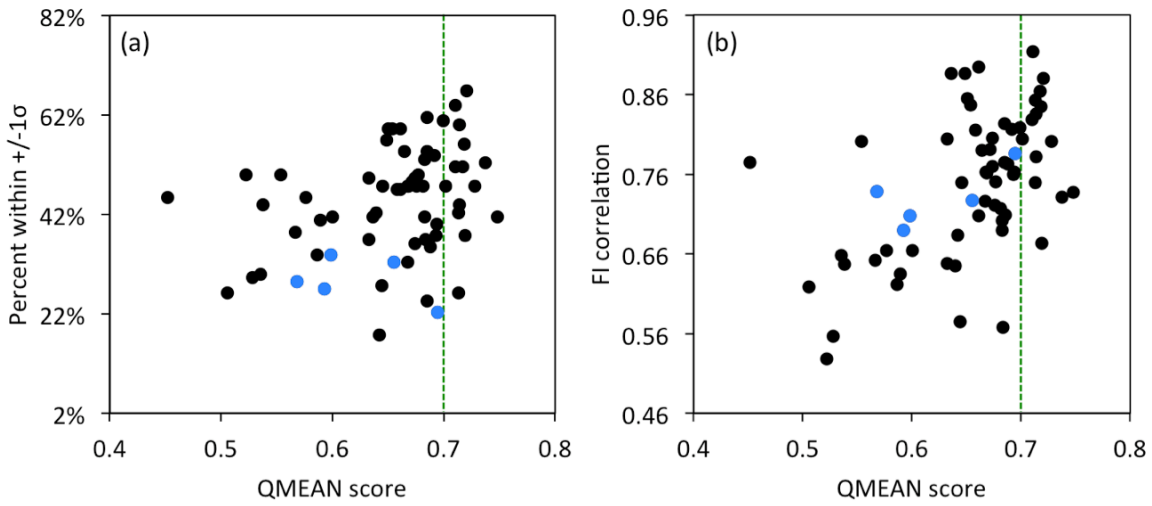

Figure 5.3: Comparison of (a) FI percent accuracy prediction and (b) FI correlation of 65 homology models against original x-ray crystal structures averaged across all 7 with their QMEAN scores. Models with high quality may not necessarily result in higher FI accuracy. Blue data points represent NMR structures. 82% within ±1σ and a FI correlation of 0.96 are the best expected scores possible comparing original x-ray structure with 7 original x-ray structures' average FI metric. The green dashed line represents an arbitrary QMEAN threshold value at 0.7.

benchmarking quantity replacing percent within ±1σ. These two datasets are qualitatively the same, the main difference being a relative upward shift in the areas of interest. More specifically, while figure 5.3a shows results between ~ 20% and 80%, the F1 correlation values (figure 5.3b) range from ~ 0.55 to 0.95. Additionally, the Spearman Rank correlation between these two quantities is 0.61 (p-value = 8.32e-8), suggesting some degree of association. This suggests that any of these methods can be convincingly used for benchmarking, however, in this study we have chosen to benchmark using ±1σ approach. The green vertical lines in figure 5.3 provide an arbitrary demarcation of model structures using QMEAN scores that should have similar backbone FI properties with the original structures. Surprisingly, this is not the case. As evident from the plot, the best homology model structure is not able to reproduce accurate QSFR calculations. Furthermore, initial analyses indicate that the highest possible FI accurate prediction of 60% does not come from the best QMEAN model. Based on our data observation we focus on the following question: *Is it possible to improve accuracy number by enriching our dataset with additional good homology models? If yes, then to what extent this accuracy is achievable sans over prediction?* To answer these questions, we have employed two different strategies for selecting good homology models. In the first case, our selection criteria is simply based upon QMEAN scores, where we select 5 top scoring models to calculate average QSFR properties. QMEAN threshold is not recommended here, as these scores are protein size dependent. In the second case, we employ clustering technique for filtering models that share common QSFR properties. The key assumption of this strategy is that models with similar properties would tend to cluster together. Along with QSFR quantities, structure quality, sequence similarity and other

thermodynamic information is also employed to build a better prediction model. We also select the best QMEAN scoring model to benchmark against original x-ray structures' average QSFR metrics.

5.4 Results and discussion

5.4.1 Expectation Maximization clustering

Expectation Maximization (EM) is a statistical algorithm based on iterative method for finding the maximum likelihood estimates of parameters of unobserved latent variable dependent statistical models. Frequently used in clustering, the EM method assigns a probability distribution to every data instance, which defines the probability of it belonging to each of the clusters. The algorithm can create its own clusters and does not require *a priori* information regarding the expected number of clusters. To find the optimum number of clusters the EM algorithm cross validates and calculates the average log-likelihood. Starting with one cluster, the numbers of clusters are increased if the average log-likelihood continuously increases at each step.

To employ EM in our clustering analysis, we collect sequence, structure, and thermodynamic and mechanical quantities for each of the 65 human model lysozymes. The initial step includes calculating percent sequence identity between template and human sequence, QMEAN structure quality score, free energy barrier height of folding and unfolding, and flexibility order parameters at native, transition and unfolded states of the models for EM clustering. The mean and standard deviation of the resulting three clusters is summarized in Table 5.2.

Cluster-2 models, with best average QMEAN score and least standard deviation (0.68±0.03), are selected for our next round of clustering. Clustering provides an indirect

Table 5.2: Clustering results from structure and thermodynamic quantities. The values represent average ± standard deviation of data points for the given quantity belonging to a defined cluster. Cluster-2 with best average QMEAN score with least standard deviation is selected.

	Cluster-1	**Cluster-2**	**Cluster-3**
Folding free energy	2.29±0.84	3.57±1.23	2.79±1.05
Unfolding free energy	1.84±0.73	3.12±1.09	2.33±0.96
θ_{nat}	0.75±0.11	0.92±0.14	0.76±0.13
θ_{trans}	1.06±0.15	1.31±0.20	1.10±0.21
θ_{dis}	1.70±0.22	2.07±0.16	1.70±0.30
QMEAN scores	0.67±0.04	0.68±0.03	0.54±0.04
Sequence identity	0.60±0.09	0.58±0.05	0.39±0.01

threshold boundary for model structures' quality assessment scores. It should be noticed that it is difficult to select a cluster based on thermodynamic information because in an actual model assortment experiment we will have no *a priori* thermodynamic information of the missing x-ray structure to benchmark upon. Figure 5.4 plots QMEAN scores versus θ_{nat}, where the points are color coded by clusters. Selecting cluster-2 outcomes in 23 models that have closely related thermodynamic properties with less standard deviation on all quantities (Table 5.2).

Our next step is to calculate an all-to-all QSFR metric correlations and all-to-all structure deviations for each model pair. 23 selected models result in 253 possible pairs, and considering each pair, sixteen different QSFR metric correlations are calculated. This data will be fed for another round of clustering where pairs with similar QSFR properties are expected to cluster together. Since QSFR quantities are structure dependent, we also feed-in model structures' RMSDs and TM-scores [210] for EM clustering. Figure 5.5 plots TM-score versus probability to rotate. Cluster-9, which has the highest correlated TM-score pairs and least paired RMSD is chosen. Finally, the paired list in the selected

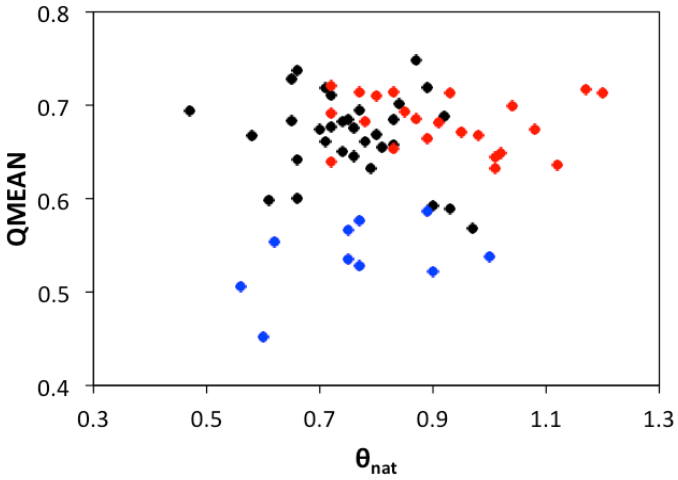

Figure 5.4: Clustering using structural and thermodynamic quantities. Clusters are represented by different colors. Models clustered in red (cluster-2) are selected for further analyses.

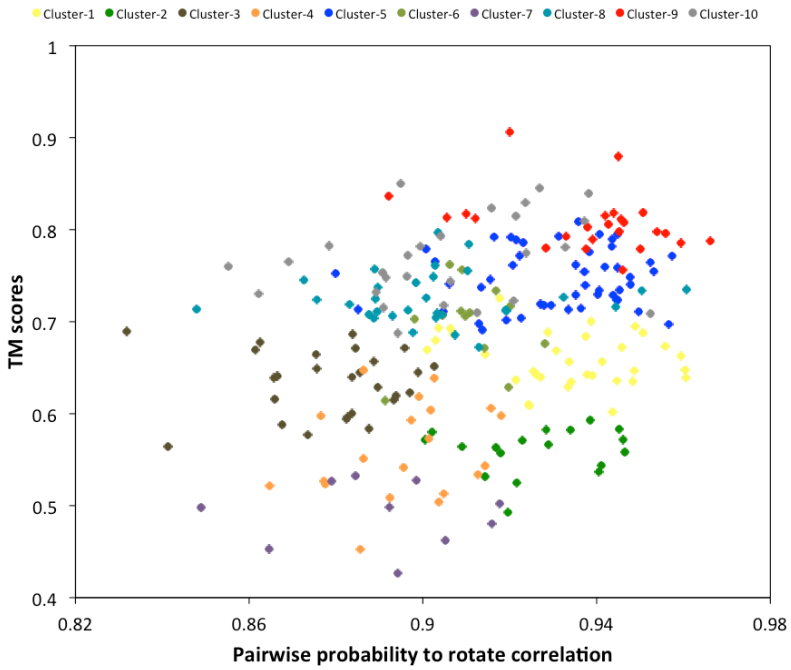

Figure 5.5: Results obtained from second round of EM clustering using structural and mechanical quantities. Clusters are represented by different colors. Model pairs clustered in red (cluster-9) are the final filtered models.

cluster is collapsed into a single list of 18 models. These 18 models constitute our third set for calculating average QSFR properties to benchmark against x-ray structures' QSFR

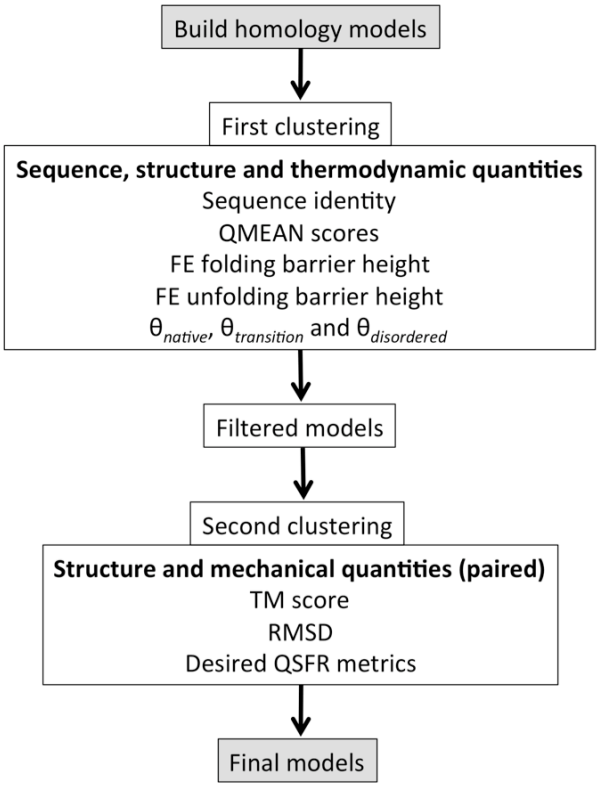

Figure 5.6: Expectation maximization clustering workflow to filter best homology models.

properties. Figure 5.6 reviews clustering workflow described above.

