The Pennsylvania State University

The Graduate School

College of Medicine

COMPUTATIONAL DESIGN AND EXPERIMENTAL CHARACTERIZATION OF PROTEINS WITH INCREASED STABILITY AND SOLUBILITY

A Dissertation in

Integrative Biosciences
by

Katrina L. Schweiker

© 2009 Katrina L. Schweiker

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2009

The dissertation of Katrina L. Schweiker was reviewed and approved* by the following:

George I. Makhatadze
Professor of Biochemistry & Molecular Biology
Dissertation Advisor
Chair of Committee

Judith S. Bond
Chair of Biochemistry & Molecular Biology

Thomas Spratt
Professor of Biochemistry & Molecular Biology
Chair, Chemical Biology Option of Integrative Biosciences Graduate Program

Philip C. Bevilacqua
Professor of Chemistry

Peter J. Hudson
Willaman Professor of Biology
Director of the Huck Institutes of the Life Sciences

*Signatures are on file in the Graduate School

ABSTRACT

The ability to design proteins from first principles will provide an efficient way to develop stabilized proteins, which could have a profound impact on a variety of biotechnological industries. For example, a biosensor made out of stable proteins would be able to be functional in harsh environmental conditions, such as the desert, where sensors made from less stable proteins would not be effective. Another example is that life-saving vaccines made from stable proteins could be stored at ambient temperatures, making it possible to distribute them more effectively to developing nations where refrigeration is not always an option. In addition to addressing the question of what forces govern thermodynamic stability, the field of protein design can also provide insight into the intramolecular interactions that are important for kinetic stability and solubility.

In the first part of this dissertation, the SH3 domain of the Fyn tyrosine kinase (FynSH3) was stabilized by the rational design of surface charge-charge interactions. Analysis of the computationally optimized distributions of surface charges showed that the increase in favorable energy per substitution begins to level off after five substitutions. One of the sequences with five substitutions (four charge reversals and one introduction of a new charge) was selected for experimental characterization. Nine additional variants were also characterized to explore the stepwise contributions of these substitutions to the stability of FynSH3. The designed sequence was found to have an increased thermostability of 12 °C and an increase in the free energy of unfolding (ΔG) of 8 kJ/mol, relative to the wild-type protein. These results suggest that a significant increase in stability can be achieved through a small number of amino acid substitutions, and argue for a seminal role of surface charge-charge interactions in modulating protein stability.

The second part of this dissertation addresses the question of how important the unfolded state of a protein is for determining its stability and whether it needs to be considered in the design approach. Some of the first attempts to address this issue tried to explain the pH-dependent changes in stability (ΔG) for several different proteins, where it was found that, in order to reproduce experimental data, a statistical (Gaussian polymer chain, GPC) representation of the unfolded state needed to be included in the calculations of ΔG. However, incorporation of this model into our design approach did not significantly improve our predictions. To determine whether this was due to an inability of the Gaussian model to accurately describe the distance distributions, and therefore the energies, observed in structural representations of the unfolded state, the distance distributions for a GPC were compared to those observed in the excluded volume limit (EV) structural libraries of two proteins: ubiquitin and NTL9. For residues that were close in sequence, where the unfolded state energies are the largest, it was found that these distributions were markedly different between the GPC and EV methods. A possible explanation for this observation is that the EV limit does not consider charge-charge interactions when creating the large-scale structural libraries. Molecular dynamics (MD) simulations were performed on the 2,000 structures in the EV libraries to model the unfolded state in the presence of charge-charge interactions, yet the Gaussian model was still unable to accurately reproduce the distance distributions of the structural library. However, very little difference was observed in the charge-charge interaction energies calculated by the Gaussian model versus directly calculating the energies in the post-MD unfolded state structural libraries, suggesting that the statistical model may be sufficient for describing the behavior of the unfolded state. Since (1) the overall charge-charge interaction energies in the unfolded state are small and (2) our design approach focuses on the differences in energies (ΔΔG) rather than absolute energies (ΔG) for selecting more stable variants, the overall effect of unfolded state can most likely be ignored without

adversely affecting the predictive ability of the algorithm. The implication of these results for a protein that has previously been thought to have specific residual interactions in the unfolded state is discussed.

In the third part of this dissertation, the question of how the thermodynamic stabilization of proteins redesigned by our approach affects the kinetics of the folding and unfolding reactions is addressed. The folding and unfolding kinetics of the wild-type and designed variants of a bacterial cold shock protein (CspB), FynSH3, tenascin, and procarboxypeptidase were examined. Since the hydrophobic collapse of the protein core is the first step in protein folding, the rate of hydrophobic collapse should drive the folding rate. All of the proteins designed in this study contain substitutions on the protein surface, while the core residues remain unchanged. Therefore, one would intuitively predict that the folding rates of the wild-type and designed variants of each protein should remain the same, so the observed increases in stability must come from much slower rates of unfolding. This is a logical conclusion because the designed proteins contain more favorable surface charge-charge interactions than the wild-type proteins, meaning that it would take more energy to break these favorable interactions once they had been formed, thus decreasing the unfolding rate. For CspB, this was indeed shown to be the case. However, the increased stability of the FynSH3, tenascin, and procarboxypeptidase variants appears to be due to a faster folding rate, while the unfolding rate remains unchanged relative to the wild-type. Based on φ-value analysis data from the literature, it appears that this affect is due to the substitutions being made at positions that have native-like structure in the transition state. The results of these experiments show that while proteins can be thermodynamically stabilized by the same method, the kinetic mechanisms of stabilization can be vastly different. By incorporating the results of existing φ-value analyses into the design algorithm, it should be possible to select for residues that would decrease the unfolding rate, rather than increase the folding rate. This

means that one could potentially design a protein that is not only thermostable, but also kinetically stable, which would have profound implications for the development of protein therapeutics.

The fourth part of the dissertation explores the role of surface charges in making proteins less susceptible to aggregation. A few recent reports suggest that adding a large number of charged moieties to proteins (supercharging) increases solubility and decreases aggregation due to thermal denaturation. While this approach seems to be an effective way to combat protein aggregation, nothing is known about the thermodynamic effects of supercharging. A supercharged variant of ubiquitin was designed by introducing charges at positions that were not predicted to have a significant impact on the thermodynamic stability. Not only was the supercharged variant of ubiquitin more soluble than the wild-type at neutral pH, but it also showed reversible thermal denaturation under conditions where wild-type ubiquitin aggregates. Interestingly, this protein was destabilized relative to the wild-type protein. While the supercharged ubiquitin was predicted to have similar thermodynamic stability to the wild-type, it is possible that our design approach cannot accurately predict charge-charge interaction energies in a highly charged molecule. Further studies on more supercharged proteins should help develop a foundation by which we can further understand the thermodynamic mechanisms, and therefore, more accurately predict, the effects of supercharging on protein both protein stability and protein aggregation.

In the fifth, and final, part of the dissertation, the effects of pressure on protein denaturation are examined. Pressure perturbation calorimetry (PPC) is a new experimental method that is being used to study the volumetric properties of proteins. PPC measures the coefficient of thermal expansion (α) of a protein in dilute solution when subjected to changes in pressure ($\Delta P \sim 80$ psi) under isothermal conditions. By measuring α as a function of temperature,

it is possible to measure the volumetric changes ($\Delta V/V$) in proteins upon unfolding. A novel method for analyzing the data using a thermodynamic two-state model of unfolding was developed, and was used to analyze PPC data for five model proteins: lysozyme, ribonuclease A, ubiquitin, cytochrome *c*, and eglinC. It was observed that the volumetric changes upon unfolding of all proteins, except cytochrome *c*, converged at high temperature. The anomalous behavior of cytochrome *c* is most likely due to the imperfect packing of the protein around the heme group. The results discussed in this chapter set a foundation for exploring how the alteration of intramolecular interactions such as packing interactions or surface charge-charge interactions will affect the volumetric properties of proteins.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

The following is a list of the abbreviations and symbols used throughout the manuscript:

The one and three letter codes for amino acids are used throughout according to the conventions set by IUPAC.

All units of measurements are also abbreviated according to IUPAC conventions.

< > – These brackets around a variable indicate an ensemble averaged value of the parameter

α – Thermal expansivity coefficient

Acp – Acylphosphatase

AMBER9 – Molecular dynamics software package

AMBER99SB – Parameterized force field for molecular dynamics solutions

ASA – Solvent accessible surface area

AUC – Analytical ultracentrifugation

CD – Circular dichroism spectroscopy

CI2 – Chymotrypsin inhibitor -2

CNBr – Cyanogen bromide; used to cleave His tags

CspB – Bacterial cold-shock protein B

CytC – Cytochrome *c*

ΔC_P – Change in heat capacity upon unfolding

ΔG – Gibbs free energy of unfolding; defines the thermodynamic stability of a protein

ΔG_{qq} – Gibbs free energy of charge-charge interactions

ΔH – Change in enthalpy upon unfolding

ΔH_{cal} – Calorimetrically measured change in enthalpy upon unfolding

ΔH_{vH} – van't Hoff change in enthalpy upon unfolding

ΔS – Change in entropy upon unfolding

ΔV – Change in partial volume of protein upon unfolding

d – root-mean-squared distance between charges

DSC – Differential scanning calorimetry

DTT – Dithiothreitol

ε – Dielectric constant

ε_{280nm} – Molar extinction coefficient at a wavelength of 280 nm; used to determine protein concentration

EDTA – Ethylenediaminetetraacetic acid

EgC – EglinC

E_{ij} – Pairwise energy of interactions between charges

EV – Excluded volume limit model of unfolded state

ExPASY – Expert protein analysis system proteomics server

F_N – Fraction of native protein

F_U – Fraction of unfolded protein

Fyn – SH3 domain of the Fyn tyrosine kinase

GA – Genetic algorithm

GB-SA – Generalized Born implicit solvent model, corrected for solvent accessibility

GFP – Green fluorescent protein

GPC – Gaussian polymer chain model of unfolded state

GST – Glutathione-S-transferase

HEWL – Hen egg-white lysozyme

I – Ionic strength of solvent

IPTG – Isopropyl-β-D-thiogalactopyranoside

K_{eq} – Equilibrium constant of unfolding

k_f – Folding rate

k_u – Unfolding rate

M – Molecular mass of a protein

MD – Molecular dynamics simulation

NLREG – Nonlinear regression software program

NTL9 – N-terminal domain of ribosomal L9 protein

OMTKY3 – Ovomucoid third domain protein

$P(E)$ – probability distribution of pairwise charge-charge interaction energies

$p(r)$ – distance distribution function

Pc – Procarboxypeptidase

PCR – Polymerase chain reaction

PDB – Protein data bank

PPC – Pressure perturbation calorimetry

q_i – Charge on residue i

R – Universal gas constant

RC – Random coil model of unfolded state

R_g – Radius of gyration

r_{ij} – Pairwise distances between atoms

RNaseA – Ribonuclease A

RNaseT1 – Ribonuclease T1 protein

SC – Supercharge

SCTKSA-GA – Computational design approach to rationally design supercharged protein variants

SDS-PAGE – SDS polyacrylimide gel electrophoresis

Ten – Tenascin

Tev – Tev protease; used to cleave His tags

TK-SA – Tanford-Kirkwood model, corrected for solvent accessibility; used to calculate the charge-charge interaction energies in a protein

TKSA-GA – Computational design approach that uses a genetic algorithm to select stable sequences based on their charge-charge interaction energies, calculated by the TK-SA model

T_m – Transition temperature; occurs where 50% of protein molecules in a sample are folded and 50% are unfolded; defines the thermostability of a protein

U1A – Ribosomal U1A protein

Ubq – Ubiquitin

$\overline{V}$ – Partial molar volume

$\overline{v}$ – Partial specific volume

WT – Wild-type

PREFACE

The work presented in Chapter 3 is the result of a collaboration with Dr. Arash Zarrine-Afsar and Dr. Alan R. Davidson, who cloned, expressed, and purified the FynSH3 domain variants. The DSC experiments and analysis were performed by Katrina Schweiker.

The work presented in Chapter 5 is from the combined efforts of Dr. Arash Zarrine-Afsar and Dr. Alan R. Davidson, who characterized the kinetics of the FynSH3 variants and Katrina Schweiker, who characterized the procarboxypeptidase and tenascin variants.

The work presented in Chapter 7 is from the combined efforts of Katrina Schweiker who characterized the effects of pressure on ubiquitin, cytochrome c, eglinC, Ubq-WT-CNBr, and Ubq-SC-Tev, and Victoria Fitz who characterized RNaseA and Lysozyme.

ACKNOWLEDGEMENTS

Figures 1.1 and 1.6 are reproduced from: Makhatadze GI, et al. (2004) Mechanism of thermostabilization in a designed cold shock protein with optimized electrostatic interactions *J Mol Biol* **336**:929-942, with permission from Elsevier.

Figures 1.3 and 1.5 are reproduced from: Strickler SS, et al. (2006) Protein stability and surface electrostatics: a charged relationship, *Biochemistry* **45**:2761-2766, with permission from the American Chemical Society.

Most of the text in Chapter 3 is reproduced from: Schweiker KL, et al. (2007) Computational design of the Fyn SH3 domain with increased stability through optimization of surface charge-charge interactions, *Protein Sci* **16**:2694-2702, with permission from The Protein Society.

The general introduction presented in Chapter 1 is a compilation of two review articles that are currently *in press*: Schweiker KL and Makhatadze GI (2009) Protein stabilization by the rational design of surface charge-charge interactions, *Meth Mol Biol* **490**:261-284; and Schweiker KL and Makhatadze GI. (2009) A computational approach for the rational design of stable proteins and enzymes: optimization of surface charge-charge interactions, *Meth Enymol* **454**

I am also indebted to many people for their roles, both professionally and personally, in my graduate education:

First, I would like to thank my advisor George Makhatadze, for guiding me on the exciting and challenging journey to becoming a critical thinker who always maintains a skeptical eye – even towards one's own work.

Next, I would like to thank the members of my committee: Dr. Judith Bond, Dr. Thomas Spratt, and Dr. Philip Bevilacqua for bringing different perspectives into the discussions of this work, which is so important for the development of good science.

I thank the members of the Makhatadze lab, specifically Marimar Lopez, Alexey Gribenko, Werner Streicher, and Mayank Patel for teaching me how to perform the careful, properly controlled experiments that are essential for testing the hypotheses presented in this thesis.

I would like to thank Alan Davidson and Arash Zarrine-Afsar for the fruitful collaboration on the study of the FynSH3 domain that is presented in Chapters 3 and 5.

The work presented in Chapter 4 on the role of the unfolded state would not have been possible without Angel Garcia and Ryan Day, who taught me how to run good simulations; and Rohit Pappu and Andreas Vitalis for providing the software necessary for generating the EV limit ensembles and for their helpful discussions about the results of this work.

I thank the US Air Force for giving me the opportunity to pursue higher education. Specifically, I need to thank Col Henry Gilman for being the one to encourage me to apply not only for graduate school, but for a PhD program, because he thought I could handle it. Also, thanks to Maj Brett

Hagen who pushed my paperwork through the sometimes slow-turning wheels of the AF bureaucracy so that I could take advantage of this opportunity.

I thank my mom, John, and Werner for their helpful comments on this work.

I would like to thank my friends and family for the moral support and encouragement that is the only way one can survive in life, and makes accomplishments like this so much sweeter.

Finally, I need to thank my parents for providing an environment where it was always cool to be smart and curious and question the nature of things. Even though my mom almost never directly answered a question, the fond memories I have growing mold in the refrigerator, growing plants in the dark, and grinding up spring and fall leaves with liquid nitrogen are even better, and provide a constant reminder to me of why I got into this business in the first place

For Kurt, who would've thought this was totally awesome

CHAPTER 1: GENERAL INTRODUCTION

1.1. Introduction to Thermodynamics

The ability to design proteins from first principles will provide an efficient way to develop stabilized proteins, which could have a profound impact on a variety of biotechnological industries. For example, a biosensor made out of stable proteins would be able to be functional in harsh environmental conditions, such as the desert, where sensors made from less stable proteins would not be effective. Another example is that life-saving vaccines made from stable proteins could be stored at ambient temperatures, making it possible to distribute them more effectively to developing nations where refrigeration is not always an option. In order to design or engineer proteins with increased stability it is necessary to have a fundamental understanding of the intramolecular forces that contribute to stabilizing the various conformations of proteins. The protein core is predominantly stabilized by the hydrophobic interactions between buried nonpolar side chains (*1-3*). Burial of polar residues in the core is unfavorable due to the high energetic cost of desolvation (*4, 5*). This energetic penalty can be offset by forming hydrogen bonds with other polar groups or buried water molecules. The core residues are further stabilized by van der Waals (packing) interactions (*4, 6-8*). Hydrogen bonding and packing interactions in the protein core have been demonstrated to be as important as hydrophobicity for stability (*8-10*). More recently it has been shown that surface residues can also modulate protein stability (*11-16*).

The term "protein stability" can have different meanings depending on the focus of the research being performed. Protein stability can refer to the change in Gibbs free energy upon unfolding (ΔG), thermostability (T_m), rates of folding or unfolding, *in vivo* degradation rates, or retention of activity after being exposed to harsh chemical or thermal conditions. The transition temperature and Gibbs free energy are measures of thermodynamic stability. They are

interrelated in such a way that it is possible to alter the stability (ΔG) of a protein without affecting the thermostability (*17*) and vice-versa (*15*) (Fig. 1.1). This is a result of the relationship between ΔG, T_m and the other thermodynamic parameters: enthalpy (ΔH), entropy (ΔS), and change in heat capacity upon unfolding (ΔC_P). The thermodynamic stability (ΔG) for a protein that unfolds via a two state transition, $N \leftrightarrow U$, can be described by the equilibrium constant, K_{eq}, which is the ratio of the fraction of unfolded protein (F_U) to the fraction of folded protein (F_N) in a sample.

$$\Delta G(T) = -RT \cdot \ln(K_{eq}) = -RT \cdot \ln\left(\frac{F_U}{F_N}\right) \qquad \textbf{(1.1)}$$

The change in thermodynamic stability (ΔG) at any temperature, T, can also be related to the enthalpy (ΔH) and entropy (ΔS) of unfolding.

$$\Delta G(T) = \Delta H(T) - T\Delta S(T) \qquad \textbf{(1.2)}$$

By assuming that the change in heat capacity upon unfolding (ΔC_P) is temperature independent, we can also relate ΔG to the thermostability (T_m) of the protein via the Gibbs-Helmholtz equation.

$$\Delta G(T) = \Delta H(T_m) + \Delta C_P \cdot (T - T_m) - T \cdot \left[\frac{\Delta H(T_m)}{T_m} + \Delta C_P \cdot ln\left(\frac{T}{T_m}\right)\right] \qquad \textbf{(1.3)}$$

where the transition temperature, T_m, is the temperature at which 50% of the protein molecules are unfolded, $\Delta H(T_m)$ is the enthalpy of unfolding at T_m, and ΔC_p is the change in heat capacity upon unfolding that characterizes the temperature dependence of both the enthalpy and entropy functions. The enthalpy and entropy at any temperature, T, are $\Delta H(T)=\Delta H(T_m)+\Delta C_p(T-T_m)$ and $\Delta S(T)=\Delta H(T_m)/T_m+\Delta C_p \cdot \ln(T/T_m)$, respectively.

The stability function defined by the Gibbs-Helmholtz equation (Eqs. 1.2 and 1.3) is a bell-shaped curve (Fig. 1.1 A), because the ΔC_p for protein unfolding is positive. The thermodynamic stability of a protein is equal to zero when 50% of the molecules are folded and

50% are unfolded. This occurs at two temperatures: T_m, the heat denaturation transition temperature and T_c, the cold denaturation transition temperature (*18*). The stability function has a maximum (ΔG_{max}) at the temperature where the entropic contribution is equal to zero (T_{max}). The changes ΔH, ΔS, and ΔC_p, in response to substitutions within the protein will also affect the temperature parameters and will define the thermodynamic mechanism by which changes in protein stability can be achieved.

Figure 1.1 shows three of the possible mechanisms, and the extreme versions of each have been modeled to illustrate the differences among them more clearly. However, one should remember that in practice, it is often more appropriate to explain experimental observations using combinations of these models (see (*19*)). If a protein is stabilized via the first mechanism, a large increase in both the maximum stability (ΔG_{max}) and thermostability (T_m) (Model 1, Fig. 1.1A) will be observed (*20, 21*). This is caused by a small decrease in the entropy function (Fig. 1.1B), while the enthalpy function (Fig. 1.1C) and the change in heat capacity upon unfolding are unchanged relative to the reference model. In the second model, a dramatic decrease in ΔC_P upon substitution creates a ΔG function with a shallower temperature dependence (*22, 23*). This results in an increase in the thermostability of the protein without affecting T_{max} or the absolute value of ΔG_{max} (Fig. 1.1A). The third model results in the entire stability function shifting to higher temperatures (*15*). This is caused by a large decrease in both the entropy and enthalpy functions, without changing their temperature dependencies (i.e. no change in ΔC_P). As a result, T_c, T_m, and T_{max} increase, while the absolute value of ΔG_{max} is not affected. In each of these three models, the stability of the protein at room temperature (ΔG_{RT}) is affected differently. In both the first and third mechanisms, ΔG_{RT} changes relative to the reference model – it increases if the protein is stabilized via Model 1, but decreases if the protein is stabilized by Model 3 (Fig 1.1A). The second model demonstrates that it is possible to increase the thermostability of a protein without

affecting ΔG_{RT}. With an understanding of the underlying thermodynamic mechanisms of stabilization, one should be able to design proteins to meet any desired thermodynamic criteria.

1.2 Protein Stabilization Approaches

The approaches to stabilizing proteins can be grouped into three major categories: directed evolution, sequence-based design, and computational design. Each has its own advantages and disadvantages that should be considered when deciding which to use for the design of stable proteins. A few of the factors to be considered include the amount of prior information required (i.e. sequence vs. 3D structure) to carry out the design, how quickly the result can be obtained, and the universal applicability of the method. A brief comparison of the three design categories in terms of these issues is provided below and summarized in Table 1.1.

1.2.1. Directed Evolution

Directed evolution uses random mutagenesis, targeted mutagenesis, or homologous recombination to introduce mutations into a gene of interest (*24, 25*). Random mutagenesis is the simplest approach, in the sense that it requires virtually no prior information about the protein. Combining error-prone PCR with screening and selection has been effective for altering the function (*26*), the stability (*27*), or both (*28*) of various proteins. Targeted mutagenesis is most effective for instances where it would be difficult to find the best mutations using random mutagenesis, such as significantly changing the function of a protein (*29*). In this case, it is necessary to have some structural or biochemical information about the protein so that mutagenesis can be directed to the appropriate active site residues. Homologous recombination between genes encoding proteins with a very high sequence identity can be used to introduce more diversity into the sequence library than is possible through random mutagenesis. It has been

used to create proteins with improved activity (*27, 28, 30, 31*), higher thermostability (*27, 28, 30, 32, 33*), or entirely new functions (*34*). Recombination has been demonstrated to be a successful approach not only when used alone (*30, 33, 34*), but also when applied in combination with targeted or random mutagenesis (*27, 28, 31, 32*).

Regardless of which directed evolution approaches are used, the first, and arguably most important, step is to create a sufficiently diverse library of sequences. Then selection pressure is applied to the library and it is screened for proteins that retain desired properties under the selected conditions. Examples of selection pressure include increasing temperature, antibiotic concentrations, or protease concentrations. Selection can also occur in a thermophilic host, which forces the protein to evolve in a biological context (*35*). The advantage of this type of selection is that the protein will not lose its natural function during the evolutionary process. Multiple rounds of mutation, screening, and selection are often necessary before the best protein variant can be identified.

Several different proteins have been stabilized using directed evolution (*35-38*). Subtilisins and p-nitrobenzyl esterase (PNE) were stabilized using random mutagenesis and then selecting for both stability at high temperatures and function at lower temperatures, with the result that thermostable variants maintained activity across a broader range of temperatures than naturally evolved enzymes (*28, 35*). These experiments suggested that stability and function are not mutually exclusive parameters. In the case of the subtilisin family of proteins, most of the stabilizing substitutions that occurred as a result of directed evolution were not found in the thermophilic proteins, and therefore would not have been selected using sequence-based approaches (*35*). One disadvantage of the method used to stabilize the subtilisins and PNE is that new functional assays had to be developed for each protein. A way to circumvent this requirement is to link selection directly to the ability of a protein to fold, rather than the ability to maintain activity (*36-44*) . The PROSIDE method (*36*), developed by Schmid and coworkers

does just that. It links the protease resistance of a protein to phage infectivity, and relies on the assumption that a stable protein will be more resistant to protease. Three different proteins, RNase T1 (*36*), CspB (*37, 43, 45*), and Gβ1 (*38, 44*) have been successfully stabilized using this method. Recently, a similar phage display approach was used to increase the thermostability of an antibody by 9 ºC (*42*).

Directed evolution is advantageous over computational design in the sense that no prior information about the protein structure is required. It is only necessary to know the protein sequence so it can be cloned appropriately and whether stability can be easily assayed. Moreover, as long as appropriate constraints are applied, stability and function can be enhanced simultaneously (*28*). The major disadvantage is that obtaining the final product can be slow because it takes time to construct libraries that are sufficiently diverse and to develop appropriate selection criteria and functional assays. In addition, important properties of the protein can be lost if they are not selected for directly (*46*). The screening process is also very labor-intensive, and often the most time consuming step (*35*). Another disadvantage is that due to the simultaneous introduction of multiple random substitutions, it is not possible to understand the mechanisms by which the protein was stabilized. In addition, there is no way of knowing whether all of the substitutions are important for stability without further study. As a result, it is difficult to learn more about why these particular substitutions were stabilizing for these proteins. Furthermore, directed evolution is not a universal approach because a different set of sequence libraries and selection criteria must be developed for each individual protein.

1.2.2. Sequence-Based Design

Sequence-based design refers to approaches that use the information contained in multiple sequence alignments to create more stable protein variants. The premise for these methods is that since the primary structure of a protein encodes all the information needed for

folding into the native tertiary structure, it also contains information about stability. In natural evolution, proteins tend to primarily be selected for function. In addition, the proteins need to be able to be easily degraded when they are no longer needed, so there is little evolutionary pressure for proteins to have high stabilities. As a result, the consensus sequence that can be obtained from a multiple sequence alignment is not always the most stable. More sophisticated statistical analyses, however, have made it possible to identify stabilizing properties from multiple sequence alignments (*47, 48*).

One of the sequence-based design approaches is based on the hypothesis that since, arguably, life originated in an extremely hot environment, the last common ancestor of all organisms is hyperthermophilic. Therefore, substituting a residue that was present in the last common ancestor into a modern protein should increase its thermostability (*49*). Since the ancestral residues are often also the consensus residue for a particular position, it raises the question: can the observed changes in stability be explained by the statistical free energy of the residue (the consensus approach), or are they due to the presence of an ancestral residue? To address this question, the enzyme 3-isopropylmalate dehydrogenase (IPMDH) was redesigned using phylogenetic analysis (*49*). The stabilities of twelve protein variants containing single site substitutions of amino acids to their ancestral residue were characterized. Eight of the ancestral residues were the same as the consensus and four were not. However, both categories had the same success rates -- half of the substitutions yielded protein variants that were more stable than the wild-type (*49*). These results suggest that stabilization by ancestral substitutions is not simply due to the statistical free energies of the residues.

Sequence-based approaches have also been used to design stable variants by making multiple substitutions simultaneously. In one example, two Bayesian statistical approaches were used to analyze a multiple sequence alignment of the subtilisin protein family. The first method, PROBE (*50*) identified a set of conserved domains that is characteristic to the protein family.

Then Classifier (*51*) was used to find a smaller subset of important residues based on specific sequence motifs. By coupling PROBE and Classifier, it was possible to identify a sequence motif that was present in some of the thermophilic subtilisins, but not the mesophilic proteins. To test whether this sixteen-residue motif was responsible for the increased stabilities of the thermophilic enzymes, the sequence was inserted into a mesophilic subtilisin, and the stability and activity of the variant were characterized. The variant had an increased thermostability of 13ºC relative to the wild-type enzyme, and was able to retain some activity at 90 ºC, a temperature where the wild-type subtilisin is completely inactive (*52*).

Sequence-based design methods are advantageous over computational design methods because no three-dimensional structure is required for design. They are also less time-consuming than directed evolution because diverse *in vivo* sequence libraries do not need to be developed and multiple rounds of selection do not need to be performed for each protein to be optimized. The successful redesign of the two different enzymes described above highlights the potential of sequence-based design to be a universal approach to protein stabilization. One of the disadvantages of sequence-based design is that the hypothesis that the ancestral protein is hyperthermophilic might not be correct for all proteins. In this case, the substitutions selected based on the ancestral protein sequence may not necessarily lead to increases in thermostability since the ancestral sequence is mesophilic. Another disadvantage of this approach is that the statistical analysis of multiple sequence alignments requires a large number of sequences. If a given protein family does not contain enough sequences to generate a statistically meaningful alignment, then it might not be possible to appropriately identify the ancestral gene. As a result, the selected substitutions might not actually be present in the ancestral protein sequence and, therefore, would not lead to increased thermostability.

1.2.3. Computational Design

Computational design refers to the stabilization of proteins by modeling the contributions of different intramolecular interactions from first principles. This approach is advantageous over directed evolution and sequence-based design methods because it has been demonstrated to be universal (*53*). Computational design, like the sequence-based design approaches, is faster than directed evolution because the energetic calculations can be performed more quickly than multiple rounds of screening and selection. It is also possible to qualitatively predict relative changes in the stabilities of proteins using computational design (*11, 15, 53*). One disadvantage of computational design is that three-dimensional structures are required to model the intramolecular interactions in the native state, so proteins that are not homologous to any known structures cannot be redesigned using computational methods. However, the advances in structural genomics projects are quickly nullifying this issue.

The contributions of hydrophobic interactions, hydrogen bonds, packing interactions, and charge-charge interactions to the stability of the native state of globular proteins have been extensively studied. The core of a globular protein typically contains a large number of nonpolar residues and is stabilized by the hydrophobic interactions between them (*1-3, 8*). The high energetic cost of desolvation of polar residues means that the burial of polar residues is usually very unfavorable (*4, 5*). However, this energetic penalty can be offset through the formation of hydrogen bonds with other polar groups or buried water molecules (*4, 5*). All buried residues are also stabilized by van der Waals (packing) interactions (*4, 6-8*). In fact, hydrogen bonding and packing interactions in the protein core have been demonstrated to be as important for protein stability as hydrophobic interactions (*8-10*).

The results of some early studies on the forces that govern protein stability suggested that residues on the surface of the protein do not provide significant contributions to stability. In one

example, a systematic set of mutations were made in T4 lysozyme, and it was observed that substitutions at many of the positions that were highly flexible and/or exposed to solvent did not have a significant affect on the stability of this protein (*3, 54*). Another example studied the interactions between charged surface residues in barnase (*55*), and found that most interactions between the solvent exposed charged residues had only weak contributions to the stability of the protein. These observations were explained by the idea that the residues on the surface of a protein are exposed to solvent in both the native and unfolded states, and as such their environments do not change significantly upon unfolding. Therefore the relative contributions of surface residues to ΔG would be smaller than for residues in the core. As a result of this hypothesis, the computational protein design field began to focus on optimizing interactions in the protein core (*56-62*). However, the attempts to stabilize proteins by redesigning the protein core have had mixed success (*6, 7, 61, 62*).

In general, core substitutions that fill cavities will enhance packing interactions and are therefore stabilizing (*61, 62*), while core substitutions that create cavities, and thus decrease the packing interactions are destabilizing (*6-8, 17, 62*). However, one should proceed with caution when making cavity filling substitutions because large, hydrophobic residues can also be destabilizing due to steric clashes within a tightly packed protein core. *De novo* attempts to redesign the protein core demonstrate how difficult it can be to model which substitutions will be stabilizing and which will be destabilizing (*6, 7, 56*). One explanation is that the core of the protein is very tightly packed, suggesting that the intramolecular interactions within the protein core are already optimized. In order to further improve the interactions within the core, one would need extremely precise modeling of the positions of the side chains. Another issue was that most early core redesign methods were modeling interactions using a fixed backbone (*6, 61, 63*). While this assumption was necessary to minimize the search space, and reduce computation time, it has been demonstrated that the backbone does indeed shift to accommodate substitutions

within the protein core (*62, 64*). However, one of the early attempts to incorporate backbone flexibility into a core design algorithm did not show significant improvements over fixed backbone methods in predicting the effects of core substitutions on protein stability (*65*).

One alternative to the computational redesign of the protein core is to focus on redesigning interactions on the protein surface. Although it had been argued that surface residues were not important for stability, support for this idea comes from a few sources. For example, when the differences between mesophilic and thermophilic proteins from the same family were examined, it was observed that the differences in stability appear to come primarily from an increase in electrostatic interactions (*15, 66-70*), which are more likely to be found on the protein surface than in the core. More evidence was provided by attempts to stabilize proteins using directed evolution, where it was observed that the stabilizing substitutions are often found on the surface of the protein (*27, 38, 43, 44, 71*). The idea that surface residues can be important for stability was also supported by a recent theoretical study on the physical origin of stability (*72*). It was suggested that as a response to evolutionary pressure, mesophilic proteins can evolve high thermostability by increasing the number of charged residues (*72*), and charged residues are much more likely to be found on the surface than in the core of the protein. Finally, experimental observations have shown that, in some cases, the interactions between charged surface residues have relatively large contributions to stability (*73-75*). A number of recent studies have exploited this information and shown that it is possible to modulate the stability of a number of proteins through altering the charge-charge interactions on the protein surface (*11-16, 76-78*). Indeed, the optimization of interactions on the protein surface can provide similar increases in stability to that obtained through the optimization of the protein core (Table 1.2).

One of the advantages of redesigning the surface of a protein is that the surface residues have greater conformational flexibility than those in the core. As a result, the modeling of surface side chains does not have to be as precise as core side chain modeling in order to obtain a good

description of the energetics of interactions. In addition, the flexibility of the surface side chains means that they are generally more tolerant to substitutions than residues in the core. For these reasons, the optimization of surface interactions can be considered a viable alternative approach to stabilizing proteins. The remainder of this chapter will discuss the optimization of surface charge-charge interactions and highlight some of the important experimental verifications of this method.

1.3. Rational Design of Surface Charge-Charge Interactions

The computational approach that will be described here is the rational design of surface charge-charge interactions. The first step to stabilizing proteins by this method is to calculate the pairwise charge-charge interaction energies in the wild-type protein. Second, the interaction energies are optimized using a genetic algorithm. The genetic algorithm will identify many sequences which are predicted to have increased stabilities relative to the wild-type. Since it is not possible to experimentally test all of these sequences, only a few are selected for characterization. Structures for the selected sequences are created using homology modeling, and the charge-charge interaction energies are calculated for the designed variants to better understand the details of how the substitutions are predicted to affect the stability. Finally, the stabilities of the selected sequences are characterized experimentally.

1.3.1. Calculating pairwise charge-charge interaction energies

The interaction energies between pairs of charges are calculated using the Tanford-Kirkwood model, corrected for solvent accessibility (TK-SA) (*79-82*). In the TK-SA model, the protein is represented by a low dielectric sphere that is impenetrable to solvent (Fig. 1.2). The charged groups in the protein are represented by point charges that occupy fixed positions in the

protein sphere. It is assumed that the interactions between charges are the only type of interaction between the groups (*79*). The energy of the charge-charge interactions between two residues on the protein surface, *i* and *j* is:

$$E_{ij} = e^2\left(\frac{A_{ij} - B_{ij}}{2b} - \frac{C_{ij}}{2a}\right)\cdot\left(1 - SA_{ij}\right) \qquad \textbf{(1.4)}$$

where *e* is the unit charge; *b* is the radius of the protein sphere and is related to the specific volume of the protein; and *a* is the radius of the ion exclusion boundary. The terms A_{ij}, B_{ij}, and C_{ij} have been previously defined by Tanford and Kirkwood (*79*). A_{ij} represents the energy between the charges in the low dielectric environment of the protein interior and is a function of the protein dielectric constant ($\varepsilon_P = 4$ for the interior of the protein) and the distance between the charges, r_{ij}. B_{ij} reflects the contributions from both the low dielectric environment of the protein the high dielectric environment of the solvent that surrounds the protein, and is a function of the solvent dielectric constant ($\varepsilon_S = 78.5$ for water), protein dielectric constant, ε_P, and the relative positions of the charges on the protein surface, which are defined by r_i, r_j, and θ_{ij}. C_{ij} is a function of the ionic strength of the solvent and the positions of the charges. The average solvent accessibility of residues *i* and *j* is represented by the term SA_{ij} and is calculated by the method of Richmond as previously described (*83, 84*).

The contribution of these pairwise charge-charge interaction energies to the Gibbs free energy of unfolding of the protein is determined from the pKa shifts of the ionizable residues in the protein relative to model compounds. At a given pH, the charges on the ionizable residues can be represented by a protonation state, χ. The energy of this protonation state for the folded protein molecule is:

$$\Delta G_N(\chi) = -RT\cdot(\ln 10)\sum_{i=1}^{n}\left(q_i + x_i\right)\cdot pK_{\text{int},i} + \frac{1}{2}\sum_{i,j=1}^{n}E_{ij}\left(q_i + x_i\right)\cdot\left(q_j + x_j\right) \qquad \textbf{(1.5)}$$

where R is the universal gas constant; T is the temperature in Kelvin (298K is the standard temperature for these calculations); x_i and x_j represent the protonation of groups i and j and will have a value of 0 or 1; q_i and q_j are the charges of groups i and j in the unprotonated state and have a value of -1 or 0; and $pK_{int,i}$ is the intrinsic pKa of group i if all other groups in the protein have zero charge. In this approach, the intrinsic pKa values for the ionizable groups of proteins are determined from model compounds and are: Asp = 4.0; Glu = 4.5; His = 6.3; Lys = 10.6; Arg = 12.0; N-ter = 7.7; and C-ter = 3.6.

There have been several reports that specific interactions between charged residues occur in the unfolded state and need to be considered to accurately predict the thermodynamic stability of proteins (*85-94*). However, the contributions of these interactions are small (~2 kJ/mol for a pKa shift of 0.4 units compared to ~20 kJ/mol for the total ΔG of unfolding for a protein) and are expected to be even smaller when comparing the unfolded state contributions to $\Delta\Delta G$ (the difference in stability between the wild-type and designed protein variants). Therefore, we assume that there are no residual charge-charge interactions in the unfolded state of the protein, and as such the energy of the protonation state, χ, in the unfolded state is:

$$\Delta G_U(\chi) = -RT \cdot (\ln 10) \sum_{i=1}^{n} (q_i + x_i) \cdot pK_{\text{int},i} \qquad \textbf{(1.6)}$$

These energy functions can then be used to define partition functions for the native (Z_N) and unfolded (Z_U) states of the protein:

$$Z_N = \sum_{\chi} \exp\left(-\frac{\Delta G_N(\chi)}{RT} - \nu(x) \cdot (\ln 10) \cdot pH \right) \qquad \textbf{(1.7)}$$

$$Z_U = \sum_{\chi} \exp\left(-\frac{\Delta G_U(\chi)}{RT} - \nu(x) \cdot (\ln 10) \cdot pH \right) \qquad \textbf{(1.8)}$$

where $\nu(\chi)$ is the number of protonated ionizable groups in the χ protonation state. By using the neutral forms of the native ($\Delta G_N(\chi) = 0$) and unfolded ($\Delta G_U(\chi) = 0$) states of the protein as

reference states for both Z_N and Z_U, the overall contribution of the charge-charge interactions to the Gibbs free energy of unfolding can be described as:

$$\Delta G_{qq} = -RT \cdot \ln\left(\frac{Z_U}{Z_N}\right) \quad \textbf{(1.9)}$$

In order to calculate the charge-charge interaction energies, one must know the distances between the charged residues, which can be determined from a high-resolution three-dimensional structural representation of the protein obtained through either X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR). Structures obtained from X-ray crystallography represent static snapshots of one possible configuration of the positions of side chains in a protein molecule. However, in solution, the surface side chains will have certain degree of conformational freedom which could alter the relative positions of the ionizable groups compared to the crystal structure. To account for the flexibility of the side chains, an ensemble of structures is created by homology modeling using Modeller software package (*95*). On the other hand, NMR structure determination experiments are performed in solution, and thus, the conformational flexibility of surface side chains is to some degree already accounted for in an ensemble of NMR structures, so no homology modeling is needed in these cases. Once the structural ensemble has been generated through homology modeling or obtained from NMR experiments, the TK-SA calculations are performed on each individual structure, and the results are then averaged over the entire ensemble. The flexibility of surface residues also makes it possible to use homology modeling to generate structural representations of proteins whose structures have not yet been solved, provided that they have a high degree of sequence similarity to known structures. Indeed, this is the approach that was used to redesign human acylphosphatase (*16*).

Figure 1.3A shows the results of the TK-SA calculations for ubiquitin (*16*). The value of ΔG_{qq} for a given residue represents the total energy of the interactions between that residue and

every other charged residue in the protein. Unfavorable interactions are represented by positive values of ΔG_{qq}, while favorable interactions are indicated by negative values of ΔG_{qq}. Note that ubiquitin has several charged residues which participate in unfavorable interactions, and therefore provide unfavorable contributions to stability. This general trend has been observed in all proteins redesigned by this approach so far (*11, 15, 16, 67, 68, 96-98*), and leads to the idea that it should be possible to increase the stability of these proteins by neutralizing or reversing the charges of the residues that participate in unfavorable interactions. In particular, one would expect that the reversal of an existing charge that participates in unfavorable interactions should yield greater increases in stability than neutralization. In fact, this behavior has been observed experimentally (*11-13*).