5.4.2 Improvement in QSFR metric prediction

To summarize up to this point; the first set consists of a single best QMEAN model, whereas the second set includes 5 best scoring QMEAN models. The third set is constructed from 18 models using EM clustering techniques. We now compare these 3 model sets and assess them by comparing against our original x-ray structures' backbone QSFR profiles. Any average quantity, per residue, resulting from a model set embraced within $\pm 1\sigma$ of the x-ray average is counted as a good prediction.

Before benchmarking the results, here is a brief description of the additional protein backbone QSFR quantities used in clustering and model assessment that are being

introduced for the first time. Susceptibility correlation describes the susceptibility for a particular residue to oscillate between a rigid and a flexible region. Regions that have higher susceptibility correlation values tend to have functional importance highlighting those residues that maintain a native/rigid structure and provide enough sloppiness to carry protein function. QSFR metrics degree of cooperative flexibility and degree of cooperative stress define local flexibility and rigidity per torsion taking into account the total number of correlated independent degrees of freedom and redundant constraints, respectively. Density of independent degrees of freedom and its probabilities signify regions that can exhibit flexible motion, whereas residues with density of redundant constraints or probability of stress exhibit rigid nature with less motion. Another QSFR metric probability to rotate quantifies local flexibility similar to flexibility index. Rotation of dihedral angles can be quantified with this metric.

Figure 5.7 gives the statistics of good prediction for three different model sets across 11

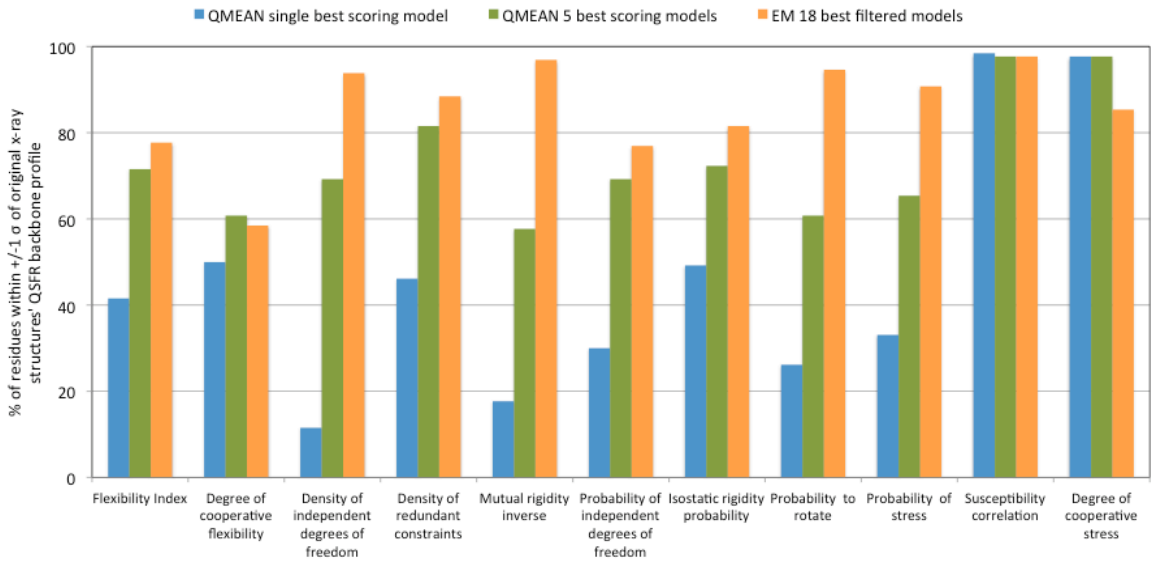

Figure 5.7: Comparison of protein backbone QSFR metric accuracy levels for different model sets. Average QSFR quantities obtained from EM 18 best filtered models have higher agreement with x-ray structures' QSFR quantities as compared to QMEAN best and 5-best models.

different backbone QSFR metrics. The single-best model set fails to give an accurate prediction for almost all the metrics. However, selecting 5-best QMEAN model improves the prediction accuracy.

Interestingly, QSFR metrics susceptibility correlation and degree of cooperative stress have high prediction accuracy even from a single best model. However, prediction of density of independent degrees of freedom resulting from single model set has the lowest accuracy of 11.5%. Lowest prediction value from 5-best QMEAN model set is 57.7% for mutual rigidity inverse. Results from other QSFR metrics also suggest that single best model and 5-best models fail to provide desired accuracy levels of at least 82% in FI. Prediction from EM 18 best filtered models significantly improves the accuracy results. Eight out of eleven QSFR metrics result in 80% or higher accuracy. These results indicate that clustering can enrich the homology model set with good models for accurate QSFR metric predictions. Average calculation of FI metric results in 78% accuracy, which is very close to our desired 82% precision (green horizontal line in figure 5.2).

Residue-residue coupled QSFR metrics, cooperativity in torsion correlation index (CC_T), cooperativity in flexibility index correlation (CC), cooperativity in independent DOF correlation (CC_{IDF}), cooperativity in probability to rotate (CC_{PR}) and cooperativity in susceptibility correlation (CC_S) show similar trends. CC and CC_{PR} provide snapshots of intramolecular couplings within the protein structure. CC_T highlight secondary structure regions that have a dense network of locked-down torsion correlated residue pairs, while CC_{IDF} highlight regions that exhibit some degree of motion, especially in the loop regions. CC_S correlates residue susceptibility or mechanical fluctuations within mechanically connected regions of the network. All the above cooperativity-coupling

maps provide insight on allostery and functionally important regions of proteins. Figure 5.8 provides qualitative comparison between x-ray structure average QSFR metric and

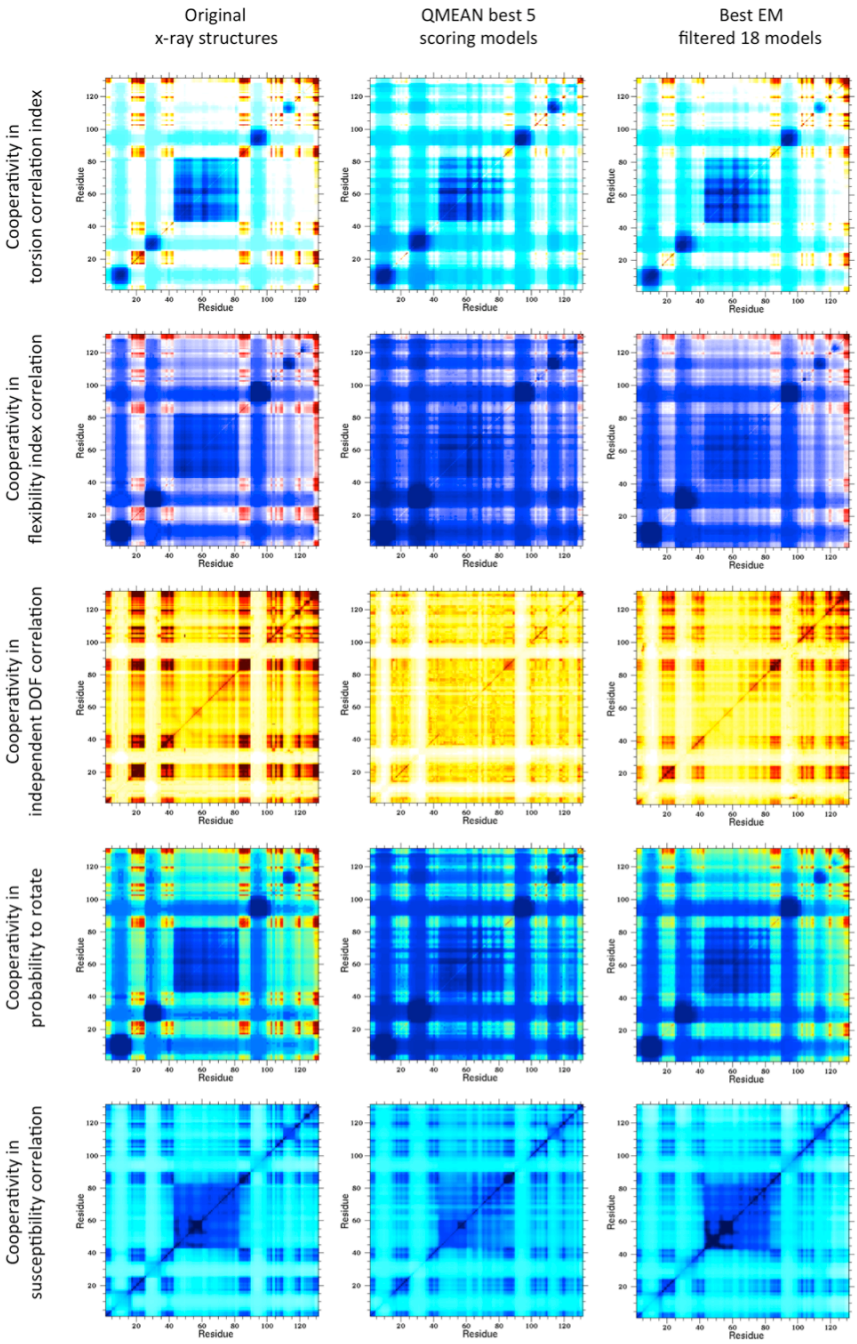

Figure 5.8: Comparison of residue-residue coupled QSFR metrics. A good qualitative resemblance is observed between EM 18-best models and original x-ray structures' average QSFR cooperativity metrics.

other homology model sets. These qualitative analyses suggest that EM filtered set has a very close agreement with x-ray structures. The homology model set derived from QMEAN fails to provide expected results. Similar benchmark quantification of CC plots is implemented using ±1σ. Calculation of cooperativity in probability to rotate metric using EM models set provide 92% accuracy, whereas QMEAN performs poorly at 44% (figure 5.9). All CC metric calculations using EM model set provide at least 70%

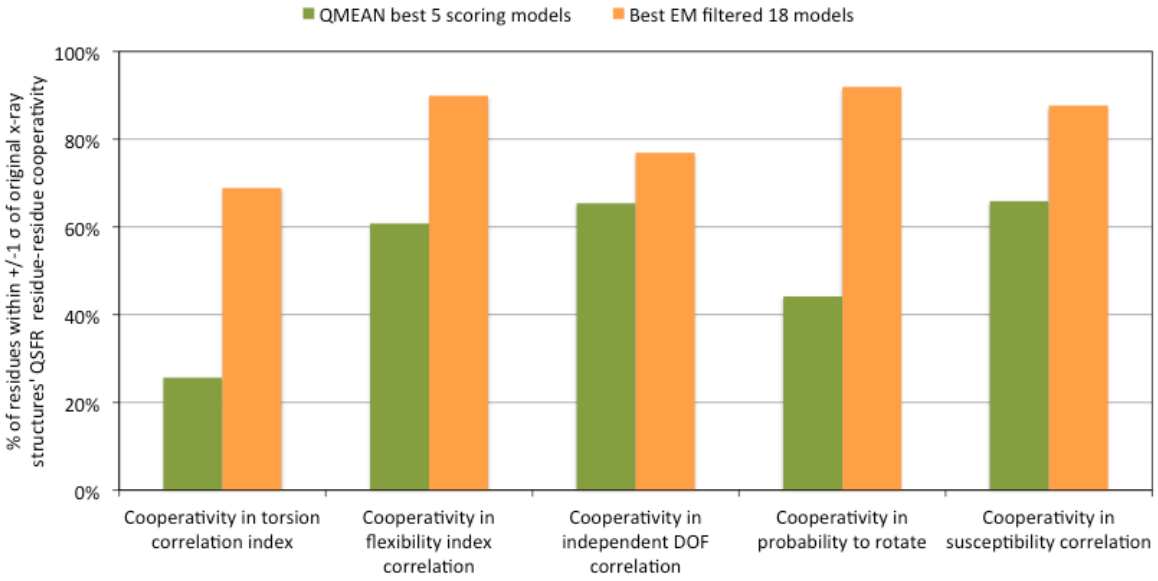

Figure 5.9: Comparison of residue-residue coupled QSFR metric accuracy levels for different model sets. Average QSFR quantities obtained from EM 18 best filtered models have better precision levels.

accuracy, whereas QMEAN 5 best models go as low as 26%. A scatter plot of CC metric values provides insight on correlation distribution (figure 5.10). Comparing QMEAN best 5 CC plot with original x-ray structure CC results in a wider distribution, while EM best 18 result in a narrow distribution underlining better CC calculation.