1.3.2. Optimization of surface charges using the Genetic Algorithm

While it is possible to make single or double substitutions and observe a significant increase in the stability of a given protein (*11, 16, 96, 98*), those variants are often not representative of the most favorable charge distribution for the protein. The ideal approach for identifying the optimal charge distribution for a protein, given its sequence, would be to use an exhaustive search algorithm to calculate the energy of every possible ionization state. However, this approach is computationally prohibitive for all but the smallest of peptides. For a protein with n charged positions on the surface, there are three possible charged states at each position (-1, 0, +1), and therefore, it would take 3^n calculations to identify all possible charge distributions. Even for a relatively small protein, such as ubiquitin, with only 23 charged surface residues, $3^{23} \approx 10^{11}$ calculations would need to be performed. Assuming one processor can perform 100 TK-SA calculations per second, it would take over 31 years to perform the exhaustive search! An excellent alternative to exhaustive calculations is the genetic algorithm (GA) (*16, 99, 100*). The genetic algorithm is faster than exhaustive calculations because it does not seek to find each and

every one of the best charge distributions, but instead identifies some of the sequences that are among the most optimal. For a protein like ubiquitin, only around $5x10^4$ calculations are required to identify some of the best sequences using the genetic algorithm. Once again, assuming that one processor can perform 100 TK-SA calculations per second, it would only take a little over eight minutes to identify optimized sequences using the genetic algorithm. This significant reduction in computation time makes it possible for the optimization by a genetic algorithm to be performed on a standard desktop PC.

Although the genetic algorithm and its implementation have been described in great detail elsewhere (*16, 99, 100*), a conceptual overview of how it works is important for understanding the computational design approach discussed here (see Fig. 1.4). In our implementation of the genetic algorithm, only residues with greater than 50% solvent accessibility are included in this optimization. The surface charge distributions available to a protein, given its sequence, are represented *in silico* by a "chromosome." The elements of these "chromosomes" are the charged states of the amino acid residues on the surface of the protein. An initial population of "chromosomes" is generated that contains a certain number of wild-type charge distributions and a certain number of randomly generated distributions (Fig. 1.4A). The "chromosomes" are scored based on their total charge-charge interaction energies, which are calculated by the TK-SA model described in the previous section. The lowest energy "chromosomes" are kept for the next generation where "crossover" events are used to finish populating the *n+1* generation (Fig. 1.4B). Once this generation has been populated, "point mutations" are used to introduce more diversity into the population (Fig. 1.4C). An energetic penalty helps to minimize the number of energetically neutral or weak "mutations" and makes the "crossover" events essential for proper sampling of the available charge distributions. This process is repeated iteratively until the lowest energy "chromosomes" have remained identical for a predetermined number of cycles (Fig. 1.4D).

The results of the genetic algorithm for ubiquitin are shown in Figure 1.5 and serves to demonstrate the effectiveness of the genetic algorithm to appropriately sample the entire sequence space available to a given protein (*16*). By examining the charge-charge interaction energies as a function of the number of amino acid substitutions relative to the wild-type sequence, one can see that, in general, an increasing number of substitutions leads to a significant increase in favorable charge-charge interactions. However, after a certain number of substitutions (eight to ten for ubiquitin), the increase in favorable energy that is obtained per additional substitution begins to level off. The observation that a large increase in stability can be obtained with a small number of substitutions also holds for other proteins (*97*), suggesting that it is possible to increase protein stability via the optimization of surface charge-charge interactions with only a few substitutions.

1.4 Experimental Verification of Computational Predictions

One of the most important facets of computational design is to experimentally test the predictions. The experimental characterization of the stabilities of the designed variants serves two important purposes. First, it is the only way to know if the physical model being used to make the predictions is appropriate or if it is lacking in some of the fundamental aspects. Second, only by testing this approach on a number of proteins with different sizes, secondary structures, and three-dimensional topologies can one determine how universal this approach is and what improvements, if any, should be made. This section will highlight the results of some of the key experiments that validated the TK-SA computational design method.

1.4.1. Single site substitutions

The first test of the hypothesis that optimizing surface charge-charge interactions will increase the stability of a protein was performed using ubiquitin as a model system (*11*). In this test, three

single site substitutions were made to neutralize the charges at positions predicted to contribute unfavorably to stability (K6Q, H68Q, and R72Q), three single substitutions were made to reverse charges at unfavorable positions (K6E, R42E, and H68E), and three single substitutions were made to neutralize charges at favorable positions (K27Q, K29Q, and K29N) to serve as controls for these experiments. The stabilities of these nine ubiquitin variants were measured by monitoring changes in secondary structure as a function of denaturant concentration using far-UV circular dichroism spectroscopy (CD) (*101*). As predicted, the neutralization of unfavorable charges was stabilizing. Furthermore, the reversal of the unfavorable charges resulted in larger increases in stability (~1 kJ/mol) than charge neutralization. In addition, the neutralization of charges predicted to contribute favorably to the stability of ubiquitin (K27Q, K29Q, and K29N) resulted in variants with significantly decreased stability relative to the wild-type, suggesting that the TK-SA model can qualitatively predict the effects of surface substitutions on the stability of ubiquitin.

One of the advantages of computational design methods over other approaches used to stabilize proteins, such as directed evolution or sequence-based design, is that it is universal. In other words, since computational design approaches model the energetics of intramolecular interactions, one should be able to use the same algorithm to redesign many different proteins without developing different selection criteria for each protein. After it had been demonstrated that the TK-SA approach could be used to successfully stabilize one model protein, the next important step was to test the robustness of this model. This was done using several proteins, and the results of each test are described in this section.

Initially, the calculated values of ΔG_{qq} were compared to the experimental stabilities reported in the literature for ubiquitin (*11*), the bacterial cold-shock protein (CspB) (*14, 102*), RNaseSA (*12*), the peripheral subunit binding domain (psbd41) (*13*), rubredoxin (*103*), barnase (*55, 74, 75, 104*), λ-repressor (*105*), T4 lysozyme (*106*), the B1 domain of protein G (GB1) (*107,*

108), and the zinc-finger domain (*109*). The changes in both the experimentally measured thermostabilites (ΔT_m) and stabilities ($\Delta\Delta G_{exp}$) of the variants relative to their wild-type proteins were compared to the changes in the charge-charge interaction energy ($\Delta\Delta G_{qq}$) expected from the substitutions (*67, 68*). It was observed in all cases that the changes in both thermostability and stability for these proteins could be qualitatively predicted based on the calculated changes in $\Delta\Delta G_{qq}$ (*67, 68*). These results provided the first indication for the robustness of the TK-SA design strategy. More extensive testing was performed by making many substitutions in three different model systems: α-lactalbumin (*96*), ribosomal protein L30e (*98*), and bacterial cold shock protein (CspB) (*15, 110*).

α-Lactalbumin is a small calcium binding protein that has recently been observed to bind electrostatically to highly basic proteins and histones. The *apo* form of the protein was predicted to have many unfavorable surface charge-charge interactions (*96*). While the presence of calcium does create favorable interactions for the residues involved in metal binding, the residues far from the binding loop maintained unfavorable interactions. However, the TK-SA approach was able to successfully predict the effects of the single site substitutions on the stability of α-lactalbumin. It was also observed that the changes in the thermostability of α-lactalbumin are in direct correlation with the changes in the calcium affinity (*96*).

In order to learn more about the extent to which surface charge-charge interactions affect stability, the *T. celer* ribosomal protein L30e was used as a model system (*98*). In this study, the TK-SA model was used to predict the effects of charge to alanine substitutions at all 26 charged positions of this protein. In addition to eliminating the charges at these positions, the alanine substitutions alter other important intramolecular interactions, such as hydrophobicity, secondary structure propensity, and side chain packing interactions. If these other types of interactions contribute more to stability than charge-charge interactions at these positions, then one would expect the calculated values of $\Delta\Delta G_{qq}$ to incorrectly predict the experimentally observed changes

in stability. However, the experimentally measured changes in stability were predicted correctly for 20 of the 26 positions studied. The remaining six positions were all located at either the N- or C-termini of α-helices, and thus, are likely to participate in specific interactions at the ends of the helix, and it has been shown previously that the identity of helix-capping residues is very important for thermodynamic stability (*111-119*). The results of the L30e experiments suggest that the non-electrostatic interactions that are important for the helix-capping motifs contribute more to stability than the charge-charge interactions at these positions.

The bacterial cold shock protein, CspB, was used as a model system to gain a better understanding of the possible thermodynamic mechanism of stabilization through the rational design of surface charges. The surface charge-charge interaction energies were calculated and compared for the CspB proteins from the mesophilic bacterium *B. subtilis* (CspB-Bs), the thermophilic *B. caldolyticus* (CspB-Bc), and the hyperthermophilic *T. maritima* (CspB-Tm) (*15, 67*). Although the sequences of these three variants of CspB are highly homologous, the distributions of the surface charges are very different (Fig. 1.6). CspB-Bs has the greatest number of unfavorable charge-charge interactions (Fig. 1.6A), whereas CspB-Tm has the most favorable charge-charge interactions (Fig. 1.6C). This trend correlates with the relative thermostabilities of these proteins. To determine whether the high stability of CspB-Tm did indeed come from the increased number of favorable surface charge-charge interactions, a cold shock protein (CspB-TB) was engineered to have the same core residues as CspB-Bs and the same surface charge distribution of CspB-Tm (Fig. 1.6D). The thermal stabilities of CspB-Bs and CspB-TB were measured using far-UV CD and it was found that CspB-TB had an increase in thermostability of 20°C relative to the CspB-Bs protein (*15*). This result further supported the idea that the rational design of surface charge-charge interactions could be a more effective way to stabilize proteins than making single substitutions at unfavorable positions.

Because CspB-Bs and CspB-TB are structurally similar, yet have dramatically different surface charge distributions, they provided a special opportunity to address two important questions regarding protein design. The first was whether or not charge-charge interactions in the unfolded state provide significant contributions to stability. To answer this question, a number of single substitutions were made in each of the proteins (*110*). For most of the substitutions, the TK-SA approach was able to semi-quantitatively (relative rank order) predict the effects of the substitutions on the stability of each protein. Since most of the positions that were predicted incorrectly were located in a β-hairpin of CspB-Bs, it is possible that residual charge-charge interactions in the unfolded state could affect the overall contributions of these residues to the stability of the native state. However, when the Gaussian chain model of the unfolded state (*93*) was incorporated into the calculations, no significant improvement in the correlation between the calculated and experimental stabilities was observed. Furthermore, when the putative unfolded state structure of CspB-Bs was disrupted, by destabilizing the β2-β3 hairpin, there was also no significant improvement between the calculations and experiments (*110*). Rather, it was found that these residues were part of a complex network of charge-charge interactions that when disrupted, led to markedly better agreement between the calculated and experimentally measured stabilities (*110*).

At first, the observation that the unfolded state of CspB did not have a significant effect on the predictions of the TK-SA model seemed to be in conflict with previous observations that consideration of the unfolded state was necessary to accurately predict experimentally measured stabilities of a number of different proteins (*85-94*). In most of these examples, it seems that there were specific non-electrostatic interactions in the unfolded state of the proteins that affected the predictions of the thermodynamic stability (ΔG). However, the TK-SA approach predicts $\Delta\Delta G_{qq}$, or the difference in charge-charge interaction energies between a wild-type protein and its designed variant. In the absence of specific interactions in the unfolded state, the high dielectric

of the solvent, to which most of the protein is exposed, will screen out interactions between charges separated by more than ten residues. Even for residues that are much closer in sequence, the calculated charge-charge interaction energies in the unfolded state are smaller than for interactions in the native state of the protein. Since the unfolded state contributions of non-specific interactions are small, even in terms of ΔG_{qq}, they would be expected to have even smaller contributions to $\Delta\Delta G_{qq}$. This idea, combined with the CspB results, would suggest that including the unfolded state charge-charge interactions is not always necessary to improve the correlation between the predicted and experimental stabilities.

The second question that the CspB proteins were uniquely suited to address is whether the surface substitutions affect interactions other than the charge-charge interactions. Ideally, to analyze only the effects of altered charge-charge interactions on protein stability, one would make substitutions that perturbed only the charge of the side chain, without affecting other factors such as size, hydrophobicity, and/or packing interactions (*13*). Incorporating non-natural amino acids into the protein sequence is an effective way to accomplish this goal (*13*), but it is only experimentally possible for small proteins. Furthermore, the relatively small number of naturally occurring amino acids offers only limited options for reversing or neutralizing charges in proteins. As such, it is often easier to use the natural amino acids lysine or glutamic acid for charge reversals and glutamine for charge neutralizations. Although this approach simplifies the design process, it is possible that observed changes in stability are actually due to changes in other important properties such as hydrogen bonding patterns, hydrophobicity, secondary structure propensity, or packing.

It is also possible that the charge reversals/neutralizations could alter short-range (i.e. salt bridges) rather than long-range charge-charge interactions. Short-range and long-range interactions are affected differently by changes in the ionic strength of a solution – long-range interactions tend to get weaker with increasing salt concentrations, whereas short-range

interactions tend to persist (*120-122*) – making it possible to determine which interactions contribute more to the observed increases in stability. To address this issue, the same substitutions were made at the same surface positions in the different electrostatic environments of CspB-Bs and CspB-TB. If the substitutions affect primarily long-range charge-charge interactions, then one would expect an inverse correlation between changes in stability and changes in the halophilicity of the proteins (*14, 41, 66*). Indeed, for most of the substitutions, this behavior was observed (*110*). The surface substitutions that did not display an inverse correlation between stability and halophilicity occurred at the same position, V20 in CspB-Bs and K20 in CspB-TB. The introduction of charge at V20 in CspB-Bs results in a protein that is both less stable and less halophilic, while the introduction of a hydrophobic residue at K20 in CspB-Tb results in a protein with increased stability and halophilicity, suggesting that hydrophobic interactions are much more important than charge-charge interactions at this position (*110*).

1.4.2. Rational design of surface charge-charge interactions using a genetic algorithm

The studies on the proteins described above provided strong evidence that the rational design of surface charge-charge interactions could be successful for many different proteins. Furthermore, the studies with the engineered CspB-TB protein not only gave important insights into the nature of how proteins are stabilized by this approach, but also led to the idea that it should be possible to computationally optimize the entire surface charge distribution for any given protein. To optimize the surface charge-charge interactions, a genetic algorithm (GA) was used to simultaneously select multiple sites for substitution. The TKSA-GA approach to stabilize proteins was tested using seven model proteins: ubiquitin, procarboxypeptidase, the Fyn SH3 domain, acylphosphatase, tenascin, U1A, and CDC42. These proteins all have different sizes, three-dimensional topologies, secondary structural composition, and surface charge distributions (*16, 97, 110*) (Fig. 1.7). The surface charge-charge interaction energies were calculated for each

protein using the TK-SA model, and then the optimal surface charge distributions were identified using the genetic algorithm. The results discussed in this section are summarized in Table 1.3.

One of the first proteins redesigned by this approach was ubiquitin (Fig. 1.7A). In this study, the stabilities of two variants with single substitutions at unfavorable positions (Ubq-6, K6E, and Ubq-72, R74E) and one variant with charge reversals at both positions (Ubq-6/72, K6E/R74E) were characterized as a reference for the magnitude of ΔΔG expected when charge-charge interactions were optimized using the genetic algorithm. It was observed that both Ubq-6 and Ubq-72 were more stable than wild-type ubiquitin ($\Delta\Delta G_{Des\text{-}WT}$ = 3.3 kJ/mol and 1.7 kJ/mol, respectively), and Ubq-6/72 was more stable than either single variant ($\Delta\Delta G_{Des\text{-}WT}$ = 5.2 kJ/mol) (*16*). Once again, this demonstrates that reversing the charges at unfavorable positions can lead to significant increases in stability. The next step was to see if the optimization of surface charges using the genetic algorithm would provide even larger increases in stability. Figure 1.5 shows an analysis of the results of the genetic algorithm for ubiquitin. In general, there is an increase in favorable charge-charge interaction energy with an increasing number of substitutions. However, after eight to ten substitutions, the increase in favorable charge-charge interaction energy gained per additional substitution begins to level off (Fig. 1.5B). As a result, it should be possible to obtain significant increases in stability with just a few mutations.

Three of the sequences that were predicted to increase the stability of ubiquitin were selected for further characterization. Figure 1.3A provides a comparison of the results of the TK-SA calculations for wild-type and designed variants of ubiquitin. One of the variants only optimized existing charges (Ubq-GA2 – four substitutions, 5.3% of the sequence), while the other two variants also allowed for neutral polar residues on the surface to be included in the optimization (Ubq-GA1 – five substitutions, 6.6% of the sequence; Ubq-GA3 – 6 substitutions, 7.9% of the sequence). From Figure 1.3A, it can be seen that several positions that have unfavorable contributions to stability in the wild-type protein are now predicted to contribute

favorably in each of the designed variants. Indeed, when the stabilities of these three designed sequences were characterized using urea-induced unfolding, it was found that, not only were they much more stable than the wild-type ($\Delta\Delta G_{UbqGA1\text{-}WT}$ = 13.2, $\Delta\Delta G_{UbqGA2\text{-}WT}$ = 18.4, and $\Delta\Delta G_{UbqGA3\text{-}WT}$ = 17.7 kJ/mol), but all three variants also had much larger increases in stability than one obtains by focusing only on one or two unfavorable positions (*16*).

Two of the sequences of procarboxypeptidase (Fig. 1.7B) that were predicted to be more stable than the wild-type were also selected for experimental characterization. One sequence contained five substitutions (6.9% of the sequence) and one contained seven substitutions (9.7% of the sequence). Both designed sequences had significantly increased stabilities relative to the wild-type protein ($\Delta\Delta G_{Des\text{-}WT}$ = 4.1 and 10.7 kJ/mol) (*16*). Two designed variants of acylphosphatase (Fig. 1.7C) were also studied. The first variant (Acp-GA1) contained four substitutions (4.1% of the sequence) and was stabilized by 7.0 kJ/mol relative to the wild-type (*16*). The second variant (Acp-GA2) contained 5 substitutions (5.1% of the sequence) and was stabilized by 11 °C relative to the wild-type (*123*).

Four sequences of the FynSH3 domain (Fig. 1.7D) were selected for characterization to understand the step-wise effects of the optimization of charge-charge interactions on protein stability (*97*). One of the variants contained five substitutions (Fyn5; 8% of the sequence) and the others contained one (Fyn1; 1.6% of the sequence), two (Fyn2; 3.2% of the sequence), or three (Fyn3; 4.8% of the sequence) of those five substitutions in their sequences. Each of the four variants was more stable than the wild-type ($\Delta\Delta G_{Fyn1}$ = 4.7 kJ/mol, $\Delta\Delta G_{Fyn2}$ = 2.3 kJ/mol, $\Delta\Delta G_{Fyn3}$ = 6.7 kJ/mol, $\Delta\Delta G_{Fyn5}$ = 7.1 kJ/mol) (*97*). Importantly, the TK-SA model was able to predict the relative rank order of the stabilities of these variants. The ubiquitin, procarboxypeptidase, acylphosphatase, and FynSH3 results demonstrate how the flexibility of this approach allows for large increases in stability to be obtained with two or three different designed sequences. This was an important observation because it suggested that it should be possible to choose sequences

that do not make substitutions in or near the active/binding site of proteins, thus providing a way to increase their stability without significantly affecting their function.

Only one optimized sequence each of tenascin (Fig. 1.7E) and U1A (Fig. 1.7F) were selected for further characterization (*16*). The designed sequence for each of these proteins had four substitutions (~4% of each sequence). Once again, for each of these proteins, the designed variants were significantly more stable than the wild-type ($\Delta\Delta G_{Des\text{-}WT}$ = 5.4 and 4.1 kJ/mol, for tenascin and U1A, respectively). Moreover, this was the first successful stabilization of the U1A ribosomal protein. The largest protein redesigned by this approach was CDC42 (190 amino acids, Fig. 7G). With only eight amino acid substitutions (4.2% of the sequence), it was possible to thermostabilize CDC42 by 10 °C (*123*), which is quite remarkable for such a large protein. These results highlight how the rational design of surface charge-charge interactions is a universal approach for stabilizing proteins of different sizes and structures.

1.4.3. Effects of stabilization on enzymatic activity

In order for any design approach to be useful for practical applications, the protein must retain its activity. To determine if this was indeed true for proteins which were stabilized by optimizing the surface charges, activity assays were performed on three of the proteins: CspB, acylphosphatase, and CDC42. The activities of each designed variant were compared to their respective wild-type proteins.

CspB-Bs is expressed by *B. subtilis* when it is exposed to cold temperatures and protects cells from these conditions by acting as an RNA chaperone. CspB can also bind polypyrimidine single-stranded DNA (ssDNA) sequences (*124-126*). The interactions between both CspB-Bs and CspB-TB (23°C more thermostable than CspB-Bs) with ssDNA templates were measured using fluorescence spectroscopy. Not only could CspB-TB bind ssDNA at higher temperatures (37°C) than CspB-Bs, but it also bound ssDNA with a higher affinity than CspB-Bs at lower

temperatures (25°C) (*15*). Based on the findings of these functional studies, the structure of CspB in complex with pT7 was solved (*127, 128*).

Acylphosphatase is an enzyme that binds its charged substrate, acylphosphate, and catalyzes the hydrolysis to produce carboxylate and inorganic phosphate. The hydrolysis of benzoylphosphate by the acylphosphatase variants was measured using a continuous UV absorption assay (*129, 130*). The Acp-GA1 variant was found to be inactive, and examination of its sequence revealed that one of the stabilizing substitutions was in the active site of the enzyme. As a result, a sequence was selected that contained only substitutions that were distant from the active site (Acp-GA2). Not only was the designed Acp-GA2 enzyme 10 °C more stable than Acp-WT, but was also able to maintain similar catalytic activity at room temperature ($k_{cat,WT}$ = (1.0 ± 0.2) x 10^{-4} s^{-1}; $k_{cat,GA2}$ = (2.3 ± 0.2) x 10^{-4} s^{-1}; $K_{M,WT}$ = (1.0 ± 0.1) x 10^{3} M; $K_{M,GA2}$ = (0.9 ± 0.1) x 10^{3} M) (*123*).

CDC42 is an important cell signaling protein that binds GTP and catalyzes the hydrolysis of GTP to GDP + P_i. This activity can be monitored using a colorimetric assay to detect the amount of free phosphate released upon hydrolysis. Since the thermal inactivation/denaturation of CDC42 is not reversible, the functional properties were characterized as the residual activity after incubation at high temperatures (~10 °C higher than the $T_{1/2}$ of wild-type CDC42). Once again, it was observed that the designed variant was not only able to retain activity at temperatures where the wild-type enzyme was inactivated, but it was also as active as the wild-type at room temperature ($k_{cat,WT}$ = (2.8 ± 0.1) hr^{-1}; $k_{cat,GA1}$ = (2.6 ± 0.1) hr^{-1}; $k_{cat,GA2}$ = (3.1 ± 0.2) hr^{-1}; $K_{M,WT}$ = (8.2 ± 0.9) x 10^{-3} M; $K_{M,GA1}$ = (8.2 ± 0.9) x 10^{-3} M; $K_{M,GA2}$ = (8.2 ± 0.9) x 10^{-3} M) (*123*).

The functional studies of the three proteins discussed in this section support two important conclusions. First, it is possible to stabilize proteins through the rational design of surface charge-charge interactions without perturbing activity. Second, it provides a strong

argument against the idea that a dichotomy exists between protein stability and function. Historically, it was believed that if a protein was thermostable, then it must be rigid at lower temperatures. Since proteins and enzymes need some flexibility to function properly, an idea developed that if a protein were stabilized, it would become more rigid, and therefore, less active at lower temperatures. This idea was supported by observations that proteins isolated from thermophilic organisms were not as active at lower temperatures as they were at higher temperatures (*28, 131-134*). However, proteins that exist in thermophilic organisms are under no evolutionary pressure to function at decreased temperatures. This does not mean *per se* that stability and flexibility/function should be mutually exclusive. Indeed, the results presented here show that by optimizing surface charge interactions at regions of the protein that are far from the active site, it is possible to increase stability and maintain activity at both higher and lower temperatures.

1.5 Practical Considerations

As evidenced in this chapter, the TK-SA method provides a simple model that is still effective for determining the qualitative changes in charge-charge interactions on the protein surface. However, there are a few assumptions that go into the model that must be taken into consideration when it is applied to the rational design of stable proteins. First, the model assumes that the protein is spherical. For globular proteins, this assumption does not appear to have adverse affects, even when the shape of the protein somewhat deviates from a sphere. For example, tenascin (Fig. 1.7D) is more cylindrical than spherical, but this approach was still able to successfully predict stabilizing substitutions (*16*). Second, the model assumes that the interactions between charges are the only electrostatic interactions that occur in the native state. This assumption could pose the biggest challenge for accurate predictions since it is known that

hydrogen bonding and partial dipoles also provide significant contributions to electrostatic interaction energies. To ameliorate this issue, surface side chains that are involved in intramolecular hydrogen bonds are not included in the optimization procedure described here. Third, it has been reported that this model ignores important parameters such as self-energy and solvation (*135-137*). While this is true, when the results of the TK-SA calculations for surface residues are compared to the results of calculations on surface residues using other continuum electrostatic models such as the finite difference solution of the Poisson-Boltzmann equation (FDPB/UHBD) (Fig. 3B) (*138, 139*), the Multi-Conformer Continuum Electrostatic model (MCCE) (Fig. 3C) (*140, 141*), the Microenvironment Modulated Screened Coulomb Potential model (MM_SCP) (Fig. 3D) (*142*), or the Langevin dipole model (PDLD) (*135, 143, 144*), they are qualitatively similar for all models. The only advantage of TK-SA over the continuum models is that it is less computationally demanding.

Another potential pitfall to the rational design of surface charges is that it is currently not possible to quantitatively predict the protein stabilities. This is largely due to the fact that only interactions between charges in the native state of the protein are being considered. While it has been shown that the unfolded state effects do not seem to be significant for the proteins redesigned by this approach, unfolded state effects have been demonstrated to be important for other proteins (*87-89, 145*), and we have not considered them in any of the designs described here. In addition, other important factors for protein stability, such as side chain hydrophobicity, secondary structure propensity, hydrogen bonding, packing interactions, and helix capping interactions are not considered (*4, 5, 8, 9, 64, 78, 116, 117, 146-148*). Nevertheless, the TK-SA model does provide excellent qualitative predictions of protein stability. In order to obtain quantitative predictions, the other factors mentioned here will need to be included in the computational optimization approach. Important questions that will need to be addressed in the development of a quantitative algorithm are: which of these interactions are the most important

for modulating stability, and how quantitative should the algorithm be to be practical? It seems likely that incorporating just a few of the factors mentioned here will give the algorithm the ability to predict the stability of designed sequences within the errors of experimental techniques.

Table 1.1 Comparison of the different approaches used to design/engineer stable proteins

	Directed Evolution	**Sequence-Based Design**	**Computational Design**
3D Structure	No	No	Yes
Speed	Slow	Fast	Fast
Labor Intensive	Yes	No	No
Guaranteed Results	Maybe	No	No
Universal	No	Maybe	Yes
Mechanism of stabilization	Kinetic Thermodynamic	Kinetic Thermodynamic	Thermodynamic

Table 1.2 Comparison of Different Computational Design Approaches

Design Method	Protein Name	# Mutations (Total Residues)	ΔT_m and/or $\Delta\Delta G$	Location of Substitutions
Rosetta	Procarboxypeptidase (*60*)	48 (71)	30 kJ/mol	Surface & Core
ORBIT	Engrailed Homeodomain (*149*)	24 (51)	33°C	Surface & Core
ORBIT	Thioredoxin (*150*)	3 (104)	10 kJ/mol	Core
Rosetta	Yeast cytosine deaminase (*61*)	3 (158)	10°C	Core
Rosetta	Procarboxypeptidase (*151*)	4 (71)	16kJ/mol	Surface
TK-SA	Fyn, Ubq, U1A, Procarb, Acp, CDC42, Ten (*53, 152*)	4 - 7 (72 - 190)	4 - 12ºC 4 - 18 kJ/mol	Surface
Poisson-Boltzmann	psbd41 (*13*)	1 (41)	9-12°C 3 kJ/mol	Surface
Altered Coulombic interactions	RnaseT1, RnaseSa (*12*)	1 (96-104)	2-7°C 2-5 kJ/mol	Surface
Sequence-based design	CspB-Bs (*14*)	2 (67)	13-21°C 8.8-14 kJ/mol	Surface

A comparison of different protein design approaches. The first two rows of the table highlight instances where the Rosetta and ORBIT algorithms were used to increase the stability of proteins by much more than what has been demonstrated to be possible by stabilizing proteins using only surface interactions. However, in both of these instances, over half of the protein sequence was subjected to substitution, resulting in optimization of both core and surface interactions. When Rosetta and ORBIT were used to make only a small number of substitutions in the protein core, similar increases in both thermodynamic stability (ΔG) and thermostability (T_m) relative to the surface redesign approaches were observed. The results highlighted in this table demonstrate that surface interactions can be as important as interactions in the protein core for modulating the stability of proteins.

Table 1.3 Summary of TK-SA/GA Results

Protein Name	Number of Substitutions (% of Sequence)	ΔΔG or T_m
Ubiquitin		
GA1	5/76 (6.6%)	13.2 kJ/mol
GA2	4/76 (5.3%)	18.4 kJ/mol
GA3	6/76 (7.9%)	17.7 kJ/mol
Procarboxypeptidase		
GA1	5/72 (6.9%)	4.1 kJ/mol
GA2	7/72 (9.7%)	10.7 kJ/mol
Acylphosphatase		
GA1	4/98 (4.1%)	7 kJ/mol
GA2	5/98 (5.1%)	11°C
Fyn SH3 Domain		
Fyn1	1/62 (1.6%)	4.7 kJ/mol
Fyn2	2/62 (3.2%)	2.3 kJ/mol
Fyn3	3/62 (4.8%)	6.7 kJ/mol
Fyn5	5/62 (8%)	7.1 kJ/mol
Tenascin		
GA1	4/90 (4.4%)	5.4 kJ/mol
U1A Protein		
GA1	4/100 (4%)	4.1 kJ/mol
CDC42		
GA1	8/190 (4.2%)	10°C

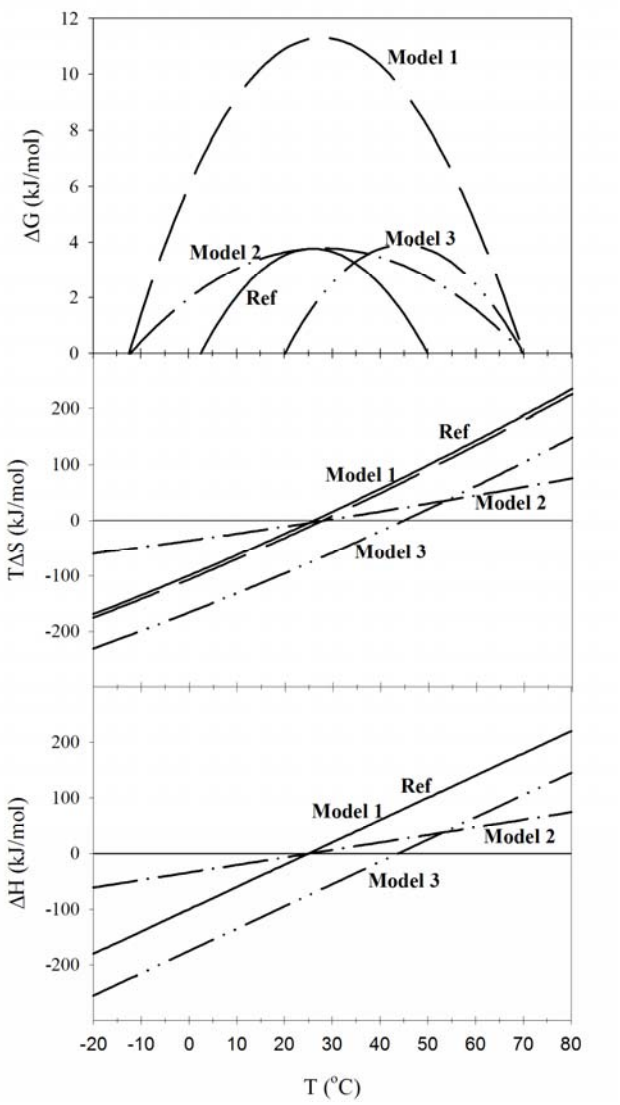

Figure 1.1 Three thermodynamic mechanisms of thermostabilization, reprinted from (*15*) with permission from Elsevier. To highlight the differences more clearly, these model functions represent extreme examples of each mechanism of stabilization. In each panel, different thermodynamic models are represented by the following lines: solid -- Reference Model; dashed -- Model 1; dash-dot-dashed -- Model 2; dash-dot-dot-dashed -- Model 3. **A.** The Gibbs free energy (ΔG) as a function of temperature. **B.** The temperature dependence of the entropic term ($T\Delta S$). **C.** The temperature dependence of the enthalpy function (ΔH).

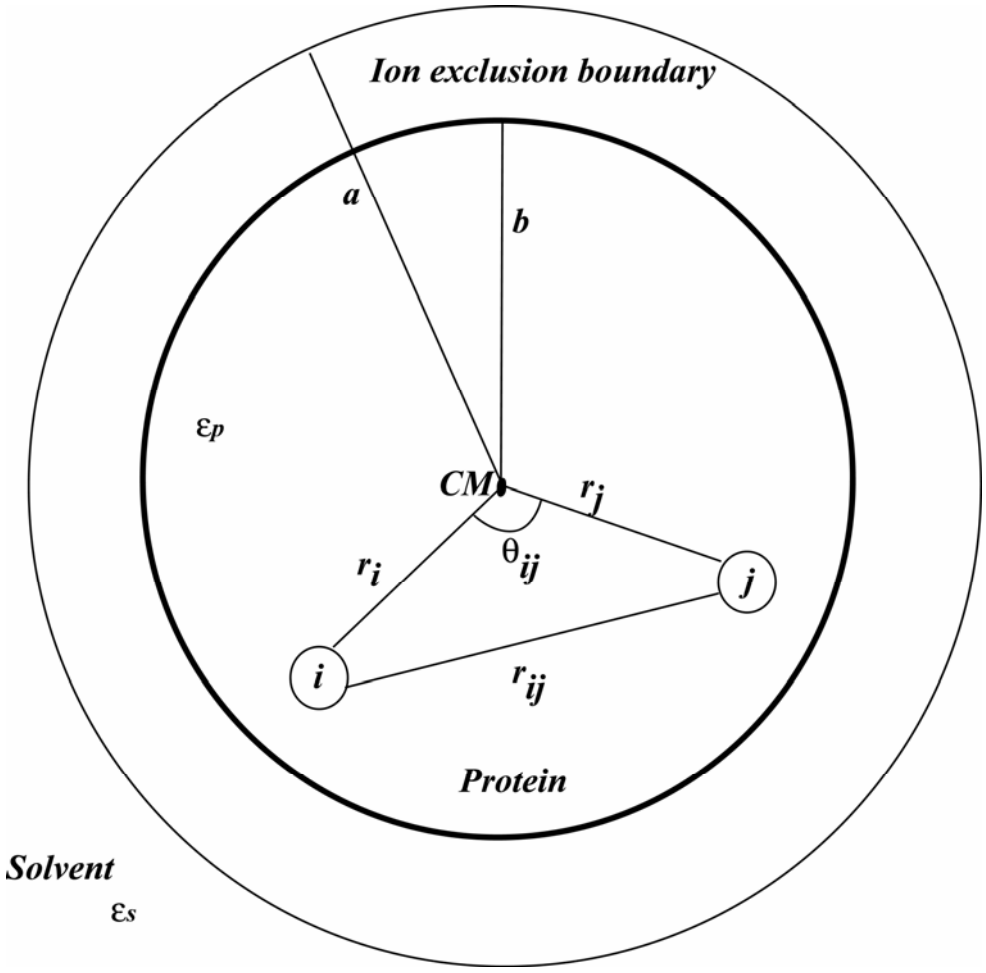

Figure 1.2 Schematic representation of the Tanford-Kirkwood model of the interactions between charged residues. The protein is represented by a hard sphere with low dielectric (ε_P) of radius *b* from the center of mass (*CM*). It is surrounded by a larger sphere of radius *a*, which is the ion exclusion boundary. These spheres are immersed in water, represented by high dielectric (ε_S). The other parameters in the TK-SA model are the identity of charges on the protein surface, represented by the small spheres *i* and *j* and the distance between them, r_{ij}.

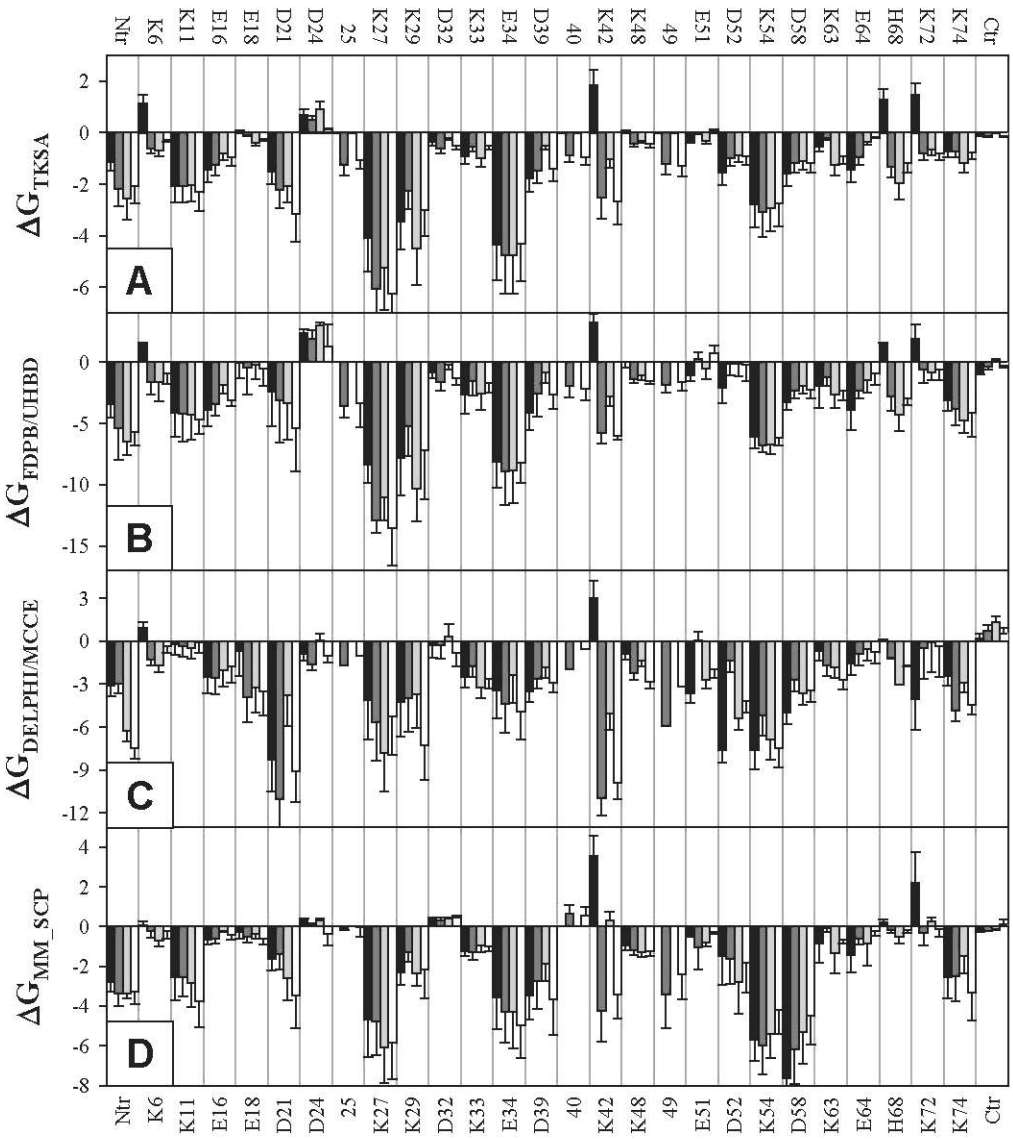

Figure 1.3 Surface charge-charge interaction energies (ΔG_{qq}) for wild-type and designed variants of ubiquitin at pH 5.5, reprinted with permission from (*53*). Copyright 2006 by the American Chemical Society. The charge-charge interactions were calculated using four different models. **A.** TK-SA **B.** FDPB-UHBD **C.** MMCE **D.** MM_SCP. Each bar represents the total energy of charge-charge interactions of the corresponding residue with every other residue in the protein, averaged over an ensemble of 11 structures. Positive values of ΔG_{qq} are indicative of unfavorable interactions, while negative values correspond to favorable interactions. Black bars -- wild-type ubiquitin. Dark grey bars -- UBQ-GA#1; light grey bars -- UBQ-GA#2; and white bars -- UBQ-GA#3 represent designed sequences that were identified by the genetic algorithm. UBQ-GA#1 and UBQ-GA#3 included uncharged polar residues in the optimization, while UBQ-GA#2 did not.