In this model exploration we are not suggesting that QMEAN model selection scoring system produces poor QSFR predictions. But, QMEAN coupled with thermodynamic and

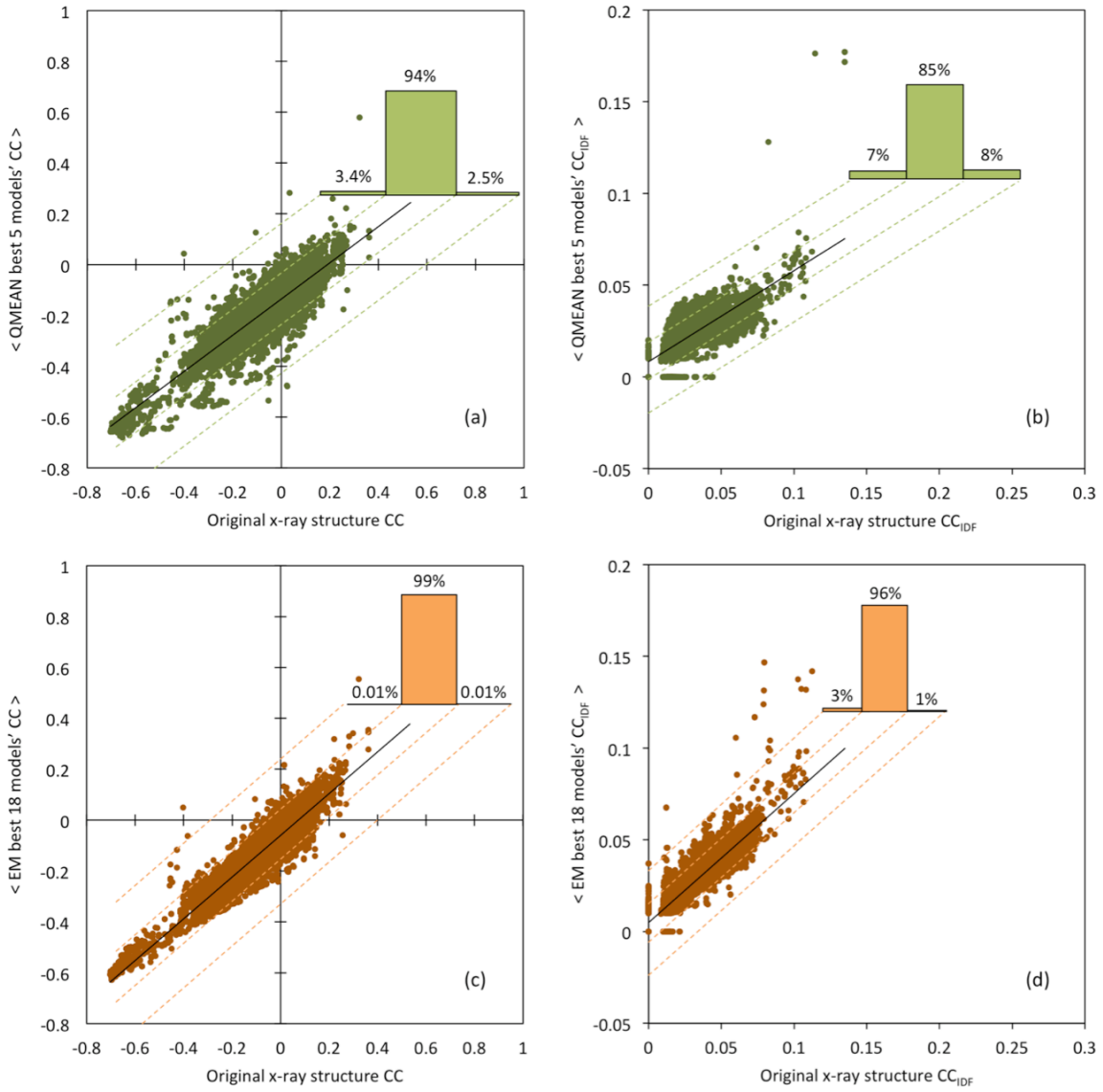

Figure 5.10: (a)(c) Residue-residue pair CC and (b)(d) CC_{IDF} values are compared against original x-ray structures. EM best 18 filtered models can reproduce highly accurate CC and CC_{IDF} properties of original x-ray structures as compared to QMEAN's best 5 models. QMEAN models' CC comparison with original x-ray structures exhibits wider data distribution, whereas EM models have a narrow distribution. Other metrics show similar trends. The black line shown is the best-fit regression across data points. The histograms are constructed by binning data points with equal intervals on y-axis on either side of the regression line. The interval size for binning is consistent across both, CC and CC_{IDF} plots, respectively.

mechanical quantities can boost statistics significantly resulting in better-estimated QSFR prediction. Since residue-residue coupled QSFR metrics are very sensitive, a robust

quantitative desired accuracy level achievement is very difficult. Yet, EM filtering does a very good job of filtering models, omitting the dependency of selecting models based purely on QMEAN information.

The above analyses do not attempt to establish thresholds. Instead, we have demonstrated that accurate filtering methods can provide precise QSFR estimates from models. The enriched data resulting from mDCM definitely provides opportunities to design other robust models. Nonetheless, this is the first step that delivers a confidence for achieving the goal of accurate QSFR prediction for unknown protein structures using homology-modeling technique.

5.5 Conclusion

Current results demonstrate that "good" homology models are able to sufficiently reproduce the x-ray QSFR data. Establishing clustering methodology on model accuracy based on their ability to reproduce QSFR metrics of x-ray structure will pave the way for comparative QSFR analysis. Expanding our analysis to over 100+ homolog structures across a family can dwarf all previous comparative exploration of protein flexibility and stability. This clearly shows that homology modeling represents a promising approach to drastically expand the "breadth" of our comparative QSFR analysis, which would help us better understand protein evolutionary relationships within families and superfamilies.

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation is an extensive investigation of protein's "Depth" and "Breadth" analysis. It provides a comprehensive assessment of variation and conservation of stability and flexibility characteristics within protein families due to mutations and evolutionary divergence. Furthermore, it provides a complete description of a protein accounting for its folding, stability, flexibility and function using Distance Constraint Model (DCM). Table 6.1 summarizes important results from each chapter.

Undertaking "depth" analysis has allowed us to better understand proteins' stability

Table 6.1: Chapter Conclusions

Chapter 2	DCM is parameterized for predicting stability of 14 point mutated c-type human lysozymes
	Parameter transferability predicts T_m's with more than 95% accuracy
	Correlation between changes in calculated and experimental T_m is 0.64
	Additional parameterization is not associated with accuracy
Chapter 3	DCM can detect subtle changes in protein flexibility due to point mutations
	Changes in flexibility upon mutation are common, large and long ranged
	β-subdomain exhibits large increase in flexibility, consistent with experimental observations
	Global dynamics of human and hen egg-white lysozyme are similar
Chapter 4	Residue pK_a shifts and active site electrostatic network is conserved across 12 β-lactamase proteins
	Electrostatic potential maps and CC properties vary, yet conserved within phylogenetic outgroups
	Protein backbone FI and H-bond network is conserved across family
	Ω-loop is marginally rigid highlighting its functional importance
Chapter 5	Comprehensive QSFR analysis across protein families is possible using homology models
	QMEAN scoring is viable but clustering is more promising to predict precise QSFR properties
	Models are EM clustered using structural, thermodynamical and mechanical properties
	Average QSFR properties of good homology models and x-ray structures are in good agreement

and flexibility changes due to single-site mutations. Under protein stability analysis, the parameters found from best fits to heat capacity curves for one or more c-type human lysozyme structures have subsequently been used to predict the heat capacity on the remaining. We have simulated a typical experimental situation, where prediction of relative stabilities in an untested mutated structure was based on known results as they accumulated. From the statistical significance of these simulations, we have established that the mDCM is a viable predictor for relative stability of protein mutants. Remarkably, using parameters from any single fitting yields an average percent error of 4.3%. Across the dataset, the mDCM reproduces experimental trends sufficiently well (R = 0.64) to be of practical value to experimentalists when making decisions about which mutations to invest time and funds for characterization. Assessing dynamical properties, our results suggest that small structural perturbations introduced by single point mutations have a frequent and pronounced affect on lysozyme flexibility that can extend over long distances. Specifically, an appreciable change occurs in backbone flexibility for 48% of the residues, and a change in cooperativity occurs in 42% of residue pairs. The average distance from mutation to a site with a change in flexibility is 17-20 Å. Interestingly, the frequency and scale of the changes within single point mutant structures are generally larger than those observed in the hen egg white lysozyme (HEWL) ortholog, which shares 61% sequence identity with human lysozyme. For example, point mutations often lead to substantial flexibility increases within the β-subdomain, which is consistent with experimental results indicating that it is the nucleation site for amyloid formation. However, β-subdomain flexibility within the human and HEWL orthologs is more similar despite the lowered sequence identity. These results suggest that compensating mutations

in HEWL reestablish desired properties.

On the other hand, the "breadth" analysis provides an insight on quantitative stability/flexibility relationship and other biophysical characterization of proteins within a family. Here, we have assessed systematic variations within physiochemical properties that underlie the different activities across twelve different class-A β-lactamases. Global conservation in per residue pK_a values, active-site electrostatic networks, and protein backbone rigidity suggests that common mechanistic strategies are employed across the family. Moreover, the Ω-loop, which is important for substrate recognition and catalysis, is consistently established to be marginally rigid. On the other hand, systematic differences within global electrostatic properties and pairwise residue-to-residue couplings are observed. Interestingly, these differences parallel evolutionary relationships, but do not reflect functional activities. These results reveal general insight into how physiochemical properties diverge during the course of enzyme family evolution, while also emphasizing that functional phenotypes can occur via multiple mechanistic approaches. Going further under "breadth" analysis, protocols for clustering/filtering of homology models based on their thermodynamic and mechanical quantities have also been developed, paving the way for comprehensive QSFR study of hundreds of proteins. Initial results indicate that homology model structures with similar structure, thermodynamic and dynamic properties yield accurate clustering using expectation maximization algorithm. Average QSFR quantities calculated from good homology models successfully reproduced x-ray structures' average QSFR properties. This is an important step towards a comprehensive QSFR analysis for hundreds of proteins. Table 6.2 highlights key results of this dissertation.