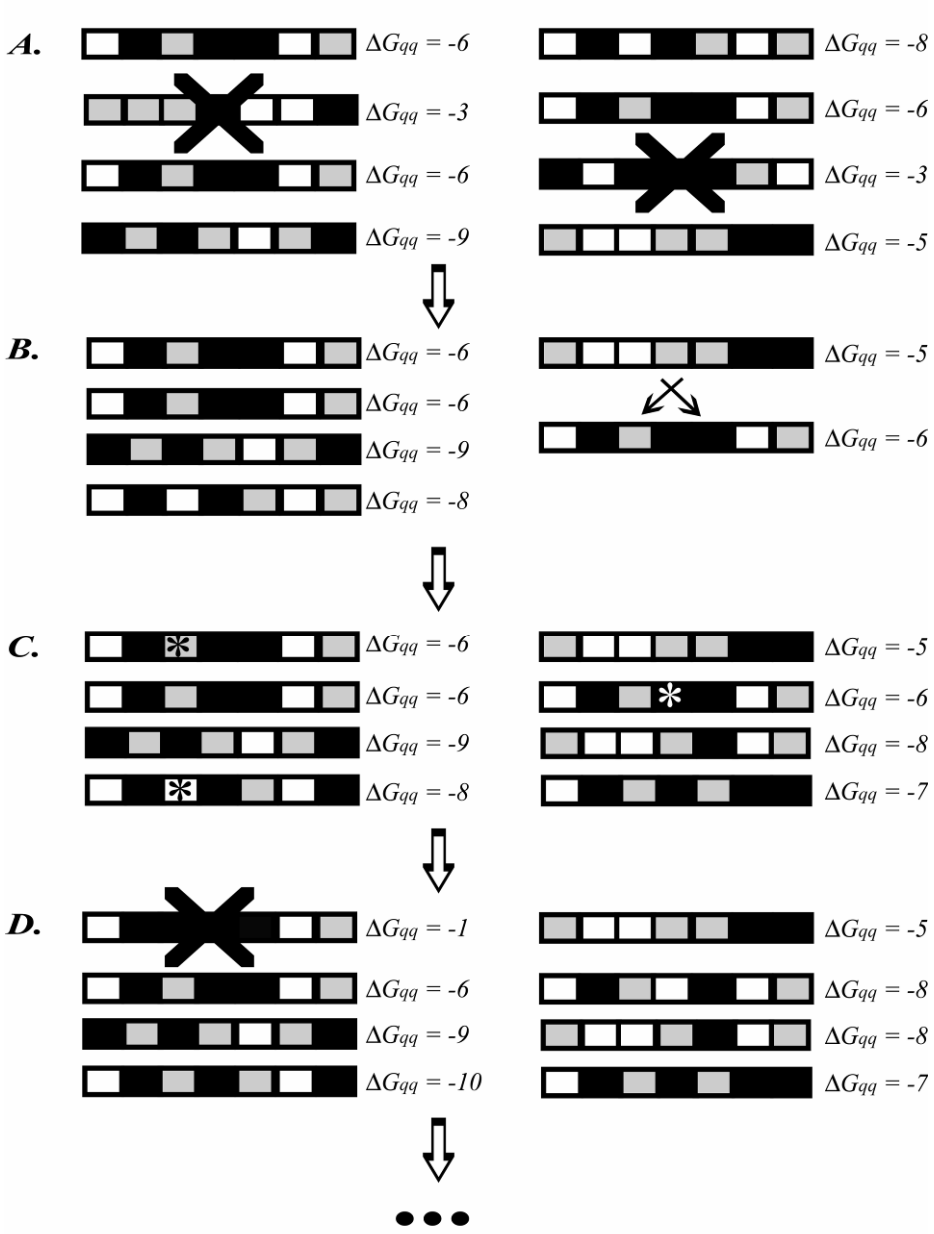

Figure 1.4 A schematic representation of the genetic algorithm. The steps of the algorithm are described in detail in the text. In each "chromosome," black boxes are representative of positive charges, grey boxes of negative charges, and white boxes indicate neutral residues. The large "X" in (**A** and **D**) indicates sequences whose energies were above the cutoff (-5 kJ/mol in this example), and were therefore not kept for the *n+1* generation. The black arrows in (**B**) show the "crossover" events that are used to finish populating the *n+1* generation. The stars in (**C**) indicate "point mutations" that are used to introduce more diversity into the population. The chromosomes in (**D**) that have energies below the pre-determined cutoff will be kept for the next generation, and these steps will be repeated iteratively until the sequences in population of "chromosomes" have reached convergence.

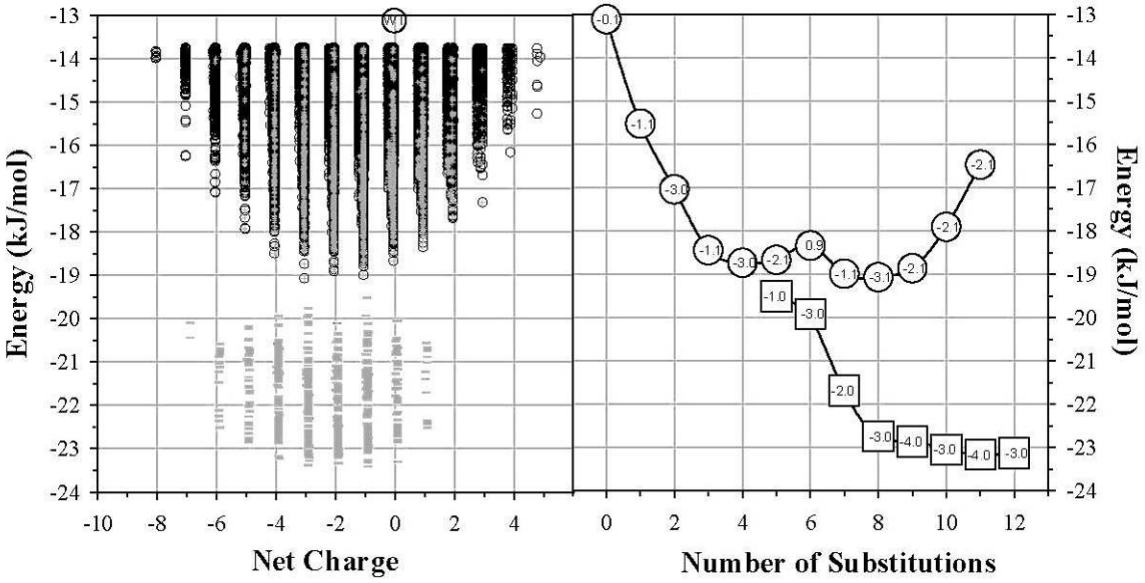

Figure 1.5 Analysis of the ability of the genetic algorithm to find the optimal sequence of ubiquitin, reprinted with permission from (*53*). Copyright 2006 by the American Chemical Society. The charge-charge interaction energies were calculated at pH 5.5 using the TK-SA model. Each sequence is characterized by the energy, net charge, and number of substitutions relative to the wild-type protein. **A.** The ability of the genetic algorithm to effectively sample the sequence space searched by more exhaustive calculations is assessed. The open black circles represent results of exhaustive calculations. The grey crosshair represents the genetic algorithm. The light grey hash marks (energies below -19 kJ/mol) represent the results of the genetic algorithm when previously uncharged surface residues were also included in optimization. **B.** The relationship between the number of substitutions in the sequence and the energy of the lowest sequence with that number of substitutions. The open circles correspond to the grey crosshairs in **A**, while the open squares represent the grey hash marks in **A**. The numbers within the symbol are the net charges of those sequences at pH 5.5.

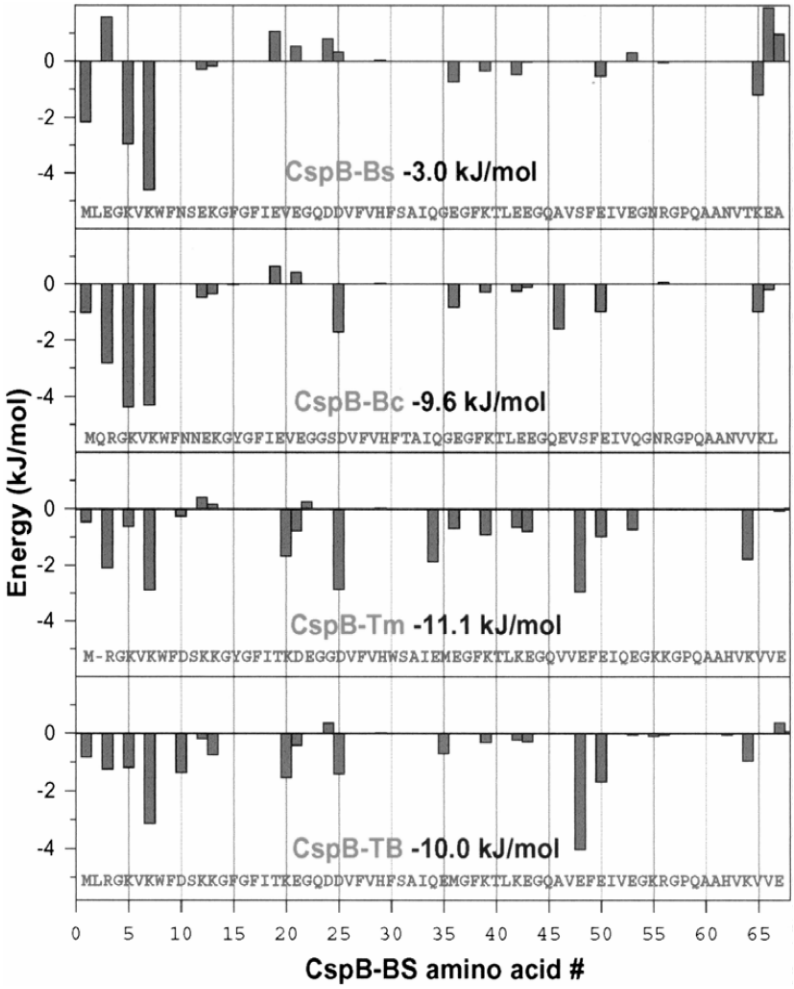

Figure 1.6 Comparison of charge-charge interaction energies for CspB-Bs, CspB-Bc, CspB-Tm, and CspB-TB, reprinted from (*15*) with permission from Elsevier. Each bar represents the total energy of charge-charge interactions of the corresponding residue with every other charged residue in the protein, averaged over an ensemble of 11 structures. Positive values of ΔG_{qq} are indicative of unfavorable interactions, while negative values correspond to favorable interactions. Of the wild-type proteins, CspB-Bs has the largest number of unfavorable interactions, while CspB-Tm has the greatest number of favorable interactions. CspB-TB was engineered to have the same core as CspB-Bs, but a similar surface charge distribution to CspB-Tm.

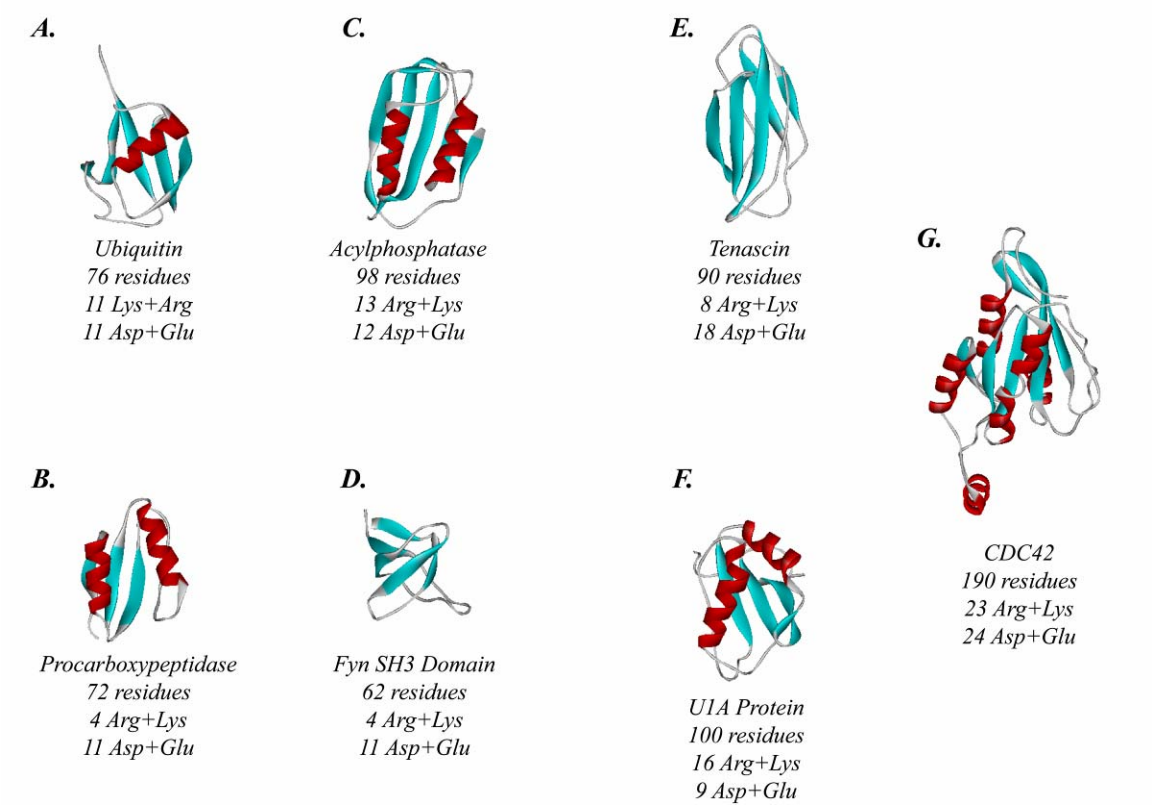

Figure 1.7 Cartoon representations of the seven proteins that have been redesigned using the TK-SA model. The surface charge-charge distribution of each protein was optimized using the genetic algorithm. The different sizes, shapes, secondary structures, and three-dimensional topologies of these proteins provide a good test of the robustness of this rational design approach. The PDB codes for the structures are **A.** 1UBQ (*153*) (ubiquitin), **B.** 1AYE (*154*) (activation domain of human procarboxypeptidase), **C.** 2ACY (*155*) (acylphosphatase), **D.** 1FYN (*156*) (Fyn SH3 domain), **E.** 1TEN (*157*) (tenascin), **F.** 1URN (*158*) (ribosomal U1A protein), and **G.** 1A4R (*159*) (CDC42).

CHAPTER 2: MATERIALS AND METHODS

2.1 Mutagenesis, Protein Expression, and Purification

2.1.1 Fyn SH3 domain variants

The FynSH3 domain mutations were generated using a PCR-based strategy. Competent *E. coli* BL21(DE3) strains were transformed by appropriate recombinant plasmid constructs (pET21d(+) vector) coding for Fyn SH3 domain variants fused to a C-terminal hexahistidine (6xHis) tag encoded by the vector. The nomenclature of the variants made is as follows: Fyn-E11K, Fyn-D16K, Fyn-H21K, Fyn-N30K, Fyn1 (E46K), Fyn2 (E11K-E46K), Fyn3 (E11K-D16K-E46K), Fyn4 (E11K-D16K-H21K-E46K), and Fyn5 (E11K-D16K-H21K-E46K-N30K). Protein expression was induced by the addition of IPTG to a final concentration of 1.5 mM to the culture media, and purification was carried out through nickel affinity column chromatography using Ni-NTA affinity matrix (Qiagen Canada, Mississauga, ON) under denaturing conditions (6M GuHCl), as previously described (*160*). Proteins were subsequently refolded through equilibrium dialysis in 50 mM sodium phosphate, 100 mM NaCl buffer. The purity of the Fyn variants was confirmed using SDS-PAGE. Protein concentrations for all experiments were determined spectrophotometrically, using a molar extinction coefficient calculated from amino acid composition (*161, 162*), of ε_{280nm} = 18,450 $M^{-1}cm^{-1}$ for all variants.

2.1.2 Tenascin and procarboxypeptidase variants

The wild-type tenascin (Ten-WT), its designed variant (Ten-GA1), and all procarboxypeptidase variants (Pc-WT, Pc-GA1, and Pc-GA2) used in the kinetics experiments were cloned, expressed, and purified as previously described (*16*). The purity of the tenascin and procarboxypeptidase variants was confirmed using SDS-PAGE. The protein concentrations for

equilibrium denaturation and kinetic experiments were determined spectrophotometrically, using molar extinction coefficients calculated from amino acid composition (*161, 162*), of ε_{280nm} = 9,970 $M^{-1}cm^{-1}$ for the tenascin variants and ε_{280nm} = 6,990 $M^{-1}cm^{-1}$ for the procarboxypeptidase variants.

2.1.3 Ubiquitin variants

Recombinant wild-type human ubiquitin (Ubq-WT) with an N-terminal 6xHis tag and a Tev protease (Tev) cleavage site in the T7 expression vector was purchased from Blue Heron® Biotechnology (Bothell, WA). The supercharged ubiquitin variant (Ubq-SC) was cloned using assembly PCR. *E. coli* BL21(DE3)pLysS cells were transformed with each plasmid, and the transformed cells were incubated in 1 L of 2YT media at 37 °C, 130 rpm, to an OD_{600} of 0.8. Protein expression was induced by adding IPTG to a final concentration of 1.5mM and incubating for five hours at 37 °C, 225 rpm. The proteins were purified by Ni-NTA affinity chromatography under denaturing conditions (8M urea) as previously described (*16*). Protein concentrations for all experiments were determined spectrophotometrically, using a molar extinction coefficient (*161, 162*), of ε_{280nm} = 1,490 $M^{-1}cm^{-1}$ for both variants.

The N-terminal 6xHis tags were removed by either cyanogen bromide (CNBr) cleavage or Tev protease cleavage. CNBr cleaves at the C-terminal side of methionine residues, resulting in the removal of the 6xHis-Tag, including the N-terminal methionine of the ubiquitin constructs. CNBr cleavage was performed by dissolving 40mg of lyophilized protein powder in 40mL of 70% formic acid. A 0.66 M solution of CNBr in 80% formic acid was freshly prepared, and 40mL of this solution was added to the protein/70% formic acid mixture under N_2. The reaction was incubated in the dark at room temperature for 24 hours. To stop the reaction, the CNBr/protein mixture was diluted 5x with water and lyophilized. Once dry, the protein powder was dissolved in 8 M urea, 100 mM sodium phosphate, 10 mM Tris-HCl, pH 8.0. The cleaved

protein was separated from uncleaved protein using Ni-NTA purification under denaturing conditions as described above. The flow-through was collected, dialyzed into 5% acetic acid, and lyophilized. The purity of the CNBr cleaved variants (Ubq-WT-CNBr and Ubq-SC-CNBr) was confirmed using SDS-PAGE.

The recognition sequence for Tev protease is –ENLYFQG– and cleavage occurs at the C-terminal side of the glutamine residue. Tev cleavage of the ubiquitin constructs removes the 6xHis-tag, while leaving a glycine residue and the N-terminal methionine residue on the N-terminus of ubiquitin constructs. Tev cleavage of Ubq-SC was performed at 4 °C in 50 mM Tris, 1 mM EDTA, 1 mM DTT, pH 8.0 buffer overnight. The cleavage of Ubq-WT was less efficient, so cleavage was performed at 4 °C in 50 mM Tris, 1 mM EDTA, 1 mM DTT, pH 8.0 buffer for 10 days, with one round of dialysis after five days. The cleavage reactions were then dialyzed extensively into 100 mM sodium phosphate, 10 mM Tris, pH 8.0 buffer to remove the EDTA and DTT. The cleaved proteins were separated from any uncleaved protein in the sample by Ni-NTA purification under denaturing conditions as described above. The flow-through was collected, dialyzed into 5% acetic acid, lyophilized, and the purity of the Tev cleaved variants (Ubq-WT-Tev and Ubq-SC-Tev) was confirmed with SDS-PAGE.

2.1.4 Purification of proteins for PPC experiments

The ribonuclease A (RNaseA), lysozyme, ubiquitin and cytochrome c proteins used for pressure perturbation calorimetry experiments were purchased from Sigma-Aldrich (St. Louis, MO) and used without further purification. Wild-type eglinC (EgC) was purified as previously described (*16, 163*). Protein concentrations for all experiments were determined spectrophotometrically, using a molar extinction coefficient (*161, 162*), of ε_{280nm} = 10,008 $M^{-1}cm^{-1}$ for RNaseA, ε_{280nm} = 38,460 $M^{-1}cm^{-1}$ for lysozyme, ε_{280nm} = 1,280 $M^{-1}cm^{-1}$ for bovine ubiquitin

(Ubq), ε_{590nm} = 11,220 $M^{-1}cm^{-1}$ for cytochrome *c*, and ε_{280nm} = 14,440 $M^{-1}cm^{-1}$ for eglinC. The purity of eglinC was confirmed using SDS-PAGE.

2.2 MALDI-TOF Mass Spectrometry

The identities of the purified protein variants were confirmed using matrix-assisted laser desorption/ionization-time of flight (MALDI-ToF) mass spectrometry (Voyager DE-PRO, PerSeptive Biosystems/Applied Biosystems). Samples were prepared for MALDI-ToF by diluting the protein stock solution (stock concentrations varied between 0.5 and 1.4 mg/mL) 1:10 with matrix. The matrix solution was prepared by washing 10mg of sinapinic acid with hexane to remove impurities, and then dissolving in 1 mL of MilliQ™ H_2O containing 30% acetonitrile and 0.1% TFA. After vortexing 1-2 minutes, the solution was centrifuged at 2500 *g* for one minute to pellet any undissolved matrix components. Three spectra were accumulated and averaged for each of the FynSH3 variants, and the averaged data were processed using Data Explorer, version 4.0 (Applied Biosystems; Penn State College of Medicine Macromolecular Core Facility). Data for the remaining proteins were also the average of three spectra, collected on a MALDI-ToF/ToF instrument at the RPI Mass Spectrometry Core Facility, and were processed using the Micromass ToF Spec 2E Mass Spectrometer software suite (RPI Mass Spectrometry Core Facility). The experimentally measured masses were compared to the theoretical masses, based on amino acid composition, calculated by the ExPASy proteomics server (*164*). The experimentally measured molecular masses of all protein variants were all within 3-5 Da of the masses calculated from the amino acid sequence.

2.3 Differential Scanning Calorimetry (DSC)

2.3.1 Fyn SH3 domain variants

The DSC experiments were performed using a VP-DSC instrument (MicroCal, Northampton, MA), at a scan rate of 90 ºC/hr as previously described (*165*). The Fyn variants were prepared for DSC by dialyzing extensively against 50mM sodium phosphate buffer, pH 7.0, containing 100 mM NaCl. The partial specific volume of the protein was calculated from amino acid composition as previously described (*166*). The values used were: 0.717 cm^3/g for Fyn-WT; 0.719 cm^3/g for all single variants; 0.720 cm^3/g for Fyn2; 0.723 cm^3/g for Fyn3; 0.725 cm^3/g for Fyn4; and 0.727 cm^3/g for Fyn5. Reversibility of unfolding of the variants was determined by stopping the DSC scan just after the transition, and then rescanning the same sample.

2.3.2 Ubiquitin variants

Experiments to test the reversibility of thermal denaturation of the Ubq-WT and Ubq-SC variants were performed under several different conditions. The partial specific volumes of each of the proteins were calculated from amino acid composition as previously described (*166*). The values used were: 0.747 cm^3/g for Ubq-WT-CNBr and Ubq-WT-Tev and 0.757 cm^3/g for Ubq-SC-CNBr and Ubq-SC-Tev. Experiments with the CNBr cleaved proteins (Ubq-WT-CNBr and Ubq-SC-CNBr) were performed at pH 7.0 in the presence and absence of salt (50 mM sodium phosphate buffer, 100 mM NaCl, pH 7.0 or 50 mM sodium phosphate buffer, pH 7.0). Both variants were reversible under all pH 7.0 conditions, so additional experiments were performed in 50mM sodium acetate buffer, pH 5.0. Experiments with the Tev cleaved proteins (Ubq-WT-Tev and Ubq-SC-Tev) were performed at both pH 5.0 and pH 7.0 in the absence of salt. A full thermodynamic characterization of Ubq-SC-Tev was obtained by performing more DSC scans using 50mM sodium acetate buffer (pH 3.5, pH 4.0, and pH 4.5).

2.3.3 Analysis of DSC experiments using a two-state model of unfolding

DSC measures the partial molar heat capacity of a protein as a function of temperature ($C_{P,\mathrm{Pr}}^{\exp}(T)$), which can be obtained by measuring the apparent difference in heat capacity ($\Delta C_P^{app}(T)$) between two identical cells – one containing only buffer (reference) and one containing a dilute protein/buffer solution (sample) (*165, 167*):

$$C_{P,\mathrm{Pr}}^{\exp}(T) = \frac{C_{p,buf}}{\overline{V}_{buf}}\overline{V}_{pr} - \frac{\Delta C_P^{app}(T)}{m_{pr}M} \qquad \textbf{(2.1)}$$

where $C_{p,buf}$ is the partial molar heat capacity of the buffer; $\overline{V}_{buf}$ is the partial molar volume of the buffer; $\overline{V}_{pr}$ is the partial molar volume of the protein; m_{Pr} is the mass of the protein in the cell, which can be calculated from the protein concentration (c) and the volume of the calorimetric cell (V_{cell}); and M is the molecular mass of the protein.

The excess heat capacity function ($C_P^{exc}(T)$) describes the heat absorbed by the protein during the unfolding reaction. It can be obtained by subtracting the progress heat capacity function ($C_P^{prg}(T)$) from the experimentally measured heat capacity:

$$C_P^{exc}(T) = C_P^{\exp}(T) - C_P^{\mathrm{Pr}\,g} \qquad \textbf{(2.2)}$$

where $C_P^{prg}(T)$ is defined by the fraction of native (F_N) and unfolded (F_U) protein in the sample:

$$C_P^{prg}(T) = F_N(T)\cdot C_{P,N}(T) + F_U \cdot C_{P,U}(T) \qquad \textbf{(2.3)}$$

where $C_{P,N}(T)$ and $C_{P,U}(T)$ are the partial molar heat capacities of the native and unfolded states, respectively. The change in heat capacity upon unfolding ($\Delta C_P = \Delta C_{P,U} - \Delta C_{P,N}$) defines the temperature dependence of the enthalpy of unfolding (ΔH). Therefore, the change in enthalpy upon unfolding ($\Delta H_{cal}(T_m)$) can be defined as the area under the excess heat capacity profile:

$$\Delta H_{cal}(T_m) = \int_0^\infty C_P^{exc}(T)dT \quad \textbf{(2.4)}$$

Another way to determine the enthalpy of unfolding in a calorimetric experiment is to determine the van't Hoff enthalpy ($\Delta H_{vH}(T_m)$) (*168*):

$$\Delta H_{vH}(T_m) = -R\frac{d(\ln(K_{eq}))}{d(1/T)} = \frac{4RT_m^2 C_P^{\max}(T_m)}{\Delta H_{cal}(T_m)} \quad \textbf{(2.5)}$$

where $C_P^{\max}(T_m)$ is the maximum value of $C_P^{exc}(T)$. For a protein that undergoes two-state unfolding ($N \Leftrightarrow U$), the values of ΔH_{vH} to ΔH_{cal} should be within 5% of each other (*5, 168*).

All DSC profiles were analyzed according to a two-state transition model. In-house scripts of the non-linear regression routine, NLREG, were used to perform global fits of the data, keeping the native and unfolded state baselines and ΔC_P the same for all experiments performed with the same protein or set of protein variants. The fitted parameters were $\Delta H(T_m)$, ΔC_p, and T_m. Using these parameters, the Gibbs free energy of unfolding ($\Delta G(T)$) is calculated by:

$$\Delta G(T) = -RT\ln(K_{eq}) = \Delta H(T) - T\Delta S(T) \quad \textbf{(2.6)}$$

$$\Delta H(T) = \int \Delta C_P dT = \Delta H(T_o) + \Delta C_P(T - T_o) \quad \textbf{(2.7)}$$

$$\Delta S(T) = \int \frac{\Delta C_P}{T} dT = \Delta S(T_o) + \Delta C_P \ln\left(\frac{T}{T_o}\right) \quad \textbf{(2.8)}$$

By defining the reference temperature (T_o), as T_m, it is possible to define ΔH_o as $\Delta H(T_m)$. For a two-state transition, the fractions of native and unfolded protein at T_m are equal ($K_{eq} = 1$), so:

$$\Delta G(T_m) = 0 = \Delta H(T_m) - T_m\Delta S(T_m) \quad \textbf{(2.9)}$$

which means that it is also possible to define the change in entropy upon unfolding ($\Delta S(T_o)$) in experimentally accessible terms:

$$\Delta S(T_m) = \frac{\Delta H(T_m)}{T_m} \quad \textbf{(2.10)}$$

Therefore, Eqs. 2.7 and 2.8 can be represented as:

$$\Delta H(T) = \int \Delta C_P dT = \Delta H(T_m) + \Delta C_P (T - T_m) \quad \textbf{(2.11)}$$

$$\Delta S(T) = \int \frac{\Delta C_P}{T} dT = \Delta S(T_m) + \Delta C_P \ln\left(\frac{T}{T_m}\right) \quad \textbf{(2.12)}$$

2.4 Spectroscopic Characterization of Protein Stability

2.4.1 Thermal denaturation – circular dichroism spectroscopy (CD)

The thermal unfolding of the Fyn variants was monitored by following the changes in the ellipticity at 220nm on an Aviv Circular Dichroism spectrometer Model 62A DS (Aviv Associates), as previously described (*160*). The thermal unfolding of the tenascin and procarboxypeptidase variants used in the kinetics studies were monitored using a Jasco-715 spectropolarimeter in 50 mM sodium phosphate buffer, pH 7.0, as previously described (*16*). The protein concentrations for the Fyn, tenascin, and procarboxypeptidase variants were 0.05 mg/mL. The unfolding of the tenascin and procarboxypeptidase variants were measured by monitoring changes in ellipticity at 230 nm (tenascin) or 222 nm (procarboxypeptidase) as a function of temperature. The protein concentrations were 0.05 mg/mL for each protein sample. The ellipticity was measured every 1 °C from 5 °C to 95 °C, with a scan rate of 1 °C per minute. The temperature was maintained in 1 cm rectangular quarts cuvettes using a Jasco PTC 424S/15 Peltier cell holder. Reversibility of unfolding was checked by allowing the samples to cool and equilibrate at 5 °C and measuring the far-UV CD spectra of the refolded proteins.

2.4.2 Analysis of thermal denaturation data

The spectroscopic thermal denaturation data were fit to a two-state model of unfolding using the nonlinear regression software, NLREG (*169*). The changes in mean residue ellipticity

($[\Theta]$) as a function of temperature can be represented by the fraction of native (F_N) and unfolded (F_U) protein in the sample (*110*):

$$[\Theta](T) = F_N(T) \cdot [\Theta]_N(T) + F_U(T) \cdot [\Theta]_U(T) \quad \textbf{(2.12)}$$

where $[\Theta]_N(T)$ and $[\Theta]_U(T)$ are the mean residue ellipticities of the native and unfolded states, respectively. The relative fraction of native protein at a given temperature is:

$$F_N(T) = 1 - F_U(T) = \frac{1}{1 + K_{eq}(T)} \quad \textbf{(2.13)}$$

From Eq. 2.6, K_{eq} can also be represented as:

$$K_{eq}(T) = \exp\left(-\frac{\Delta G(T)}{RT}\right) \quad \textbf{(2.14)}$$

The thermodynamic parameters of T_m, $\Delta G(T)$, $\Delta H(T_m)$, and ΔC_P can then be obtained by incorporating Eqs. 2.6-2.9 into the analysis.

2.4.3 Urea-induced denaturation (CD & fluorescence spectroscopy)

Equilibrium urea-induced denaturation of the tenascin and procarboxypeptidase variants were monitored using a Jasco-715 spectropolarimeter in 50 mM sodium phosphate buffer, pH 7.0 as previously described (*16*). The unfolding of the tenascin variants was monitored by following changes in ellipticity at 230 nm, 37 ºC, while the procarboxypeptidase variants were monitored at 222 nm, 25 ºC. The protein concentrations used in all experiments were 0.05 mg/mL, with an initial volume of 2 mL of the protein/buffer solution in a 1 cm rectangular quartz cuvette. Small volumes of 0.05 mg/mL protein in 9 M urea/buffer solution were titrated into the sample using the Jasco ATS-429S/15 automatic titration system, controlled by scripts contained in the Jasco software. The samples were allowed to equilibrate by stirring for 15 minutes before the change in ellipticity was recorded at 230 nm for tenascin and 222 nm for procarboxypeptidase.

Urea denaturation of the tenascin and procarboxypeptidase variants were also studied by following changes in intrinsic tryptophan fluorescence using a FluoroMax fluorimeter in 50 mM sodium phosphate buffer, pH 7.0. These experiments were also performed at 37 ºC for the tenascin variants and 25 ºC for the procarboxypeptidase variants. The protein concentrations used for these experiments were 5 μM protein, with an initial volume of 2 mL protein/buffer solution in a 1 cm cuvette. A MicroLab500 automatic titration system, controlled by scripts in the FluoroMax software, was used to titrate small volumes of a stock solution of 5 μM protein in a 9 M urea/buffer solution into the cuvette. The samples were allowed to equilibrate for 20 minutes before the spectra were acquired using an excitation wavelength of 290 nm and monitoring the emission from 310-450 nm.

2.4.4 Analysis of urea denaturation data using linear extrapolation method

The stabilities of the tenascin and procarboxypeptidase variants due to urea denaturation were determined using the linear extrapolation method (*101, 170*). For a protein that unfolds via a two-state mechanism, the equilibrium constant (K_{eq}) can be represented by:

$$K_{eq}([Urea]) = \frac{y_N([Urea]) - y([Urea])}{y([Urea]) - y_U([Urea])} \qquad \textbf{(2.15)}$$

where *y([Urea])* is the experimental observable (mean residue ellipticity for CD or intensity for fluorescence) at a given urea concentration, y_N*([Urea])* is the value of that observable in the native state of the protein, and y_U*([Urea])* is the value of the observable in the unfolded state of the protein. The mid-point of the transition, c_m, is the concentration of urea where $K_{eq} = 1$. By substituting this representation of K_{eq} into Eq. 2.6, it is possible to determine the stability of the protein as a function of urea concentration. ΔG varies linearly with [*Urea*], making it possible to determine the stability of the protein in the absence of denaturant by extrapolating the data to 0 M *[Urea]* (ΔG_{H2O}) (*170*):

$$\Delta G = \Delta G_{H_2O} - m \cdot [Urea] \quad \textbf{(2.16)}$$

where the slope of the line (*m*-value) represents the strength of the denaturant, and has been found to be proportional to the amount of surface area exposed upon unfolding (*171*).

2.5 Analytical Ultracentrifugation (AUC)

Analytical ultracentrifugation experiments were performed on a Beckman XLA ultracentrifuge. The absorbances of the wild-type (Fyn-WT) and designed (Fyn5) Fyn variants were monitored at 280 nm and samples were allowed to equilibrate at three different rotor speeds (22,000, 28,000 and 37,000 rpm) at 20 °C. The Fyn variant experiments were performed in 50 mM sodium phosphate, 100 mM NaCl buffer, pH 7.0. For Ubq-WT-CNBr and Ubq-SC-Tev, the absorbances were monitored at 276 nm and the samples were allowed to equilibrate at three different rotor speeds (20,000, 25,000, and 37,000 rpm) at 20 °C. The ubiquitin experiments were performed in 50 mM sodium phosphate buffer, pH 7.0. Absorbance data were globally fitted to a single species model:

$$A(r) = A_o \cdot \exp\left(M \cdot \frac{(1 - \bar{v} \cdot \rho) \cdot \omega^2}{2RT} \cdot (r^2 - r_o^2) \right) + E \quad \textbf{(2.17)}$$

where *A(r)* is the absorbance at a given radius *r*; A_o is the absorbance at a reference radius r_o; *M* is the molecular mass of the species in the cell; *E* is the baseline offset; $\bar{v}$ is the partial specific volume (0.717 cm^3/g for Fyn-WT, 0.727 cm^3/g for Fyn5, 0.747 cm^3/g for Ubq-WT-CNBr and 0.757 cm^3/g for Ubq-SC-Tev); calculated according to (*166*); ρ is the density of the solution (assumed to be 1g/ml); and ω is the rotor angular velocity.

2.6 Generation of Unfolded State Ensembles

2.6.1 Random-coil library (RC)

The random-coil libraries of ubiquitin (16,000 structures), apomyoglobin (8,000 structures), and staphylococcal nuclease (5,000 structures) were generated by the method of Sosnick and co-workers as previously described (*172*). Briefly, a library of backbone dihedral angles is generated based on the frequency of occurrence in the PDB. All dihedrals that are contained in, or are adjacent to, regions of regular secondary structural elements (i.e. α-helices, β-sheets, or turns) are not included in the library. Dihedral angles are assigned to each amino acid based on the probabilities that the φ/ψ angles exist in the library. The statistical potential also includes a parameter to account for nearest neighbor effects due to the adjacent residues. Steric overlap that can occur due to building the backbone by successively adding one residue to the C-terminus of the chain is removed by nudging the φ /ψ angles slightly. To account for potential steric clashes between backbone, and side chains of the protein, the side chains are modeled as a soft sphere with 90% of the volume of the original side chain, and possible rotameric conformations are sampled. After the lowest energy backbone conformations are built, the side chains are added using homology modeling.

2.6.2 Excluded-volume limit library (EV)

The excluded-volume limit libraries of ubiquitin (2,000 and 100,000 structures), apomyoglobin (20,000 structures), tenascin (2,000 structures), and ribosomal N-terminal L9 protein (NTL9, 2,000 structures) were generated by the method of Pappu and co-workers, as previously described (*173*). In this model of the unfolded state, all intramolecular interactions are ignored with the exception of excluded volume requirements. The polypeptide chain is assigned random, sterically allowed torsional angles. In each step of Metropolis Monte Carlo simulation,

one residue is randomly selected to have either its side chain or backbone torsional angles altered. If this rotation minimizes the excluded volume energy function, then the change is kept for the next round of simulation, otherwise the residue reverts to the configuration in which it started the simulation step. This process is repeated iteratively until the simulations have converged. For the purposes of these simulations, convergence is determined by calculating the ensemble averaged radius of gyration (<R_g>) and asphericity (<δ>) as a function of the number of structures. When <R_g> and <δ> no longer change with the addition of more structures to the ensembles, then convergence has been reached.

2.7 Calculation of Charge-Charge Interactions in the Unfolded State

The charge-charge interactions in the unfolded state were calculated using Debye-Hückel theory:

$$E_{ij} = \frac{q_i q_j}{4\pi\varepsilon_o}\frac{\exp(-\kappa r)}{r} \qquad \textbf{(2.18)}$$

where q_i and q_j are the charges of residues i and j; ε_o is the dielectric constant of the solvent, taken to be 78.5 for water; κ is a constant related to the ionic strength of the solvent (I) and is equal to $I^{1/2}/3.04$ A^{-1} at room temperature; and r is the distance between charges. The charge-charge interactions were calculated for each individual structure in the library. The pairwise charge-charge interaction energies of each structure were averaged over the entire population of structures the library ($E_{ij,wavg}$) by weighting each pairwise energy (E_{ij}) by the probability that E_{ij} occurs in the population ($P(E)$):

$$E_{ij,wavg} = \frac{\sum E_{ij} \cdot P(E)}{\sum P(E)} \qquad \textbf{(2.19)}$$

The average energies of charge-charge interactions calculated based on the structures in the EV limit library were compared to the energies expected based on the Gaussian polymer chain model of the unfolded state (*93*):

$$E_{ij} = 322\int_0^\infty \frac{p(r)\exp(-\kappa r)}{\varepsilon r}dr = \frac{332\left(\frac{6}{\pi}\right)^{1/2}\left[1-\pi^{1/2}x\exp\left(x^2\right)erfc(x)\right]}{\varepsilon d} \qquad \textbf{(2.20)}$$

where $p(r) = 4\pi r^2\left(\frac{3}{2\pi d}\right)^{3/2}\exp\left(-\frac{3r^2}{2d^2}\right)$; d is the root mean square distance, defined as $d = 7.5l^{1/2} + 5$, where l is the number of residues separating i and j; $x = \kappa d/6^{1/2}$, where κ is related to the ionic strength of the solvent, as described above; and $erfc(x)$ is the complementary error function.

2.8 Molecular Dynamics Simulations (MD)

The molecular dynamics simulations of individual structures in the ubiquitin, tenascin, and NTL9 EV limit libraries were performed using the AMBER99SB (*174*) force-field contained in the AMBER9 molecular dynamics software package (*175*). Simulations were run using the GB-SA solvent model (*176*) at an ionic strength of 100mM. The low pH limit ("pH 2" in Section 4.2) was modeled by neutralizing the charges on all acidic residues (including the C-terminus) and protonating the histidines. The high pH limit ("pH 14" in Section 4.2) was modeled by neutralizing the charges on all basic residues (including the N-terminus). The neutral pH structures ("pH 7" in Section 4.2) were modeled by charging all residues and termini and deprotonating the histidines. The radius of gyration (R_g) of each structure was calculated as a function of simulation time using scripts in AMBER9. These values were then averaged over the

population of the library, and simulations were run until the average R_g did not change significantly as a function time. The R_g of the EV limit library of 2,000 ubiquitin structures appeared to converge after 300 ps of simulation time. Therefore, all subsequent simulations were run for 300 ps. The calculations were run on the supercomputer available through RPI's Computational Center for Nanotechnology Innovations (CCNI). Using 50 nodes of four 2.6 GHz AMD Opteron processors, it was possible to run 300 ps simulations for 100 structures in 10-12 hours for a total of 20,000 to 24,000 processor hours of computational time and 600 ns of simulation time per 2,000 structure library.

2.9 Kinetic Measurement of Protein Folding and Unfolding Reactions

2.9.1 Stopped-flow fluorescence and circular dichroism spectroscopy

Stopped-flow experiments were performed on a Jasco-815 spectropolarimeter with a BioLogic SFM300 stopped-flow mixer attachment. The photodetector can be set up in either fluorescence mode or CD mode. All fluorescence experiments were performed in 50 mM sodium phosphate buffer, pH 7.0, while all CD experiments were performed in 10 mM sodium phosphate buffer, pH 7.0. For experiments performed in fluorescence mode, a solution of 10 μM protein was diluted 1:11 with buffer and 8 M urea in buffer to a final concentration of 1 μM. For CD experiments, a 2-5 mg/mL solution of protein was diluted 11-fold with buffer and urea to a final concentration of 0.2-0.5 mg/mL. More concentrated protein solutions yield better signal to noise ratios in the CD mode. Protein stock solutions for the refolding arm of the chevron were made by directly dissolving the lyophilized protein powder into 8 M urea, 10 mM or 50 mM sodium phosphate buffer, pH 7.0. For the unfolding arm of the chevron, the protein stock solutions were

made by dissolving the protein into 5% acetic acid, and then dialyzing extensively into the appropriate buffer solution.