Table 6.2: Dissertation Highlights

Depth Analysis	Stability of proteins with single site mutations have been predicted with an accuracy of more than 95%
	Flexibility changes in proteins due to point mutations that have frequent, large and long-range effects have been elucidated
Breadth Analysis	Quantitative stability/flexibility relationships and biophysical properties of proteins in a family have been characterized
	Protocols for comprehensive stability/flexibility relationship analysis using homology models have been developed

With a good understanding of this top-down approach of proteins' sequence/structure/function relationships, researchers can employ a bottom-up approach in designing new proteins. That is, an ensemble of protein structures can be developed computationally and can be clustered based upon desired stability and flexibility properties, consequently paving the way for design of novel proteins using DCM.

REFERENCES

[1] K. A. Dill, "Dominant forces in protein folding," *Biochemistry,* vol. 29, pp. 7133-55, Aug 7 1990.

[2] K. Muller-Dethlefs and P. Hobza, "Noncovalent interactions: a challenge for experiment and theory," *Chem Rev,* vol. 100, pp. 143-68, Jan 12 2000.

[3] G. D. Rose and R. Wolfenden, "Hydrogen bonding, hydrophobicity, packing, and protein folding," *Annu Rev Biophys Biomol Struct,* vol. 22, pp. 381-415, 1993.

[4] D. E. Koshland, Jr., "Enzyme flexibility and enzyme action," *J Cell Comp Physiol,* vol. 54, pp. 245-58, Dec 1959.

[5] K. Linderstrom-Lang, "Structure and enzymatic break-down of proteins," *Cold Spring Harb Symp Quant Biol,* vol. 14, pp. 117-26, 1950.

[6] F. B. Straub, "FORMATION OF THE SECONDARY AND TERTIARY STRUCTURE OF ENZYMES," *Adv Enzymol Relat Areas Mol Biol,* vol. 26, pp. 89-114, 1964.

[7] *Protein Structure, Stability, and Folding*, 2001.

[8] M. R. Kasimova, S. M. Kristensen, P. W. Howe, T. Christensen, F. Matthiesen, J. Petersen, H. H. Sorensen, and J. J. Led, "NMR studies of the backbone flexibility and structure of human growth hormone: a comparison of high and low pH conformations," *J Mol Biol,* vol. 318, pp. 679-95, May 3 2002.

[9] D. J. Jacobs and S. Dallakyan, "Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity," *Biophys J,* vol. 88, pp. 903-15, Feb 2005.

[10] D. R. Livesay, S. Dallakyan, G. G. Wood, and D. J. Jacobs, "A flexible approach for understanding protein stability," *FEBS Lett,* vol. 576, pp. 468-76, Oct 22 2004.

[11] H. Meirovitch, "Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation," *Curr Opin Struct Biol,* vol. 17, pp. 181-6, Apr 2007.

[12] K. A. Dill, "Additivity principles in biochemistry," *J Biol Chem,* vol. 272, pp. 701-4, Jan 10 1997.

[13] A. E. Mark and W. F. van Gunsteren, "Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies," *J Mol Biol,* vol. 240, pp. 167-76, Jul 8 1994.

[14] D. J. Jacobs, S. Dallakyan, G. G. Wood, and A. Heckathorne, "Network rigidity at finite temperature: relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems," *Phys Rev E Stat Nonlin Soft Matter Phys,* vol. 68, p. 061109, Dec 2003.

[15] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe, "Protein flexibility predictions using graph theory," *Proteins,* vol. 44, pp. 150-65, Aug 1 2001.

[16] D. J. Jacobs and M. F. Thorpe, "Generic rigidity percolation: The pebble game," *Phys Rev Lett,* vol. 75, pp. 4051-4054, Nov 27 1995.

[17] O. K. Vorov, D. R. Livesay, and D. J. Jacobs, "Helix/coil nucleation: a local response to global demands," *Biophys J,* vol. 97, pp. 3000-9, Dec 2 2009.

[18] B. I. Dahiyat, D. B. Gordon, and S. L. Mayo, "Automated design of the surface positions of protein helices," *Protein Sci,* vol. 6, pp. 1333-7, Jun 1997.

[19] A. Fernandez, J. Kardos, and Y. Goto, "Protein folding: could hydrophobic collapse be coupled with hydrogen-bond formation?," *FEBS Lett,* vol. 536, pp. 187-92, Feb 11 2003.

[20] D. R. Livesay and D. J. Jacobs, "Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair," *Proteins,* vol. 62, pp. 130-43, Jan 1 2006.

[21] D. J. Jacobs, D. R. Livesay, J. Hules, and M. L. Tasayco, "Elucidating quantitative stability/flexibility relationships within thioredoxin and its fragments using a distance constraint model," *J Mol Biol,* vol. 358, pp. 882-904, May 5 2006.

[22] D. R. Livesay, D. H. Huynh, S. Dallakyan, and D. J. Jacobs, "Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family," *Chem Cent J,* vol. 2, p. 17, 2008.

[23] J. M. Mottonen, M. Xu, D. J. Jacobs, and D. R. Livesay, "Unifying mechanical and thermodynamic descriptions across the thioredoxin protein family," *Proteins,* vol. 75, pp. 610-27, May 15 2009.

[24] D. Verma, D. J. Jacobs, and D. R. Livesay, "Predicting the melting point of human C-type lysozyme mutants," *Curr Protein Pept Sci,* vol. 11, pp. 562-72, Nov 2010.

[25] D. Verma, D. J. Jacobs, and D. R. Livesay, "Changes in Lysozyme Flexibility upon Mutation Are Frequent, Large and Long-Ranged," *PLoS Comput Biol,* vol. 8, p. e1002409, 2012.

[26] L. T. Huang, K. Saraboji, S. Y. Ho, S. F. Hwang, M. N. Ponnuswamy, and M. M. Gromiha, "Prediction of protein mutant stability using classification and regression tool," *Biophys Chem,* vol. 125, pp. 462-70, Feb 2007.

[27] L. T. Huang, M. M. Gromiha, and S. Y. Ho, "Sequence analysis and rule development of predicting protein stability change upon mutation using decision tree model," *J Mol Model,* vol. 13, pp. 879-90, Aug 2007.

[28] L. T. Huang, M. M. Gromiha, and S. Y. Ho, "iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations," *Bioinformatics,* vol. 23, pp. 1292-3, May 15 2007.

[29] J. Cheng, A. Randall, and P. Baldi, "Prediction of protein stability changes for single-site mutations using support vector machines," *Proteins,* vol. 62, pp. 1125-32, Mar 1 2006.

[30] E. Capriotti, P. Fariselli, R. Calabrese, and R. Casadio, "Predicting protein stability changes from sequences using support vector machines," *Bioinformatics,* vol. 21 Suppl 2, pp. ii54-8, Sep 1 2005.

[31] E. Capriotti, P. Fariselli, and R. Casadio, "I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure," *Nucleic Acids Res,* vol. 33, pp. W306-10, Jul 1 2005.

[32] B. Rost and C. Sander, "Combining evolutionary information and neural networks to predict protein secondary structure," *Proteins,* vol. 19, pp. 55-72, May 1994.

[33] J. Caballero, L. Fernandez, J. I. Abreu, and M. Fernandez, "Amino Acid Sequence Autocorrelation vectors and ensembles of Bayesian-Regularized Genetic Neural Networks for prediction of conformational stability of human lysozyme mutants," *J Chem Inf Model,* vol. 46, pp. 1255-68, May-Jun 2006.

[34] E. Capriotti, P. Fariselli, and R. Casadio, "A neural-network-based method for predicting protein stability changes upon single point mutations," *Bioinformatics,* vol. 20 Suppl 1, pp. i63-8, Aug 4 2004.

[35] T. Alber, "Mutational effects on protein stability," *Annu Rev Biochem,* vol. 58, pp. 765-98, 1989.

[36] S. S. Strickler, A. V. Gribenko, T. R. Keiffer, J. Tomlinson, T. Reihle, V. V. Loladze, and G. I. Makhatadze, "Protein stability and surface electrostatics: a charged relationship," *Biochemistry,* vol. 45, pp. 2761-6, Mar 7 2006.

[37] G. I. Makhatadze, V. V. Loladze, D. N. Ermolenko, X. Chen, and S. T. Thomas, "Contribution of surface salt bridges to protein stability: guidelines for protein engineering," *J Mol Biol,* vol. 327, pp. 1135-48, Apr 11 2003.

[38] G. I. Makhatadze, V. V. Loladze, A. V. Gribenko, and M. M. Lopez, "Mechanism of thermostabilization in a designed cold shock protein with optimized surface electrostatic interactions," *J Mol Biol,* vol. 336, pp. 929-42, Feb 27 2004.

[39] M. Torrez, M. Schultehenrich, and D. R. Livesay, "Conferring thermostability to mesophilic proteins through optimized electrostatic surfaces," *Biophys J,* vol. 85, pp. 2845-53, Nov 2003.

[40] F. Dong, M. Vijayakumar, and H. X. Zhou, "Comparison of calculation and experiment implicates significant electrostatic contributions to the binding stability of barnase and barstar," *Biophys J,* vol. 85, pp. 49-60, Jul 2003.

[41] H. X. Zhou and F. Dong, "Electrostatic contributions to the stability of a thermophilic cold shock protein," *Biophys J,* vol. 84, pp. 2216-22, Apr 2003.

[42] F. Dong and H. X. Zhou, "Electrostatic contributions to T4 lysozyme stability: solvent-exposed charges versus semi-buried salt bridges," *Biophys J,* vol. 83, pp. 1341-7, Sep 2002.

[43] D. J. Jacobs and S. Dallakyan, "Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity," *Biophys J,* vol. 88, pp. 903-15, Feb 2005.

[44] D. R. Livesay, S. Dallakyan, G. G. Wood, and D. J. Jacobs, "A flexible approach for understanding protein stability," *FEBS Lett,* vol. 576, pp. 468-76, Oct 22 2004.

[45] A. Razvi and J. M. Scholtz, "Lessons in stability from thermophilic proteins," *Protein Sci,* vol. 15, pp. 1569-78, Jul 2006.

[46] Y. Yamagata, M. Kubota, Y. Sumikawa, J. Funahashi, K. Takano, S. Fujii, and K. Yutani, "Contribution of hydrogen bonds to the conformational stability of human lysozyme: calorimetry and X-ray analysis of six tyrosine --> phenylalanine mutants," *Biochemistry,* vol. 37, pp. 9355-62, Jun 30 1998.

[47] K. Takano, Y. Yamagata, S. Fujii, and K. Yutani, "Contribution of the hydrophobic effect to the stability of human lysozyme: calorimetric studies and X-ray structural analyses of the nine valine to alanine mutants," *Biochemistry,* vol. 36, pp. 688-98, Jan 28 1997.

[48] K. Takano, M. Ota, K. Ogasahara, Y. Yamagata, K. Nishikawa, and K. Yutani, "Experimental verification of the 'stability profile of mutant protein' (SPMP) data using mutant human lysozymes," *Protein Eng,* vol. 12, pp. 663-72, Aug 1999.

[49] K. Takano, K. Tsuchimori, Y. Yamagata, and K. Yutani, "Effect of foreign N-terminal residues on the conformational stability of human lysozyme," *Eur J Biochem,* vol. 266, pp. 675-82, Dec 1999.

[50] J. Funahashi, K. Takano, Y. Yamagata, and K. Yutani, "Contribution of amino acid substitutions at two different interior positions to the conformational stability of human lysozyme," *Protein Eng,* vol. 12, pp. 841-50, Oct 1999.

[51] J. Funahashi, K. Takano, K. Ogasahara, Y. Yamagata, and K. Yutani, "The structure, stability, and folding process of amyloidogenic mutant human lysozyme," *J Biochem,* vol. 120, pp. 1216-23, Dec 1996.

[52] T. Herning, K. Yutani, K. Inaka, R. Kuroki, M. Matsushima, and M. Kikuchi, "Role of proline residues in human lysozyme stability: a scanning calorimetric study combined with X-ray structure analysis of proline mutants," *Biochemistry,* vol. 31, pp. 7077-85, Aug 11 1992.

[53] A. D. Robertson and K. P. Murphy, "Protein Structure and the Energetics of Protein Stability," *Chem Rev,* vol. 97, pp. 1251-1268, Aug 5 1997.

[54] J. C. Gordon, J. B. Myers, T. Folta, V. Shoja, L. S. Heath, and A. Onufriev, "H++: a server for estimating pKas and adding missing hydrogens to macromolecules," *Nucleic Acids Res,* vol. 33, pp. W368-71, Jul 1 2005.

[55] J. Gomez, V. J. Hilser, D. Xie, and E. Freire, "The heat capacity of proteins," *Proteins,* vol. 22, pp. 404-12, Aug 1995.

[56] D. J. Jacobs, D. R. Livesay, J. Hules, and M. L. Tasayco, "Elucidating quantitative stability/flexibility relationships within thioredoxin and its fragments using a distance constraint model," *J Mol Biol,* vol. 358, pp. 882-904, May 5 2006.

[57] D. R. Livesay, D. H. Huynh, S. Dallakyan, and D. J. Jacobs, "Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family," *Chem Cent J,* vol. 2, p. 17, 2008.

[58] D. R. Livesay and D. J. Jacobs, "Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair," *Proteins,* vol. 62, pp. 130-43, Jan 1 2006.

[59] J. M. Mottonen, M. Xu, D. J. Jacobs, and D. R. Livesay, "Unifying mechanical and thermodynamic descriptions across the thioredoxin protein family," *Proteins,* vol. 75, pp. 610-27, May 15 2009.

[60] C. N. Pace, B. A. Shirley, M. McNutt, and K. Gajiwala, "Forces contributing to the conformational stability of proteins," *FASEB J,* vol. 10, pp. 75-83, Jan 1996.

[61] R. Guerois, J. E. Nielsen, and L. Serrano, "Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations," *J Mol Biol,* vol. 320, pp. 369-87, Jul 5 2002.

[62] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern, "Intrinsic dynamics of an enzyme underlies catalysis," *Nature,* vol. 438, pp. 117-21, Nov 3 2005.

[63] S. D. Khare and N. V. Dokholyan, "Common dynamical signatures of familial amyotrophic lateral sclerosis-associated structurally diverse Cu, Zn superoxide

dismutase mutants," *Proc Natl Acad Sci U S A,* vol. 103, pp. 3147-52, Feb 28 2006.

[64] C. J. Tsai, A. del Sol, and R. Nussinov, "Allostery: absence of a change in shape does not imply that allostery is not at play," *J Mol Biol,* vol. 378, pp. 1-11, Apr 18 2008.

[65] S. T. Hsu, G. Blaser, and S. E. Jackson, "The folding, stability and conformational dynamics of beta-barrel fluorescent proteins," *Chem Soc Rev,* vol. 38, pp. 2951-65, Oct 2009.

[66] A. Razvi and J. M. Scholtz, "Lessons in stability from thermophilic proteins," *Protein Sci,* vol. 15, pp. 1569-78, Jul 2006.

[67] N. Tokuriki and D. S. Tawfik, "Stability effects of mutations and protein evolvability," *Curr Opin Struct Biol,* vol. 19, pp. 596-604, Oct 2009.

[68] A. Krushelnitsky, T. Zinkevich, N. Mukhametshina, N. Tarasova, Y. Gogolev, O. Gnezdilov, V. Fedotov, P. Belton, and D. Reichert, "13C and 15N NMR study of the hydration response of T4 lysozyme and alphaB-crystallin internal dynamics," *J Phys Chem B,* vol. 113, pp. 10022-34, Jul 23 2009.

[69] G. A. Cook, H. Zhang, S. H. Park, Y. Wang, and S. J. Opella, "Comparative NMR studies demonstrate profound differences between two viroporins: p7 of HCV and Vpu of HIV-1," *Biochim Biophys Acta,* vol. 1808, pp. 554-60, Feb 2011.

[70] B. O. Brandsdal, E. S. Heimstad, I. Sylte, and A. O. Smalas, "Comparative molecular dynamics of mesophilic and psychrophilic protein homologues studied by 1.2 ns simulations," *J Biomol Struct Dyn,* vol. 17, pp. 493-506, Dec 1999.

[71] E. S. Heimstad, L. K. Hansen, and A. O. Smalas, "Comparative molecular dynamics simulation studies of salmon and bovine trypsins in aqueous solution," *Protein Eng,* vol. 8, pp. 379-88, Apr 1995.

[72] A. Pang, Y. Arinaminpathy, M. S. Sansom, and P. C. Biggin, "Comparative molecular dynamics--similar folds and similar motions?," *Proteins,* vol. 61, pp. 809-22, Dec 1 2005.

[73] A. Brigo, K. W. Lee, F. Fogolari, G. I. Mustata, and J. M. Briggs, "Comparative molecular dynamics simulations of HIV-1 integrase and the T66I/M154I mutant: binding modes and drug resistance to a diketo acid inhibitor," *Proteins,* vol. 59, pp. 723-41, Jun 1 2005.

[74] K. Cox and M. S. Sansom, "One membrane protein, two structures and six environments: a comparative molecular dynamics simulation study of the bacterial outer membrane protein PagP," *Mol Membr Biol,* vol. 26, pp. 205-14, May 2009.

[75] L. Liu, L. M. Koharudin, A. M. Gronenborn, and I. Bahar, "A comparative analysis of the equilibrium dynamics of a designed protein inferred from NMR, X-ray, and computations," *Proteins,* vol. 77, pp. 927-39, Dec 2009.

[76] M. C. Zwier and L. T. Chong, "Reaching biological timescales with all-atom molecular dynamics simulations," *Curr Opin Pharmacol,* vol. 10, pp. 745-752, Oct 7 2010.

[77] D. J. Jacobs and S. Dallakyan, "Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity," *Biophys J,* vol. 88, pp. 903-15, Feb 2005.

[78] D. R. Livesay, S. Dallakyan, G. G. Wood, and D. J. Jacobs, "A flexible approach for understanding protein stability," *FEBS Lett,* vol. 576, pp. 468-76, Oct 22 2004.

[79] D. J. Jacobs, D. R. Livesay, J. Hules, and M. L. Tasayco, "Elucidating quantitative stability/flexibility relationships within thioredoxin and its fragments using a distance constraint model," *J Mol Biol,* vol. 358, pp. 882-904, May 5 2006.

[80] D. R. Livesay and D. J. Jacobs, "Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair," *Proteins,* vol. 62, pp. 130-43, Jan 1 2006.

[81] D. Verma, D. J. Jacobs, and D. R. Livesay, "Predicting the melting point of human C-type lysozyme mutants," *Curr Protein Pept Sci,* vol. 11, pp. 562-72, Nov 2010.

[82] D. R. Booth, M. Sunde, V. Bellotti, C. V. Robinson, W. L. Hutchinson, P. E. Fraser, P. N. Hawkins, C. M. Dobson, S. E. Radford, C. C. Blake, and M. B. Pepys, "Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis," *Nature,* vol. 385, pp. 787-93, Feb 27 1997.

[83] S. E. Radford, C. M. Dobson, and P. A. Evans, "The folding of hen lysozyme involves partially structured intermediates and multiple pathways," *Nature,* vol. 358, pp. 302-7, Jul 23 1992.

[84] S. E. Radford and C. M. Dobson, "Insights into protein folding using physical techniques: studies of lysozyme and alpha-lactalbumin," *Philos Trans R Soc Lond B Biol Sci,* vol. 348, pp. 17-25, Apr 29 1995.

[85] L. C. Wu, Z. Y. Peng, and P. S. Kim, "Bipartite structure of the alpha-lactalbumin molten globule," *Nat Struct Biol,* vol. 2, pp. 281-6, Apr 1995.

[86] M. Dumoulin, D. Canet, A. M. Last, E. Pardon, D. B. Archer, S. Muyldermans, L. Wyns, A. Matagne, C. V. Robinson, C. Redfield, and C. M. Dobson, "Reduced global cooperativity is a common feature underlying the amyloidogenicity of pathogenic lysozyme mutations," *J Mol Biol,* vol. 346, pp. 773-88, Feb 25 2005.

[87] Y. Joti, M. Nakasako, A. Kidera, and N. Go, "Nonlinear temperature dependence of the crystal structure of lysozyme: correlation between coordinate shifts and thermal factors," *Acta Crystallogr D Biol Crystallogr,* vol. 58, pp. 1421-32, Sep 2002.

[88] M. Muraki, K. Harata, N. Sugita, and K. Sato, "Origin of carbohydrate recognition specificity of human lysozyme revealed by affinity labeling," *Biochemistry,* vol. 35, pp. 13562-7, Oct 22 1996.

[89] Y. Yamagata, M. Kubota, Y. Sumikawa, J. Funahashi, K. Takano, S. Fujii, and K. Yutani, "Contribution of hydrogen bonds to the conformational stability of human lysozyme: calorimetry and X-ray analysis of six tyrosine --> phenylalanine mutants," *Biochemistry,* vol. 37, pp. 9355-62, Jun 30 1998.

[90] K. Takano, Y. Yamagata, S. Fujii, and K. Yutani, "Contribution of the hydrophobic effect to the stability of human lysozyme: calorimetric studies and X-ray structural analyses of the nine valine to alanine mutants," *Biochemistry,* vol. 36, pp. 688-98, Jan 28 1997.

[91] K. Takano, K. Tsuchimori, Y. Yamagata, and K. Yutani, "Effect of foreign N-terminal residues on the conformational stability of human lysozyme," *Eur J Biochem,* vol. 266, pp. 675-82, Dec 1999.

[92] K. Takano, M. Ota, K. Ogasahara, Y. Yamagata, K. Nishikawa, and K. Yutani, "Experimental verification of the 'stability profile of mutant protein' (SPMP) data using mutant human lysozymes," *Protein Eng,* vol. 12, pp. 663-72, Aug 1999.

[93] J. Funahashi, K. Takano, Y. Yamagata, and K. Yutani, "Contribution of amino acid substitutions at two different interior positions to the conformational stability of human lysozyme," *Protein Eng,* vol. 12, pp. 841-50, Oct 1999.

[94] J. Funahashi, K. Takano, K. Ogasahara, Y. Yamagata, and K. Yutani, "The structure, stability, and folding process of amyloidogenic mutant human lysozyme," *J Biochem,* vol. 120, pp. 1216-23, Dec 1996.

[95] T. Herning, K. Yutani, K. Inaka, R. Kuroki, M. Matsushima, and M. Kikuchi, "Role of proline residues in human lysozyme stability: a scanning calorimetric study combined with X-ray structure analysis of proline mutants," *Biochemistry,* vol. 31, pp. 7077-85, Aug 11 1992.

[96] J. C. Gordon, J. B. Myers, T. Folta, V. Shoja, L. S. Heath, and A. Onufriev, "H++: a server for estimating pKas and adding missing hydrogens to macromolecules," *Nucleic Acids Res,* vol. 33, pp. W368-71, Jul 1 2005.

[97] J. W. Ponder and D. A. Case, "Force fields for protein simulations," *Adv Protein Chem,* vol. 66, pp. 27-85, 2003.

[98] J. M. Mottonen, D. J. Jacobs, and D. R. Livesay, "Allosteric response is both conserved and variable across three CheY orthologs," *Biophys J,* vol. 99, pp. 2245-54, Oct 6 2010.

[99] M. F. Thorpe and P. M. Duxbury, *Rigidity Theory and Applications*. New York: Kluwer Academic / Plenum Publishers, 1999.