2.9.2 Manual mixing

The unfolding rates of the tenascin variants were too slow (>1,000 sec) to be monitored using the stopped-flow method. Therefore, unfolding kinetics of the tenascin variants at 37 ºC were monitored using manual mixing. The samples were excited at a wavelength of 295 nm and changes in the emission fluorescence at 350 nm were followed as a function of time in a 1 cm cuvette. A concentrated stock solution of 9 M urea in 50 mM sodium phosphate buffer, pH 7.0 was diluted with buffer, inverted five times to mix, and placed in the fluorimeter to measure a buffer blank. The buffer was allowed equilibrate to the experimental temperature of 37 ºC by stirring for 1 minute. A buffer baseline spectrum was then collected for 1 minute. After acquisition of the buffer spectrum, 20 μL of the urea/buffer mixture was removed, and the same volume of a concentrated stock solution of protein in 50 mM phosphate buffer, pH 7.0 was added to the solution for a final protein concentration of 5 μM. The protein solution was mixed by pipetting up and down repeatedly, and then the emission at 350 nm was followed as a function of time for 600-1200 seconds. Each data point in the chevron represents the average of three individual experiments.

2.9.3 Analysis of kinetic data

The tenascin stopped-flow and manual mixing data were fit to a single exponential for unfolding and a double exponential for the refolding data points, as previously described (*177*). The procarboxypeptidase stopped-flow data were fit to single exponentials for both the unfolding and refolding experiments. In order to obtain the folding and unfolding rates in the absence of

denaturant, the chevrons for the tenascin and procarboxypeptidase variants were fit to a two-state model, as previously described (*178*):

$$\ln k_{obs} = \ln\left(k_{f,H_2O}\exp\left(-m_f\left[Urea\right]\right)+k_{u,H_2O}\exp\left(-m_u\left[Urea\right]\right)\right) \quad \textbf{(2.21)}$$

where k_{obs} is the experimentally observed rate constant at a given urea concentration ([*Urea*]); $k_{f,H2O}$ is the folding rate constant in the absence of denaturant; $k_{u,H2O}$ is the unfolding rate constant in the absence of denaturant; m_f is the slope of the refolding arm of the chevron; and m_u is the slope of the unfolding arm of the chevron.

2.10 Pressure Perturbation Calorimetry (PPC)

Ribonuclease A (RNaseA), hen egg white lysozyme (HEWL), horse heart cytochrome *c* (CytC), and bovine ubiquitin (Ubq) were purchased from Sigma-Aldrich (St. Louis, MO). Wild-type eglinC (EgC), Tev-cleaved supercharged ubiquitin (Ubq-SC-Tev), and CNBr cleaved human ubiquitin (Ubq-WT-CNBr) were purified as described in section 2.1.

Parallel DSC experiments were performed on all proteins studied by PPC, except eglinC which had been previously characterized (*163*). All experiments were performed in 20-100 mM glycine buffer, pH 2.4-3.6, with a protein concentration of 0.6-1.0 mg/mL. In order to examine what cold denaturation would look like with PPC, DSC experiments were performed to screen for conditions where Ubq-SC-Tev would cold denature while still giving a good signal. These experiments were performed in 50 mM glycine (pH 2.6-3.4) or 50 mM sodium acetate buffer (pH 3.5-4.75). It was found that pH 3.5 would be the best condition for studying the cold denaturation of Ubq-SC-Tev by PPC. The partial specific volume of the proteins were calculated from amino acid composition as previously described (*166*). The values used were: 0.721 cm^3/g for RNaseA, 0.729 cm^3/g for lysozyme, 0.783 cm^3/g for cytochrome *c*, 0.747 cm^3/g for Ubq, and 0.734 cm^3/g for eglinC.

PPC experiments for RNase, HEWL, and Ubq were performed in 50 mM glycine buffer, pH 2.2-3.4. CytC experiments were performed in 50 mM glycine, pH 2.4-3.4 and 100 mM glycine, pH 3.6. EgC experiments were performed in 20 mM glycine, pH 2.5-3.5 to compare with previously published data (*163*). The Ubq-SC-Tev experiment was performed in 50 mM sodium acetate, pH 3.5 to determine a cold denaturation profile for pressure perturbation calorimetry. The Ubq-WT-CNBr experiment was performed in 50 mM $NaPO_4$ buffer, pH 7.0 to determine the shape of the PPC native state baseline over a broad temperature range.

The PPC experiments were performed on a MicroCal VP-DSC with a PPC attachment (MicroCal, LLC, Northampton, MA). The protein solutions were dialyzed extensively against the corresponding buffers at 25 °C using Spectrapor3 dialysis membranes with a 3.5 kDa molecular weight cutoff. Samples were centrifuged at 13,000 rpm in an Eppendorf 5417R microcentrifuge for 20-30 minutes at 25 °C to remove insoluble material present in the solution after dialysis. The experiments were performed using protein concentrations between 0.8 and 4.0 mg/mL. The partial specific volumes of the proteins were calculated as previously described (*166*). The values used were: 0.721 cm^3/g for RNaseA, 0.729 cm^3/g for HEWL, 0.783 cm^3/g for CytC, 0.747 cm^3/g for Ubq, 0.734 cm^3/g for EgC, 0.757 cm^3/g for Ubq-SC-Tev, and 0.747 cm^3/g for Ubq-WT-CNBr.

In order to calculate the coefficient of thermal expansion (α_{Pr}) for a protein, the expansion effects of water and buffer salts must be taken into account. Therefore, control runs of water/water, buffer/water, and buffer/buffer were also performed before each experiment. The temperature range for the control experiments was 5 °C – 110 °C, with data collected every 5 °C. The protein/buffer experiments were also performed in the temperature range of 5 °C – 110 °C, with data collected every 5 °C in the baseline range, and every 2 °C in the transition region, which was determined by corresponding DSC experiments.

Although the foundations for analyzing data from a PPC experiment have already been described (*179*), a brief overview of the thermodynamic relationship between changes in pressure, volume, and heat, will help make the analysis presented in Chapter 7 more clear. A PPC experiment measures small changes in the heat absorbed/released by the solution in the calorimetric cell as small perturbations in pressure (ΔP) are applied. Starting from the second law of thermodynamics, and differentiating with respect to pressure, we can represent this change as:

$$dQ = dS \cdot T \Rightarrow \left(\frac{\partial Q}{\partial P}\right)_T = T\left(\frac{\partial S}{\partial P}\right)_T \quad \textbf{(2.22)}$$

where dS is the change in entropy at temperature, T, for a reversible change in heat, dQ. The Maxwell relationship $\left(\frac{\partial S}{\partial P}\right)_T = -\left(\frac{\partial V}{\partial T}\right)_P$, makes it possible to express the pressure-induced changes in heat to the change in the volume (V) of the system:

$$\left(\frac{\partial Q}{\partial P}\right)_T = -T\left(\frac{\partial V}{\partial T}\right)_P = -TV\alpha \quad \textbf{(2.23)}$$

where α is the coefficient of thermal expansion and equal to $\frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_P$. Integrating Eq. 2.23 yields:

$$Q = -TV\alpha\Delta P \quad \textbf{(2.24)}$$

For a solution containing m_{Pr} grams of protein, and m_S grams of solvent, the total volume of the system, V_{Tot}, can be represented by:

$$V_{Tot} = m_{pr} \cdot \bar{v}_{pr} + m_S \cdot v_S \quad \textbf{(2.25)}$$

where $\bar{v}_{pr}$ is the partial specific volume of protein in the cell, and v_S is the specific volume of buffer. Differentiating Eq. 2.25 with respect to temperature and substituting into Eq. 2.23 yields:

$$\left(\frac{\partial Q}{\partial P}\right)_T = -T\left(m_{\mathrm{Pr}}\bar{v}_{\mathrm{Pr}}\alpha_{\mathrm{Pr}} + m_S v_S \alpha_S\right) \quad \textbf{(2.26)}$$

By integrating Eq. 2.26 over a small pressure range, we can obtain an expression for the thermal expanssivity coefficient of the protein (α_{Pr}) at temperature, T, in terms of the thermal expansion of the buffer (α_S), the change in the heat of the calorimetric cell (ΔQ) and the change in the pressure (ΔP) of the system:

$$\alpha_{\mathrm{Pr}} = \alpha_S - \frac{\Delta Q}{T \Delta P m_{\mathrm{Pr}} \bar{v}_{\mathrm{Pr}}} \qquad \textbf{(2.27)}$$

The thermal expansion coefficient of the buffer is determined from the buffer/water scans in a similar manner:

$$\alpha_S = \alpha_{H_2O} - \frac{\Delta Q}{T \Delta P V_{cell}} \qquad \textbf{(2.28)}$$

where V_{cell} is the volume of the calorimetric cell. The water/water scans, mentioned above, determine the value of the thermal expansion coefficient of water (α_{H2O}). The raw data from the PPC experiments were processed using the scripts in the Origin PPC data analysis software supplied by MicroCal (Northampton, MA) to obtain values for α_{Pr} as a function of temperature. The novel analysis of the $\alpha_{Pr}(T)$ profiles to obtain the volumetric changes upon unfolding is presented in detail in Chapter 7.

CHAPTER 3: COMPUTATIONAL DESIGN OF THE FYN SH3 DOMAIN WITH INCREASED STABILITY THROUGH THE OPTIMIZATION OF SURFACE CHARGE-CHARGE INTERACTIONS

3.1 Introduction

The design of proteins with improved thermodynamic stability has been the focus of many protein engineering studies. Due to the widely accepted notion that the interactions in the core of a protein play a major role in determining protein stability (*2, 5*), most design approaches have been focused on optimizing interactions in the core and, as a result, the protein surface had often been ignored in such studies. However, core optimization algorithms have challenges associated with accurately modeling interactions in the tightly packed interior of proteins (*6, 7, 62*). The protein surface, on the other hand, offers a much smaller set of interactions to be optimized but was largely ignored in design procedures due to the belief that residues on the surface do not contribute significantly to stability, since their solvent exposure in the native and unfolded states are similar. However, in the native state of a protein, surface residues do participate in a number of tertiary interactions, such as hydrogen bonding or long-range electrostatic interactions. Residues that participate in these types of interactions will contribute differently to the stability of the native and the unfolded states of a protein. Since surface residues are more amenable to substitution than those in the core, they should provide effective means to manipulate the stability of a protein without affecting the structural integrity of the protein. Indeed, it has been demonstrated that surface charge-charge interactions can be successfully exploited to modulate protein stability. For example, it has been shown that neutralizing or reversing the charges of individual residues with unfavorable interaction energies successfully enhances the stability (*11-14, 67, 96, 98, 110*). In addition, it has been shown that

further increases in stability can be gained by optimizing the entire surface charge distribution (*16*).

SH3 domains are small protein-protein interaction modules that have been the subject of numerous folding studies. Structurally, the SH3 domains are comprised of two three-stranded β-sheets, orthogonally packed against one another (Fig. 3.1A). The folding of SH3 domains is well approximated by a simple two-state reaction, where a polypeptide chain folds into its native state by passing through a high energy transition state barrier in the absence of populated folding intermediates (*180-184*). In the present study, the Fyn SH3 domain is used as a model system for the rational optimization of surface charge-charge interactions to increase the stability of this protein. The Fyn SH3 domain with increased stability serves an important purpose. The wild-type Fyn SH3 domain has a higher calculated energy of charge-charge interactions (ΔG_{qq}), compared to any other protein previously optimized by this approach (*11, 15, 16, 96, 98, 110*) hereby providing insights into how much stability can be gained through the optimization of proteins possessing highly unfavorable charge-charge interaction energies. Furthermore, the kinetics of the folding and unfolding reactions of the FynSH3 domain is well characterized, providing an excellent model system for understanding how the optimization of surface charge-charge interactions affects the folding kinetics of proteins. In this chapter, the results of the experimental thermodynamic studies on the stabilities of the computationally redesigned variants of Fyn SH3 domain with optimized surface charge-charge interactions are presented.

3.2 Results and Discussion

3.2.1 Modeling charge-charge interactions in Fyn SH3 domain

To explore the possibility of optimizing surface charge-charge interactions in the Fyn SH3 domain, we first evaluated the energetics of the charge-charge interactions in this protein.

Figure 3.1C shows the energies of charge-charge interactions for the wild-type Fyn SH3 domain (Fyn-WT), as calculated using the TK-SA model (*84*). These results indicate that the wild-type protein has many unfavorable charge-charge interactions, defined by positive values of ΔG_{qq}, suggesting that the charge-charge interactions of Fyn are not fully optimized. The neutralization or reversal of the existing charges should lead to the unfavorable interactions becoming favorable. Additional favorable interactions can also be gained by introducing new charges at previously uncharged positions on the protein surface. To find the most favorable combinations of surface charges, a genetic algorithm of search (*16, 100*) combined with the TK-SA model for calculation of energy of charge-charge interactions was used as previously described (*16*). In addition to the existing charged residues, three neutral polar positions on the surface of the Fyn SH3 domain (Q27, N30, and Q53) were included in the optimization algorithm.

The dependence of the predicted energy of the charge-charge interactions on the number of the total substitutions made is shown in Figure 3.2. It is evident that the interaction energy initially becomes more favorable with increasing the number of substitutions, but begins to level off after five substitutions. One of the sequences with five substitutions (Fyn5 - E11K/D16K/H21K/N30K/E46K, see also Fig. 3.1B) that was predicted to have favorable charge-charge interactions, was selected for further experimental characterization. In the Fyn5 variant, four of the substitutions were at existing charged residues, while the fifth introduced a new charge at N30. Structurally, most of these sites are found in the loop regions of the protein (Fig. 3.1A). E11, D16, and H21 are all located in the distal loop, between the first and the second β-strands, D16 is located near the β-turn in this region, while E11 is closer to the first β-strand, and H21 is near the second β-strand. N30 is the first residue in the turn region between the second and third β-strands. E46 is considered to be the N-terminal residue of the fourth β-strand. The effect of the substitutions on the predicted energy of charge-charge interactions (on a per residue basis) in the optimized sequence, Fyn5, is compared to that of the wild-type in Figure 3.1C. It is evident that

the E11, D16, and E46 residues have unfavorable interaction energies in the wild-type protein that are predicted to become favorable upon charge reversals in the optimized Fyn5 variant. The H21 residue is already favorable in the wild-type protein but is expected to become much more so in the context of the other substitutions in the Fyn5 variant. The positive charge introduction at the N30 position is also predicted to contribute favorably to the total ΔG_{qq} of the Fyn SH3 domain.

In order to examine the additivity of the contributions of the E11K, D16K, H21K, N30K and E46K substitutions to the stability of the designed Fyn SH3 domain, we also characterized all single variants containing these substitutions (E11K, D16K, H21K, N30K, and E46K (Fyn1)), as well as Fyn2 - E46K/E11K, Fyn3 - E46K/E11K/D16K, and Fyn4 – E46K/E11K/D16K/H21K (see also Fig. 3.1B). The stepwise contributions of the five substitutions in Fyn5 were examined starting with the Fyn1 variant, which contains only one substitution: E46K. This position is unfavorable in the wild-type protein, so reversing the charge at this position is expected to cause the interaction energies to become favorable (See Fig. 3.1C). The Fyn2 construct has two substitutions: E11K/E46K, and is predicted to have favorable interaction energies at both positions, as illustrated in Figure 3.1C. The increase in the overall favorable energy of charge-charge interactions, however, is not predicted to be significantly different between the Fyn1 and Fyn2 variants (Fig. 3.2). The Fyn3 variant has three substitutions: E11K/D16K/E46K and exhibits favorable predicted interaction energies at all three positions, unlike the WT protein that possessed unfavorable interactions at these positions (Fig. 3.1C). This construct is also predicted to have a much more favorable overall energy of charge-charge interactions compared to both Fyn1 and Fyn2 variants (Fig. 3.2). Fyn4 contains four substitutions: E11K/D16K/H21K/E46K, and is predicted to be more stable than Fyn3 (Fig. 3.2). Therefore, we predict that the experimentally measured stabilities of the Fyn variants examined in this study will conform to the following rank order: Fyn5 > Fyn4 > Fyn3 > Fyn2 ≈ Fyn1 > WT.

3.2.2 Experimental evaluation of the role of charge-charge interactions in the stability of the Fyn SH3 domain

The predicted rank order in stability for the designed variants of Fyn SH3 domain was experimentally tested using several biophysical methods, as described below. For clarity, Figure 3.3 only shows the differential scanning calorimetry (DSC) profiles of the wild-type Fyn SH3 domain with four of the designed variants, obtained at neutral pH. The data for the remaining variants are given in Table 3.1. These profiles clearly show that the variants have increased thermostability relative to the WT protein, as evidenced by an increase in the temperature of the heat absorption maximum (See Fig. 3.3 and Table 3.1). The Fyn4 and Fyn5 variants have the highest transition temperatures (T_m), which is consistent with the predictions based on the calculations of charge-charge interactions (Fig. 3.2). Interestingly, it appears that the effects of the substitutions in Fyn4 and Fyn5 can be explained by the principles of additivity (ΔG_{add}, Table 3.1). The stabilities measured by DSC are in good agreement, within experimental error, with those predicted by summing the stabilities of the single variants comprising Fyn4 and Fyn5. Although Fyn5 was predicted to be more stable than Fyn4, the experimentally measured T_m and ΔG values of these variants were actually quite similar. This result can be explained by the experimental observation that the single variant containing the N30K substitution appears to have little effect on the stability of Fyn (Table 3.1). The Fyn1 and Fyn2 variants also have similar transition temperatures (T_m), which is in agreement with the prediction results given in Figure 3.2, suggesting that they do not possess significantly different charge-charge interaction energies. The similar stabilities of Fyn1 and Fyn2 can also be explained by the principles of additivity because the E11K single substitution has very little effect on the stability of the protein (Table 3.1). Furthermore, the experimentally obtained T_m of the Fyn3 construct conforms to the predicted order of stability based on the computed energies of charge-charge interactions.

Importantly this result can also be explained by additivity because the D16K substitutio has a fairly significant contribution to stability on its own (Table 3.1).

The observed increase in the transition temperature of Fyn variants could be due to the optimized energetics of charge-charge interactions, but it might also result from changes in the structure and/or oligomerization state of the protein due to the substitutions. The structural properties of the designed variants of Fyn were characterized by circular dichroism (CD) spectroscopy. The CD spectra of the wild-type and designed variants were similar, illustrating that the substitutions did not have a significant effect on the protein structure (Fig 3.4). Analytical ultracentrifugation experiments were carried out to eliminate the possibility that the amino acid substitutions changed the oligomerization state of the proteins. Analysis revealed that proteins remain monomeric under experimental conditions (Fig. 3.5). Therefore, the observed differences in the transition temperature cannot be attributed to changes in the structure or in the oligomerization state of the Fyn variants.

In addition to the transition temperature (T_m) values, DSC scans can also provide insight into whether the two-state folding mechanism of the WT protein is retained in the designed variants. For this purpose, DSC profiles were fit to a two-state unfolding model, and the validity of this model was tested in two ways. First, the van't Hoff enthalpies (ΔH_{VH}) extracted from fitting the data were compared to the calorimetric enthalpies (ΔH_{cal}) that are measured directly by DSC (Table 3.1). The ΔH_{VH} and ΔH_{cal} are within the experimental error of 5% for each variant, suggesting that these proteins do unfold via a two-state mechanism (*185*), as previously noted for WT Fyn SH3 domain (*180-182*). Second, the thermal unfolding of each of the designed variants was monitored using far-UV circular dichroism spectroscopy and demonstrated that the transition temperature (T_m) obtained from these experiments are similar to the T_m value measured by DSC. The far-UV CD unfolding experiments monitor changes in the secondary structure upon unfolding, while the DSC experiments measure the energetics of global changes in the protein

conformation. If the Fyn variants unfold via a two-state mechanism, then the stabilities and T_m measured by CD and DSC should be similar. Comparison of the fractions of unfolded proteins as a function of temperature obtained from CD and DSC experiments (Fig. 3.3, inset) shows that they are indeed similar, providing further evidence that the two-state unfolding model is valid for all Fyn variants analyzed in this study.

In order to obtain a clearer picture of the mechanism of stabilization of the designed sequences, we also measured and compared the changes in heat capacity upon unfolding (ΔC_p) of the Fyn variants to that of the WT protein. The change in heat capacity of a protein defines the temperature dependence of the enthalpy of unfolding of the protein, $\Delta C_p = (d\Delta H/dT)$. Empirically, it has been noted that ΔC_p is defined by the amount of polar and non-polar surface area that is buried in the native state (*5, 171*). Consequently, substitutions on the surface of the Fyn SH3 domain are not expected to have a large effect on the ΔC_p value of the domain. Figure 3.6 shows the temperature dependence of the enthalpies of unfolding $\Delta H(T_m)$ for all Fyn variants. It appears that the enthalpies of unfolding of all variants follow the same function, suggesting that the ΔC_p value is the same for all the of the Fyn variants used in this study. The ΔC_p estimated from the slope of this function is 3.4 ± 0.4 kJ/mol·K, which is consistent with ΔC_p values of other proteins of similar size (*16, 84*). Moreover, it corresponds well to two previous estimates of ΔC_p for the WT Fyn, 3.3±0.4 kJ/mol·K (*186*) and 3.5 kJ/mol·K (*160*). This consistency in the ΔC_p value of the Fyn SH3 variants, together with the similar enthalpy of unfolding function, suggests that enthalpic effects are not the primary mechanism of stabilization.