[100] D. R. Livesay, D. H. Huynh, S. Dallakyan, and D. J. Jacobs, "Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family," *Chem Cent J,* vol. 2, p. 17, 2008.

[101] J. M. Mottonen, M. Xu, D. J. Jacobs, and D. R. Livesay, "Unifying mechanical and thermodynamic descriptions across the thioredoxin protein family," *Proteins,* vol. 75, pp. 610-27, May 15 2009.

[102] J. Higo and M. Nakasako, "Hydration structure of human lysozyme investigated by molecular dynamics simulation and cryogenic X-ray crystal structure analyses: on the correlation between crystal water sites, solvent density, and solvent dipole," *J Comput Chem,* vol. 23, pp. 1323-36, Nov 15 2002.

[103] P. J. Artymiuk and C. C. Blake, "Refinement of human lysozyme at 1.5 A resolution analysis of non-bonded and hydrogen-bond interactions," *J Mol Biol,* vol. 152, pp. 737-62, Nov 15 1981.

[104] H. Song, K. Inaka, K. Maenaka, and M. Matsushima, "Structural changes of active site cleft and different saccharide binding modes in human lysozyme co-crystallized with hexa-N-acetyl-chitohexaose at pH 4.0," *J Mol Biol,* vol. 244, pp. 522-40, Dec 16 1994.

[105] T. Durek, V. Y. Torbeev, and S. B. Kent, "Convergent chemical synthesis and high-resolution x-ray structure of human lysozyme," *Proc Natl Acad Sci U S A,* vol. 104, pp. 4846-51, Mar 20 2007.

[106] D. J. Jacobs, "Predicting protein flexibility and stability using network rigidity: a new modeling paradigm," in *Recent Research Developments in Biophysics*. vol. 5, ed Trivandrum, India: Transworld Research Network, 2006, pp. 71-131.

[107] G. Rhodes, *Crystallography Made Crystal Clear*, 3rd Edition ed.: Academic Press, 2006.

[108] D. K. Smith, P. Radivojac, Z. Obradovic, A. K. Dunker, and G. Zhu, "Improved amino acid flexibility parameters," *Protein Sci,* vol. 12, pp. 1060-72, May 2003.

[109] M. Guzman-Casado, A. Parody-Morreale, S. Robic, S. Marqusee, and J. M. Sanchez-Ruiz, "Energetic evidence for formation of a pH-dependent hydrophobic cluster in the denatured state of Thermus thermophilus ribonuclease H," *J Mol Biol,* vol. 329, pp. 731-43, Jun 13 2003.

[110] J. Hollien and S. Marqusee, "A thermodynamic comparison of mesophilic and thermophilic ribonucleases H," *Biochemistry,* vol. 38, pp. 3831-6, Mar 23 1999.

[111] J. Hollien and S. Marqusee, "Structural distribution of stability in a thermophilic enzyme," *Proc Natl Acad Sci U S A,* vol. 96, pp. 13674-8, Nov 23 1999.

[112] S. Robic, M. Guzman-Casado, J. M. Sanchez-Ruiz, and S. Marqusee, "Role of residual structure in the unfolded state of a thermophilic protein," *Proc Natl Acad Sci U S A,* vol. 100, pp. 11345-9, Sep 30 2003.

[113] H. S. Cho, S. Y. Lee, D. Yan, X. Pan, J. S. Parkinson, S. Kustu, D. E. Wemmer, and J. G. Pelton, "NMR structure of activated CheY," *J Mol Biol,* vol. 297, pp. 543-51, Mar 31 2000.

[114] X. Zhu, C. D. Amsler, K. Volz, and P. Matsumura, "Tyrosine 106 of CheY plays an important role in chemotaxis signal transduction in Escherichia coli," *J Bacteriol,* vol. 178, pp. 4208-15, Jul 1996.

[115] D. R. Livesay, K. E. Kreth, and A. A. Fodor, "A critical evaluation of correlated mutation algorithms and coevolution within allosteric mechanisms " in *Allostery: Methods and Protocols*, A. W. Fenton, Ed., ed: Humana Press, In press.

[116] A. J. Rader, B. M. Hespenheide, L. A. Kuhn, and M. F. Thorpe, "Protein unfolding: rigidity lost," *Proc Natl Acad Sci U S A,* vol. 99, pp. 3540-5, Mar 19 2002.

[117] F. A. Mulder, B. Hon, D. R. Muhandiram, F. W. Dahlquist, and L. E. Kay, "Flexibility and ligand exchange in a buried cavity mutant of T4 lysozyme studied by multinuclear NMR," *Biochemistry,* vol. 39, pp. 12614-22, Oct 17 2000.

[118] J. Liu and J. Song, "Insights into protein aggregation by NMR characterization of insoluble SH3 mutants solubilized in salt-free water," *PLoS One,* vol. 4, p. e7805, 2009.

[119] H. J. Lee, Y. J. Yoon, S. Jang do, C. Kim, H. J. Cha, B. H. Hong, K. Y. Choi, and H. C. Lee, "15N NMR relaxation studies of Y14F mutant of ketosteroid isomerase: the influence of mutation on backbone mobility," *J Biochem,* vol. 144, pp. 159-66, Aug 2008.

[120] Y. Wen, J. Li, M. Xiong, Y. Peng, W. Yao, J. Hong, and D. Lin, "Solution structure and dynamics of the I214V mutant of the rabbit prion protein," *PLoS One,* vol. 5, p. e13273, 2010.

[121] X. Yuan, J. M. Werner, J. Lack, V. Knott, P. A. Handford, I. D. Campbell, and A. K. Downing, "Effects of the N2144S mutation on backbone dynamics of a TB-cbEGF domain pair from human fibrillin-1," *J Mol Biol,* vol. 316, pp. 113-25, Feb 8 2002.

[122] A. Mittermaier and L. E. Kay, "The response of internal dynamics to hydrophobic core mutations in the SH3 domain from the Fyn tyrosine kinase," *Protein Sci,* vol. 13, pp. 1088-99, Apr 2004.

[123] J. L. Battiste, R. Li, and C. Woodward, "A highly destabilizing mutation, G37A, of the bovine pancreatic trypsin inhibitor retains the average native conformation but greatly increases local flexibility," *Biochemistry,* vol. 41, pp. 2237-45, Feb 19 2002.

[124] R. S. Johnson, "Mass Spectrometric Measurement of Changes in Protein Hydrogen Exchange Rates that Result from Point Mutations," *J Am Soc Mass Spectrom,* vol. 7, pp. 515-521, 1996.

[125] J. A. Boyer and A. L. Lee, "Monitoring aromatic picosecond to nanosecond dynamics in proteins via 13C relaxation: expanding perturbation mapping of the rigidifying core mutation, V54A, in eglin c," *Biochemistry,* vol. 47, pp. 4876-86, Apr 29 2008.

[126] O. Millet, A. Mittermaier, D. Baker, and L. E. Kay, "The effects of mutations on motions of side-chains in protein L studied by 2H NMR dynamics and scalar couplings," *J Mol Biol,* vol. 329, pp. 551-63, Jun 6 2003.

[127] T. I. Igumenova, A. L. Lee, and A. J. Wand, "Backbone and side chain dynamics of mutant calmodulin-peptide complexes," *Biochemistry,* vol. 44, pp. 12627-39, Sep 27 2005.

[128] M. W. Clarkson and A. L. Lee, "Long-range dynamic effects of point mutations propagate through side chains in the serine protease inhibitor eglin c," *Biochemistry,* vol. 43, pp. 12448-58, Oct 5 2004.

[129] A. K. Chamberlain, V. Receveur, A. Spencer, C. Redfield, and C. M. Dobson, "Characterization of the structure and dynamics of amyloidogenic variants of human lysozyme by NMR spectroscopy," *Protein Sci,* vol. 10, pp. 2525-30, Dec 2001.

[130] J. D. Bloom, A. Raval, and C. O. Wilke, "Thermodynamics of neutral protein evolution," *Genetics,* vol. 175, pp. 255-66, Jan 2007.

[131] J. D. Bloom, C. O. Wilke, F. H. Arnold, and C. Adami, "Stability and the evolvability of function in a model protein," *Biophys J,* vol. 86, pp. 2758-64, May 2004.

[132] P. D. Williams, D. D. Pollock, B. P. Blackburne, and R. A. Goldstein, "Assessing the Accuracy of Ancestral Protein Reconstruction Methods," *PLoS Comput Biol,* vol. 2, p. e69, 2006.

[133] X. Gu, "Functional divergence in protein (family) sequence evolution," *Genetica,* vol. 118, pp. 133-41, Jul 2003.

[134] A. M. Altenhoff, R. A. Studer, M. Robinson-Rechavi, and C. Dessimoz, "Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs," *PLoS Comput Biol,* vol. 8, p. e1002514, May 2012.

[135] S. Donadio, S. Maffioli, P. Monciardini, M. Sosio, and D. Jabes, "Antibiotic discovery in the twenty-first century: current trends and future perspectives," *J Antibiot (Tokyo),* vol. 63, pp. 423-30, Aug 2010.

[136] E. P. Abraham and E. Chain, "An enzyme from bacteria able to destroy penicillin. 1940," *Rev Infect Dis,* vol. 10, pp. 677-8, Jul-Aug 1988.

[137] F. K. Majiduddin, I. C. Materon, and T. G. Palzkill, "Molecular analysis of beta-lactamase structure and function," *Int J Med Microbiol,* vol. 292, pp. 127-37, Jul 2002.

[138] S. O. Meroueh, G. Minasov, W. Lee, B. K. Shoichet, and S. Mobashery, "Structural aspects for evolution of beta-lactamases from penicillin-binding proteins," *J Am Chem Soc,* vol. 125, pp. 9612-8, Aug 13 2003.

[139] G. A. Jacoby and L. S. Munoz-Price, "The new beta-lactamases," *N Engl J Med,* vol. 352, pp. 380-91, Jan 27 2005.

[140] D. L. Paterson, K. M. Hujer, A. M. Hujer, B. Yeiser, M. D. Bonomo, L. B. Rice, and R. A. Bonomo, "Extended-spectrum beta-lactamases in Klebsiella pneumoniae bloodstream isolates from seven countries: dominance and widespread prevalence of SHV- and CTX-M-type beta-lactamases," *Antimicrob Agents Chemother,* vol. 47, pp. 3554-60, Nov 2003.

[141] C. L. Emery and L. A. Weymouth, "Detection and clinical significance of extended-spectrum beta-lactamases in a tertiary-care medical center," *J Clin Microbiol,* vol. 35, pp. 2061-7, Aug 1997.

[142] J. Y. Kim, H. I. Jung, Y. J. An, J. H. Lee, S. J. Kim, S. H. Jeong, K. J. Lee, P. G. Suh, H. S. Lee, S. H. Lee, and S. S. Cha, "Structural basis for the extended substrate spectrum of CMY-10, a plasmid-encoded class C beta-lactamase," *Mol Microbiol,* vol. 60, pp. 907-16, May 2006.

[143] J. F. Fisher, S. O. Meroueh, and S. Mobashery, "Bacterial resistance to beta-lactam antibiotics: compelling opportunism, compelling opportunity," *Chem Rev,* vol. 105, pp. 395-424, Feb 2005.

[144] K. Bush, G. A. Jacoby, and A. A. Medeiros, "A functional classification scheme for beta-lactamases and its correlation with molecular structure," *Antimicrob Agents Chemother,* vol. 39, pp. 1211-33, Jun 1995.

[145] D. R. Livesay, P. Jambeck, A. Rojnuckarin, and S. Subramaniam, "Conservation of electrostatic properties within enzyme families and superfamilies," *Biochemistry,* vol. 42, pp. 3464-73, Apr 1 2003.

[146] D. R. Livesay and D. La, "The evolutionary origins and catalytic importance of conserved electrostatic networks within TIM-barrel proteins," *Protein Sci,* vol. 14, pp. 1158-70, May 2005.

[147] J. Lamotte-Brasseur, J. Knox, J. A. Kelly, P. Charlier, E. Fonze, O. Dideberg, and J. M. Frere, "The structures and catalytic mechanisms of active-site serine beta-lactamases," *Biotechnol Genet Eng Rev,* vol. 12, pp. 189-230, 1994.

[148] R. M. Gibson, H. Christensen, and S. G. Waley, "Site-directed mutagenesis of beta-lactamase I. Single and double mutants of Glu-166 and Lys-73," *Biochem J,* vol. 272, pp. 613-9, Dec 15 1990.

[149] J. Lamotte-Brasseur, F. Jacob-Dubuisson, G. Dive, J. M. Frere, and J. M. Ghuysen, "Streptomyces albus G serine beta-lactamase. Probing of the catalytic mechanism via molecular modelling of mutant enzymes," *Biochem J,* vol. 282 (Pt 1), pp. 189-95, Feb 15 1992.

[150] J. Lamotte-Brasseur, G. Dive, O. Dideberg, P. Charlier, J. M. Frere, and J. M. Ghuysen, "Mechanism of acyl transfer by the class A serine beta-lactamase of Streptomyces albus G," *Biochem J,* vol. 279 (Pt 1), pp. 213-21, Oct 1 1991.

[151] N. C. J. Strynadka, Martin, R., Jensen, S. E., Gold, M. and Jones, J. B. (1996) 3, 688–695, *Nature Struct. Biol. ,* vol. 3, pp. 688-695, 1996.

[152] O. a. M. Herzberg, J., *Curr. Opin. Struct. Biol. ,* vol. 1, pp. 946-953, 1991.

[153] N. C. Strynadka, H. Adachi, S. E. Jensen, K. Johns, A. Sielecki, C. Betzel, K. Sutoh, and M. N. James, "Molecular structure of the acyl-enzyme intermediate in beta-lactam hydrolysis at 1.7 A resolution," *Nature,* vol. 359, pp. 700-5, Oct 22 1992.

[154] H. Adachi, T. Ohta, and H. Matsuzawa, "Site-directed mutants, at position 166, of RTEM-1 beta-lactamase that form a stable acyl-enzyme intermediate with penicillin," *J Biol Chem,* vol. 266, pp. 3186-91, Feb 15 1991.

[155] P. Swarén, L. Maveyraud, V. Guillet, J.-M. Masson, L. Mourey, and J.-P. Samama, "Electrostatic analysis of TEM1 β-lactamase: effect of substrate binding, steep potential gradients and consequences of site-directed mutations," *Structure,* vol. 3, pp. 603-613, 1995.

[156] A. Matagne and J. M. Frere, "Contribution of mutant analysis to the understanding of enzyme catalysis: the case of class A beta-lactamases," *Biochim Biophys Acta,* vol. 1246, pp. 109-27, Jan 19 1995.

[157] C. Damblon, X. Raquet, L. Y. Lian, J. Lamotte-Brasseur, E. Fonze, P. Charlier, G. C. Roberts, and J. M. Frère, "The catalytic mechanism of beta-lactamases: NMR titration of an active-site lysine residue of the TEM-1 enzyme," *Proceedings of the National Academy of Sciences,* vol. 93, pp. 1747-1752, March 5, 1996 1996.

[158] B. Joris, P. Ledent, O. Dideberg, E. Fonze, J. Lamotte-Brasseur, J. A. Kelly, J. M. Ghuysen, and J. M. Frere, "Comparison of the sequences of class A beta-lactamases and of the secondary structure elements of penicillin-recognizing proteins," *Antimicrob Agents Chemother,* vol. 35, pp. 2294-301, Nov 1991.

[159] J. M. Ghuysen, "Serine beta-lactamases and penicillin-binding proteins," *Annu Rev Microbiol,* vol. 45, pp. 37-67, 1991.

[160] B. Joris, J. M. Ghuysen, G. Dive, A. Renard, O. Dideberg, P. Charlier, J. M. Frere, J. A. Kelly, J. C. Boyington, P. C. Moews, and et al., "The active-site-serine penicillin-recognizing enzymes as members of the Streptomyces R61 DD-peptidase family," *Biochem J,* vol. 250, pp. 313-24, Mar 1 1988.

[161] F. Sanschagrin, F. Couture, and R. C. Levesque, "Primary structure of OXA-3 and phylogeny of oxacillin-hydrolyzing class D beta-lactamases," *Antimicrob Agents Chemother,* vol. 39, pp. 887-93, Apr 1995.

[162] H. Yi, K.-H. Cho, Y. S. Cho, K. Kim, W. C. Nierman, and H. S. Kim, "Twelve Positions in a β-Lactamase That Can Expand Its Substrate Spectrum with a Single Amino Acid Substitution," *PLoS ONE,* vol. 7, p. e37585, 2012.

[163] D. R. Livesay and D. J. Jacobs, "Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair," *Proteins,* vol. 62, pp. 130-43, Jan 1 2006.

[164] D. R. Livesay, D. H. Huynh, S. Dallakyan, and D. J. Jacobs, "Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family," *Chem Cent J,* vol. 2, p. 17, 2008.

[165] D. J. Jacobs, D. R. Livesay, J. Hules, and M. L. Tasayco, "Elucidating quantitative stability/flexibility relationships within thioredoxin and its fragments using a distance constraint model," *J Mol Biol,* vol. 358, pp. 882-904, May 5 2006.

[166] J. M. Mottonen, M. Xu, D. J. Jacobs, and D. R. Livesay, "Unifying mechanical and thermodynamic descriptions across the thioredoxin protein family," *Proteins,* vol. 75, pp. 610-27, May 15 2009.

[167] D. Verma, D. J. Jacobs, and D. R. Livesay, "Predicting the melting point of human C-type lysozyme mutants," *Curr Protein Pept Sci,* vol. 11, pp. 562-72, Nov 2010.

[168] D. Verma, D. J. Jacobs, and D. R. Livesay, "Changes in Lysozyme Flexibility upon Mutation Are Frequent, Large and Long-Ranged," *PLoS Comput Biol,* vol. 8, p. e1002409, 2012.

[169] P. Y. Savard and S. M. Gagne, "Backbone dynamics of TEM-1 determined by NMR: evidence for a highly ordered protein," *Biochemistry,* vol. 45, pp. 11414-24, Sep 26 2006.

[170] Kanlikili, x00E, P. er, x, E. O. lmez, Bu, x, deyri, x, N., and B. S. Akbulut, "Investigation of TEM-1 and SHV-1 beta-lactamase ligand binding," in *Health Informatics and Bioinformatics (HIBIT), 2010 5th International Symposium on*, 2010, pp. 167-172.

[171] F. Bos and J. Pleiss, "Multiple molecular dynamics simulations of TEM beta-lactamase: dynamics and water binding of the omega-loop," *Biophys J,* vol. 97, pp. 2550-8, Nov 4 2009.

[172] S. Banerjee, U. Pieper, G. Kapadia, L. K. Pannell, and O. Herzberg, "Role of the omega-loop in the activity, substrate specificity, and structure of class A beta-lactamase," *Biochemistry,* vol. 37, pp. 3286-96, Mar 10 1998.

[173] Y. Chen and B. K. Shoichet, "Molecular docking and ligand specificity in fragment-based inhibitor discovery," *Nat Chem Biol,* vol. 5, pp. 358-64, May 2009.

[174] I. Trehan, F. Morandi, L. C. Blaszczak, and B. K. Shoichet, "Using steric hindrance to design new inhibitors of class C beta-lactamases," *Chem Biol,* vol. 9, pp. 971-80, Sep 2002.

[175] J. M. Mottonen, D. J. Jacobs, and D. R. Livesay, "Allosteric response is both conserved and variable across three CheY orthologs," *Biophys J,* vol. 99, pp. 2245-54, Oct 6 2010.

[176] O. Fisette, S. Morin, P. Y. Savard, P. Lague, and S. M. Gagne, "TEM-1 backbone dynamics-insights from combined molecular dynamics and nuclear magnetic resonance," *Biophys J,* vol. 98, pp. 637-45, Feb 17 2010.

[177] K. Bush and G. A. Jacoby, "Updated functional classification of beta-lactamases," *Antimicrob Agents Chemother,* vol. 54, pp. 969-76, Mar 2010.

[178] P. C. Moews, J. R. Knox, O. Dideberg, P. Charlier, and J. M. Frere, "Beta-lactamase of Bacillus licheniformis 749/C at 2 A resolution," *Proteins,* vol. 7, pp. 156-71, 1990.

[179] O. Herzberg, "Refined crystal structure of beta-lactamase from Staphylococcus aureus PC1 at 2.0 A resolution," *J Mol Biol,* vol. 217, pp. 701-19, Feb 20 1991.

[180] E. Sauvage, E. Fonze, B. Quinting, M. Galleni, J. M. Frere, and P. Charlier, "Crystal structure of the Mycobacterium fortuitum class A beta-lactamase: structural basis for broad substrate specificity," *Antimicrob Agents Chemother,* vol. 50, pp. 2516-21, Jul 2006.

[181] A. P. Kuzin, M. Nukaga, Y. Nukaga, A. M. Hujer, R. A. Bonomo, and J. R. Knox, "Structure of the SHV-1 beta-lactamase," *Biochemistry,* vol. 38, pp. 5720-7, May 4 1999.

[182] M. Nukaga, K. Mayama, A. M. Hujer, R. A. Bonomo, and J. R. Knox, "Ultrahigh resolution structure of a class A beta-lactamase: on the mechanism and specificity of the extended-spectrum SHV-2 enzyme," *J Mol Biol,* vol. 328, pp. 289-301, Apr 18 2003.

[183] T. R. Walsh, A. P. MacGowan, and P. M. Bennett, "Sequence analysis and enzyme kinetics of the L2 serine beta-lactamase from Stenotrophomonas maltophilia," *Antimicrob Agents Chemother,* vol. 41, pp. 1460-4, Jul 1997.

[184] M. C. Orencia, J. S. Yoon, J. E. Ness, W. P. Stemmer, and R. C. Stevens, "Predicting the emergence of antibiotic resistance by directed evolution and structural analysis," *Nat Struct Biol,* vol. 8, pp. 238-42, Mar 2001.

[185] D. Lim, F. Sanschagrin, L. Passmore, L. De Castro, R. C. Levesque, and N. C. Strynadka, "Insights into the molecular basis for the carbenicillinase activity of PSE-4 beta-lactamase from crystallographic and kinetic studies," *Biochemistry,* vol. 40, pp. 395-402, Jan 16 2001.

[186] W. Sougakoff, G. L'Hermite, L. Pernot, T. Naas, V. Guillet, P. Nordmann, V. Jarlier, and J. Delettre, "Structure of the imipenem-hydrolyzing class A beta-lactamase SME-1 from Serratia marcescens," *Acta Crystallogr D Biol Crystallogr,* vol. 58, pp. 267-74, Feb 2002.

[187] P. Swaren, L. Maveyraud, X. Raquet, S. Cabantous, C. Duez, J. D. Pedelacq, S. Mariotte-Boyer, L. Mourey, R. Labia, M. H. Nicolas-Chanoine, P. Nordmann, J. M. Frere, and J. P. Samama, "X-ray analysis of the NMC-A beta-lactamase at 1.64-A resolution, a class A carbapenemase with broad substrate specificity," *J Biol Chem,* vol. 273, pp. 26714-21, Oct 9 1998.

[188] C. Jelsch, L. Mourey, J. M. Masson, and J. P. Samama, "Crystal structure of Escherichia coli TEM1 beta-lactamase at 1.8 A resolution," *Proteins,* vol. 16, pp. 364-83, Aug 1993.

[189] O. Dideberg, P. Charlier, J. P. Wery, P. Dehottay, J. Dusart, T. Erpicum, J. M. Frere, and J. M. Ghuysen, "The crystal structure of the beta-lactamase of Streptomyces albus G at 0.3 nm resolution," *Biochem J,* vol. 245, pp. 911-3, Aug 1 1987.

[190] P. Arriaga, M. Menendez, J. M. Villacorta, and J. Laynez, "Differential scanning calorimetric study of the thermal unfolding of .beta.-lactamase I from Bacillus cereus," *Biochemistry,* vol. 31, pp. 6603-6607, 1992/07/01 1992.

[191] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research,* vol. 25, pp. 3389-3402, September 1, 1997 1997.

[192] G. Wang and R. L. Dunbrack, "PISCES: a protein sequence culling server," *Bioinformatics,* vol. 19, pp. 1589-1591, August 12, 2003 2003.

[193] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research,* vol. 32, pp. 1792-1797, March 1, 2004 2004.

[194] K. Kelly. (1996). *Multiple sequence and structural alignment in MOE. Chemical Computing Group.* Available: http://www.chemcomp.com/Journal_of_CCG

[195] J. C. Gordon, J. B. Myers, T. Folta, V. Shoja, L. S. Heath, and A. Onufriev, "H++: a server for estimating pKas and adding missing hydrogens to macromolecules," *Nucleic Acids Res,* vol. 33, pp. W368-71, Jul 1 2005.

[196] J. Antosiewicz, J. A. McCammon, and M. K. Gilson, "Prediction of Ph-dependent Properties of Proteins," *Journal of Molecular Biology,* vol. 238, pp. 415-436, 1994.

[197] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, "Electrostatics of nanosystems: application to microtubules and the ribosome," *Proc Natl Acad Sci U S A,* vol. 98, pp. 10037-41, Aug 28 2001.

[198] W. T. Wong, K. C. Chan, P. K. So, H. K. Yap, W. H. Chung, Y. C. Leung, K. Y. Wong, and Y. Zhao, "Increased structural flexibility at the active site of a fluorophore-conjugated beta-lactamase distinctively impacts its binding toward diverse cephalosporin antibiotics," *J Biol Chem,* vol. 286, pp. 31771-80, Sep 9 2011.

[199] D. R. Livesay, D. H. Huynh, S. Dallakyan, and D. J. Jacobs, "Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family," *Chem Cent J,* vol. 2, p. 17, 2008.

[200] J. M. Mottonen, M. Xu, D. J. Jacobs, and D. R. Livesay, "Unifying mechanical and thermodynamic descriptions across the thioredoxin protein family," *Proteins,* vol. 75, pp. 610-27, May 15 2009.

[201] D. Verma, D. J. Jacobs, and D. R. Livesay, "Predicting the melting point of human C-type lysozyme mutants," *Curr Protein Pept Sci,* vol. 11, pp. 562-72, Nov 2010.

[202] D. Verma, D. J. Jacobs, and D. R. Livesay, "Changes in Lysozyme Flexibility upon Mutation Are Frequent, Large and Long-Ranged," *PLoS Comput Biol,* vol. 8, p. e1002409, 2012.

[203] R. Sanchez and A. Sali, "Evaluation of comparative protein structure modeling by MODELLER-3," *Proteins,* vol. Suppl 1, pp. 50-8, 1997.

[204] J. C. Gordon, J. B. Myers, T. Folta, V. Shoja, L. S. Heath, and A. Onufriev, "H++: a server for estimating pKas and adding missing hydrogens to macromolecules," *Nucleic Acids Res,* vol. 33, pp. W368-71, Jul 1 2005.

[205] D. J. Jacobs and S. Dallakyan, "Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity," *Biophys J,* vol. 88, pp. 903-15, Feb 2005.

[206] D. R. Livesay, S. Dallakyan, G. G. Wood, and D. J. Jacobs, "A flexible approach for understanding protein stability," *FEBS Lett,* vol. 576, pp. 468-76, Oct 22 2004.

[207] K. Takano, Y. Yamagata, S. Fujii, and K. Yutani, "Contribution of the hydrophobic effect to the stability of human lysozyme: calorimetric studies and X-ray structural analyses of the nine valine to alanine mutants," *Biochemistry,* vol. 36, pp. 688-98, Jan 28 1997.

[208] P. Gniewek, S. P. Leelananda, A. Kolinski, R. L. Jernigan, and A. Kloczkowski, "Multibody coarse-grained potentials for native structure recognition and quality assessment of protein models," *Proteins,* vol. 79, pp. 1923-9, Jun 2011.

[209] S. M. Gopal, K. Klenin, and W. Wenzel, "Template-free protein structure prediction and quality assessment with an all-atom free-energy model," *Proteins,* vol. 77, pp. 330-41, Nov 1 2009.

[210] J. Xu and Y. Zhang, "How significant is a protein structure similarity with TM-score = 0.5?," *Bioinformatics,* vol. 26, pp. 889-95, Apr 1 2010.

[211] P. Benkert, M. Kunzli, and T. Schwede, "QMEAN server for protein model quality estimation," *Nucleic Acids Res,* vol. 37, pp. W510-4, Jul 2009.

[212] P. Benkert, T. Schwede, and S. C. Tosatto, "QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information," *BMC Struct Biol,* vol. 9, p. 35, 2009.

[213] D. Cozzetto, A. Kryshtafovych, K. Fidelis, J. Moult, B. Rost, and A. Tramontano, "Evaluation of template-based models in CASP8 with standard measures," *Proteins,* vol. 77 Suppl 9, pp. 18-28, 2009.

VITA

EDUCATION

PhD Candidate - Bioinformatics and Computational Biology
Fall '07 – Fall '12
University of North Carolina at Charlotte, NC
Dissertation: Elucidating the effects of mutation and evolutionary divergence upon protein structure quantitative stability/flexibility relationships (*using Distance Constraint Model*).
GPA: 3.9/4.0

B. Tech. - Bioinformatics
Fall '03 – Spring '07
Jaypee University of IT, Solan HP, India
Thesis: 3D-QSAR, molecular docking and drug designing studies of HIV-I non-nucleoside reverse transcriptase inhibitors (NNRTIs).
GPA: 8.3/10.0

PUBLICATIONS

Changes in lysozyme dynamics upon mutation are frequent, large and long-ranged. Verma D, Jacobs DJ, Livesay DR (2012). *PLoS Computational Biology*, 8:e1002409.

WaveMap: Interactively discovering features from protein flexibility matrices using wavelet-based visual analytics. Barlow S, Liu Y, Yang J, Livesay DR, Jacobs DJ, Mottonen J, Verma D (2011). *Computer Graphics Forum*, 30:1001-1010.

Ensemble properties of network rigidity reveal allosteric mechanisms. Jacobs DJ, Livesay DR, Mottonen JM, Vorov OK, Istomin AY, Verma D (2011). In *Allostery: Methods and Protocols,* Fenton A (Ed.), Springer, ISBN: 978-94-007-0880-8.

Predicting the melting point of human c-type lysozyme mutants. Verma D, Jacobs DJ, Livesay DR (2010). *Current Protein and Peptide Science*, 11:562-572.

Docking, MM-GB/SA and ADME screening of HIV-1 NNRTI inhibitor: Nevirapine and its analogs. Naik PK, Verma D, Sengupta D (2008). *In-Silico Biology,* 8, 0023.

Comparative analysis of MMFF94x and AMBER99 using different protein class data sets: Aspartic proteases, serine proteases, metallo-proteases and sugar-binding proteins. Singh H, Marla S, Verma D (2007). *Online Journal of Bioinformatics*, 8(1): 45-55.

Docking mode of delvardine and its analogues into P66 domain of HIV-1 reverse transcriptase: Screening using MM-GB/SA and ADME screening. Naik PK, Verma D, Sengupta D (2007). *J Biosciences,* 32 (7):1307-16.

The binding modes, binding affinities and ADME screening of HIV-1 NNRTI inhibitor: Efavirnez and its analogs. Naik PK, Verma D, Sengupta D (2007). *Online Journal of Bioinformatics,* 8(1): 99-114.

Clustering of HIV-I subtype: Study of molecular diversity using phylogenetic analysis. Naik PK, Verma D, Sengupta D, Mishra VS (2006). *Bioinformatics Trends*, 19-26.

CONFERENCE PRESENTATIONS

Is rigidity conserved across the class A β-lactamase family? Verma D and Livesay DR, *Biophysical Society 56th Annual meeting*, 2012.

Elucidating the effects of mutation upon c-type lysozyme through quantitative stability/flexibility relationships. Verma D, Jacobs DJ and Livesay DR, *Biophysical Journal,* Volume 100, Issue 3, 2011, 400a.

Towards comprehensive analysis of protein family quantitative stability/flexibility relationships. Verma D, Jacobs DJ, Guo J and Livesay DR, *Biophysical Journal,* Volume 98, Issue 3, Supplement 1, 2010, 637a.

Predicting protein mutant stability with a combined

experimental/theoretical approach. Verma D, Jacobs DJ and Livesay DR, *Biophysical Journal,* Volume 96, Issue 3, Supplement 1, 2009, 301a-302a.

ONGOING (unpublished) PROJECTS

Developing thresholds on QSFR metrics for homology model quality assessment (*manuscript under preparation*).

Deciphering QSFR changes in MHC class II protein upon peptide binding.

SKILLS

Research Interests: Molecular Dynamics, Quantitative Stability/Flexibility Relationship, QSAR and *In-silico* Drug Designing,
Bioinformatics Software Skills: MOE package, Pymol, Phylip, TreeView, Clustal, MODELLER, BLAST, H++, ProtParam and other online Bioinformatics tools.
Wet Lab Skills: PCR & RAPD analysis, Agarose Gel Electrophoresis, SDS PAGE Electrophoresis, Quantitative Precipitant Assay Technique, Immunodiffusion Techniques – ODD, RID, Sandwich ELISA techniques, Qualitative & Quantitative estimation of proteins.
Programming: Java, C/C++ and PERL.
Other Tools: UNIX, HTML, R statistical Package, ChemOffice, Gnuplot, WaveMap and DCM.

OTHER WORK EXPERIENCE

Doctoral Candidate/Graduate Researcher - UNC Charlotte (Aug ’07 – Dec ‘12)
Teaching Assistant - Dept. of Bioinformatics and Genomics, UNC-Charlotte (Jan ’10 – May ’10)
Mentored undergraduate students in their projects – Responsibilities included preparing extensive self-learning tutorials and assisting with teaching of bioinformatics tools
Intern - Dade Behring (currently Siemens Healthcare) (Dec ’06 – Feb ’07)
Intern - Adroit Life Sciences (May ’05 – Jul ’05)

PROFESSIONAL SOCIETIES

Member, Biophysical Society (BPS) (2007-)
Member, International Society for Computational Biology (ISCB) (2011-)
Secretary, Bioinformatics Assembly of Students - UNC Charlotte (Aug ’10 – May ‘12)
Technical Coordinator, Bioinformatics Club - Jaypee University of IT (Aug ’05 – Jul ’06)

AWARDS

1st place in Poster Competition, 2012 CBES Molecular Engineering and Design Category
Graduate Assistant Support Plan Award – UNC Charlotte (Sep ‘07 – May ’12)
College of Computing and Informatics Student Travel Award – UNC Charlotte (2011)
The Center for Biomedical Engineering Systems Student Travel Award – UNC Charlotte (2012, 2011, 2010 & 2009)