3.2.3 Comparison between theory and experiment

To obtain quantitative insight into the additivity of the contribution of charge-charge interactions to the experimentally measured changes in stability one needs to compare the change

in stability upon stepwise substitutions. The Gibbs free energy of unfolding for each variant was calculated at the T_m of the wild type using the Gibbs-Helmholtz equation:

$$\Delta G(T) = \Delta H(T_m) + \Delta C_P \cdot (T - T_m) - T \cdot \left[\frac{\Delta H(T_m)}{T_m} + \Delta C_P \cdot \ln \frac{T}{T_m} \right] \qquad \textbf{(3.1)}$$

where the enthalpy of unfolding (ΔH) is represented by $\Delta H(T_m) + \Delta C_P \cdot (T - T_m)$ and the entropy of unfolding (ΔS) is equal to $(\Delta H(T_m)/T_m) + \Delta C_P \cdot \ln(T/T_m)$. Figure 3.7 compares the experimentally measured differences in stability, $\Delta\Delta G_{exp} = \Delta G_{var} - \Delta G_{WT}$, with the calculated differences in stability expected from changes in the energy of charge-charge interactions, $\Delta\Delta G_{qq} = \Delta G_{WT,qq} - \Delta G_{var,qq}$. Interestingly, two of the single variants (E11K and N30K) were slightly destabilizing, relative to the wild-type, which is not what was predicted from our calculations. This is probably due to the fact that the TK-SA model only considers interactions between charges. The deviation of the experimentally measured stabilities from our predictions suggests that other intramolecular interactions such as hydrophobic interactions, secondary structure propensity, or packing interactions are more important than charge-charge interactions for stability at these positions. However, the effects of the E11K and N30K substitutions in the context of the Fyn2, Fyn3, Fyn4, and Fyn5 variants were very well predicted by the TK-SA model. These results suggest that multiple charge reversals offset the potential effects of other intramolecular interactions that destabilized the E11K and N30K single variants relative to Fyn-WT. In fact, the experimental data from all other variants correlate very well with the calculations (R=0.88), suggesting that the optimization of surface charge-charge interactions is a valid approach to stabilizing proteins. The deviation of the slope of the best-fit line from unity (m = 0.74) suggests that the calculated changes in the energies of charge-charge interactions describe the overall changes in stability qualitatively but not quantitatively. The data presented in Figure 3.7 also suggest that changes in thermostability can be qualitatively predicted using this computational approach, evidenced by

the very good correlation between computed $\Delta\Delta G_{qq}$ and experimentally measured differences in thermostability ($\Delta T_m = T_{m,var} - T_{m,WT}$) for all variants except Fyn-E11K and Fyn-N30K.

The correlations between the experimental stability data ($\Delta\Delta G_{exp}$) and the theoretical calculations ($\underline{\Delta\Delta G_{qq}}$) are generally not quantitative. This can be attributed to the simplicity and insufficient accuracy of the computational model used to calculate the energetics of charge-charge interactions. In addition, this computational model does not attempt to quantify the effects of other types of interactions that are also important for stability such as side chain hydrophobicity and secondary structure propensity. However, the correlation between the results of calculations and experiments (Fig. 3.7) signals for a seminal role of charge-charge interactions in determining the stability of the Fyn SH3 protein. It is also evident that the experimental stability data conforms to the relative rank-order of the variants stability observed in the calculations. Furthermore, the computational modeling has been able to successfully predict both the sign and, to a reasonable degree, the magnitude of the contribution of charge-charge interactions to the total protein stability. Of particular interest, the calculations predicted that the Fyn1 and Fyn2 variants would have comparable stabilities (Fig. 3.2) and the experimentally measured stabilities for these pairs were found to be similar within the experimental error (Fig. 3.7, Table 3.1).

Finally, it has been demonstrated that it is possible to computationally identify a more energetically favorable combination of surface charge-charge interactions that leads to a significant increase in thermostability (stability) of over 12 ºC (~8 kJ/mol). More importantly, the data presented here suggest that an increase in stability of such magnitude can be achieved with a small number of substitutions, as only four or five surface residues have been substituted in the most stable designed variants of the Fyn SH3 domain.

While previously published design approaches have reported larger increases in stability than we observed here, these studies substituted over half of the amino acid residues of the

protein in their design (*60, 149*). Such dramatic changes in the sequence sometimes lead to unexpected consequences. For example, it is reported that one of the these designed proteins was a dimer at the experimental concentrations and, hence, the dimerization of the protein partly contributed to the observed increase in the stability (*151*). In the case of the Fyn SH3 domain, the designed variant is monomeric in solution as determined by the analytical ultracentrifugation experiment and appears to be have very similar structure to the wild-type protein (Fig. 3.4 and Fig. 3.5). These data collectively suggest that the stabilization observed in the designed variant of the Fyn SH3 protein is likely to stem from the optimization of surface charge-charge interactions and is not from an altered dimeric state or a dramatic change in the protein structure. More interestingly, the magnitude of the increase in T_m of the designed sequences relative to the wild-type sequence with only a few surface mutations observed in this work (and elsewhere (*16*)) is comparable to studies that have engineered stability through making a few substitutions in the hydrophobic cores of model proteins (reporting increases in stability of 10 ºC (*61*) or 10 kJ/mol (*150*)). These observations serve to further support the idea that the rational design of surface charge-charge interactions is an effective strategy to complement core optimization algorithms to enhance protein stability.

3.3 Implications for Protein Design Strategies

In the present work, the rational optimization of charge-charge interactions successfully increased the thermostability of the Fyn SH3 domain sequence with only four or five substitutions. Furthermore, it was possible to qualitatively predict the stepwise effects of substitutions on the stability of each variant. For Fyn, the energy of favorable charge-charge interactions was predicted to decrease after thirteen substitutions (Fig. 3.2). A similar trend has

also been observed for ubiquitin: exhaustive calculations performed on every ionizable residue in ubiquitin indicate that the favorable energy begins to decrease after ten substitutions (*16*).

The finding that the increase in favorable interactions begins to level off suggests that there is a limit to the amount of stability that can be gained for a protein through the optimization of surface charge-charge interactions. This limit is a result of the fact that the native topology defined by a given protein sequence occupies a finite space. The addition of new charges into this space will always involve the introduction of both favorable and unfavorable interaction energies. If the substitution sites are chosen appropriately, the energy of favorable interactions will be larger than the unfavorable interaction energy. However, when the charge density increases beyond a certain point, the introduction of a new charge into the limited space of the native topology will lead to a balance between favorable and unfavorable interactions and no further increase in stability will be observed. Eventually, the charge density will become such that the introduction of new charges can only be unfavorable, so the energy of favorable charge-charge interactions (and predicted stability) will decrease. As a result, only a few sequences will produce optimal surface charge-charge interactions. To increase the stability of a protein beyond what is possible through optimization of surface charge-charge interactions, it would be necessary to optimize other types of interactions, such as hydrogen bonding, packing, or hydrophobicity.

Table 3.1: Thermodynamic parameters of unfolding for the Fyn variants at pH 7.0.

	Amino Acid Substitutions	T_m (ºC)	$\Delta H_{cal}(T_m)$ (kJ/mol)	$\Delta H_{VH}(T_m)$ (kJ/mol)	ΔG (71.6 ºC) (kJ/mol)	ΔG_{add} (71.6 ºC) (kJ/mol)	ΔG (25 ºC) (kJ/mol)
Fyn	Wild-type	71.6	232	239	0	-	19.5
	E11K	70.6	234	236	-0.7	-	19.6
	D16K	77.1	256	259	3.9	-	23.4
	H21K	76.6	249	252	3.4	-	22.3
	N30K	71.2	224	227	-0.3	-	18.4
Fyn 1	E46K	77.7	261	260	4.4	-	24.2
Fyn 2	E11K/E46K	76.2	246	252	3.1	3.7	21.8
Fyn 3	E11K/D16K/ E46K	81.9	272	270	7.4	7.6	26.2
Fyn 4	E11K/D16K/ H21K/E46K	84.5	269	274	8.9	11.0	25.9
Fyn 5	E11K/D16K/ H21K/N30K/ E46K	83.3	274	262	8.3	10.7	26.6

ΔG (T=71.6 ºC) and ΔG (T=25 ºC) represent the stabilities of each of the variants at the transition temperature of wild-type Fyn and at 25 ºC, respectively. ΔG_{add}(71.6 ºC) represents the stability you would expect for each designed variant based on the stabilities of the single variants at that temperature. These values were calculated using a ΔC_P value of 3.4 ± 0.4 kJ/(mol·K) obtained from the temperature dependence of $\Delta H(T_m)$ vs. T_m. The thermodynamic parameters have the following estimated errors: T_m: ± 0.1 ºC, $\Delta H(T_m)$: ± 5%, ΔG (T=71.6ºC): ± 1.2 kJ/mol, and ΔG (T=25ºC): ± 2.2 kJ/mol.

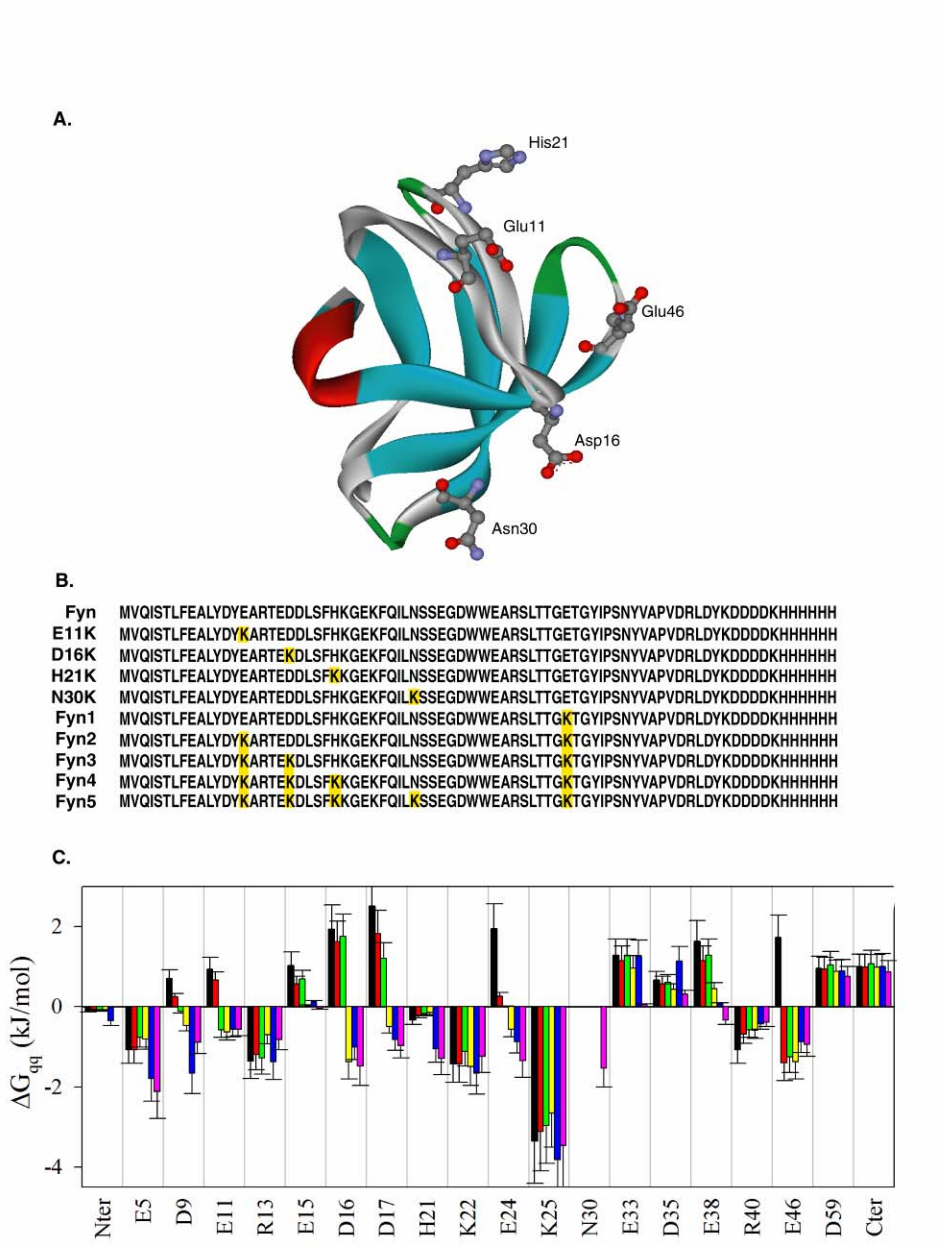

Figure 3.1 Primary sequence and the tertiary structure and charge-charge interaction energies of Fyn. **A.** Cartoon representation of the three dimensional structure of the Fyn SH3 domain (1FYN). The sites selected for substitution are represented with the ball-and-stick model. **B.** The sequence alignment of the Fyn variants with the selected substitution sites highlighted in yellow. **C.** Comparison of the energies of charge-charge interactions in the wild-type and five designed sequences of Fyn at pH 7.0. Each bar represents the total energy of charge-charge interactions for that residue with all other charged residues in the protein, averaged over the ensemble of eleven structures. The error bars represent the standard deviations of the averaged values. Favorable interactions have negative values of ΔG_{qq}, while positive values represent unfavorable ones. Black bars - wild-type, Red bars - Fyn1, Green bars - Fyn2, Yellow bars - Fyn3, Blue bars – Fyn4, and Purple bars – Fyn5 (see text for construct nomenclature).

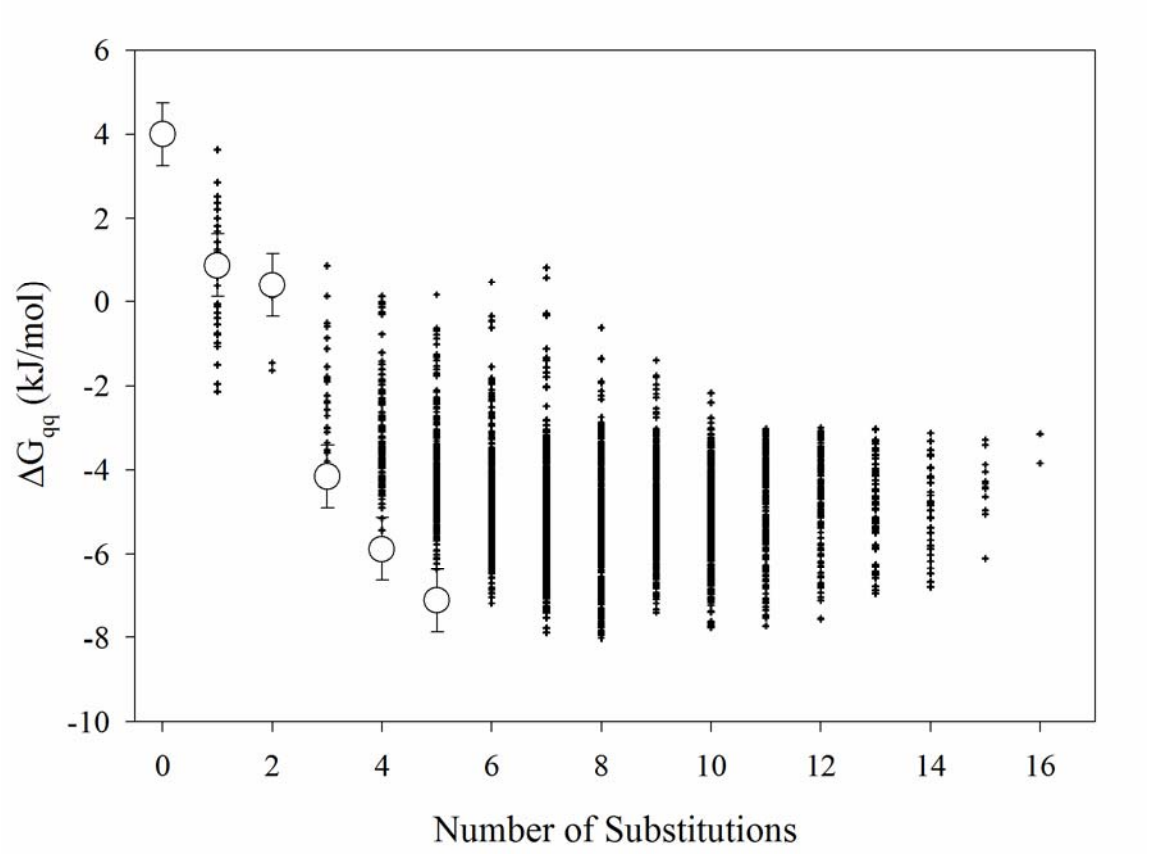

Figure 3.2 Evaluation of the effectiveness of the genetic algorithm to find favorable charge distributions at pH 7.0 with increased favorable charge-charge interaction energies relative to wild-type Fyn. The interactions energies are calculated by the TK-SA model. Each sequence, represented by black crosshairs, is characterized by the energy of charge-charge interactions and the number of substitutions relative to the wild-type protein. Note that the more favorable energies have smaller values of ΔG_{qq}. The designed (Fyn1-Fyn5) and wild-type sequences that were characterized experimentally are represented by white circles.

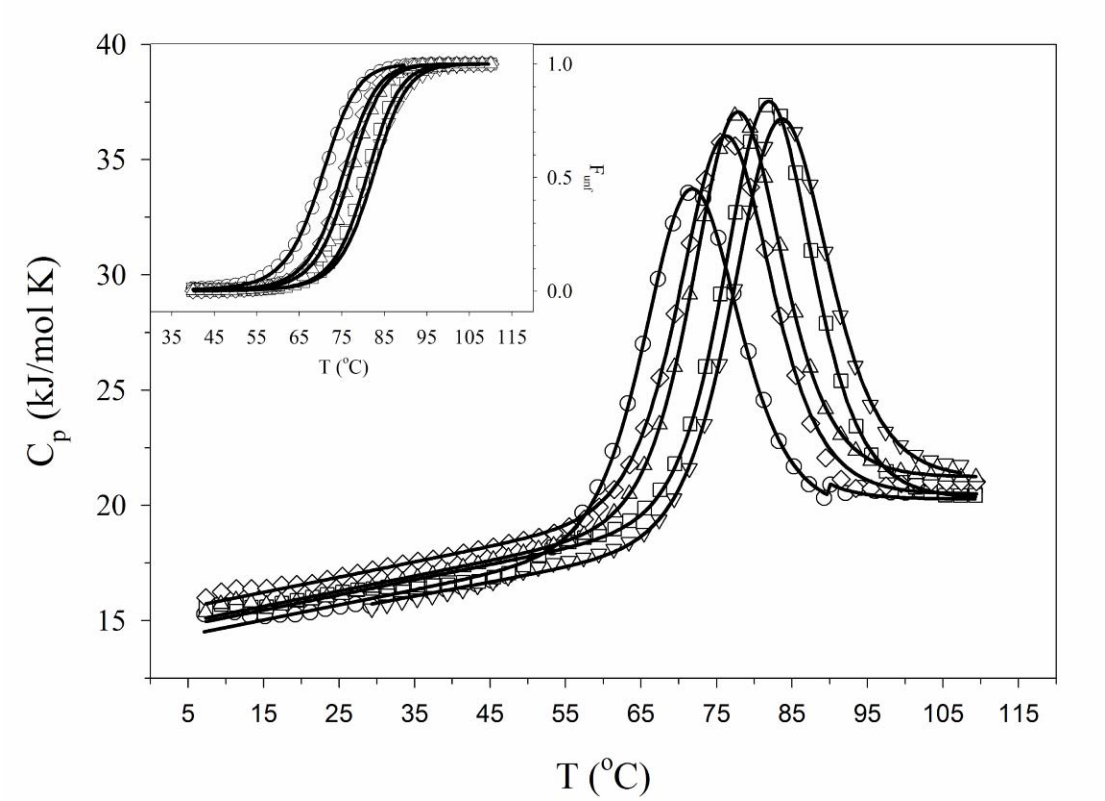

Figure 3.3 Comparison of stabilities of the Fyn variants. DSC profiles of Fyn variants at pH 7.0. The open symbols represent experimental data (circles - wild-type; triangles - Fyn1; diamonds - Fyn2; squares - Fyn3; inverted triangles - Fyn5). Only every fifth data point is shown, for clarity. The solid lines represent the global fit of the data to a two-state unfolding model. **Inset:** The fraction of unfolded protein (F_U) as a function of temperature for CD (symbols, same as above) and DSC (solid lines).

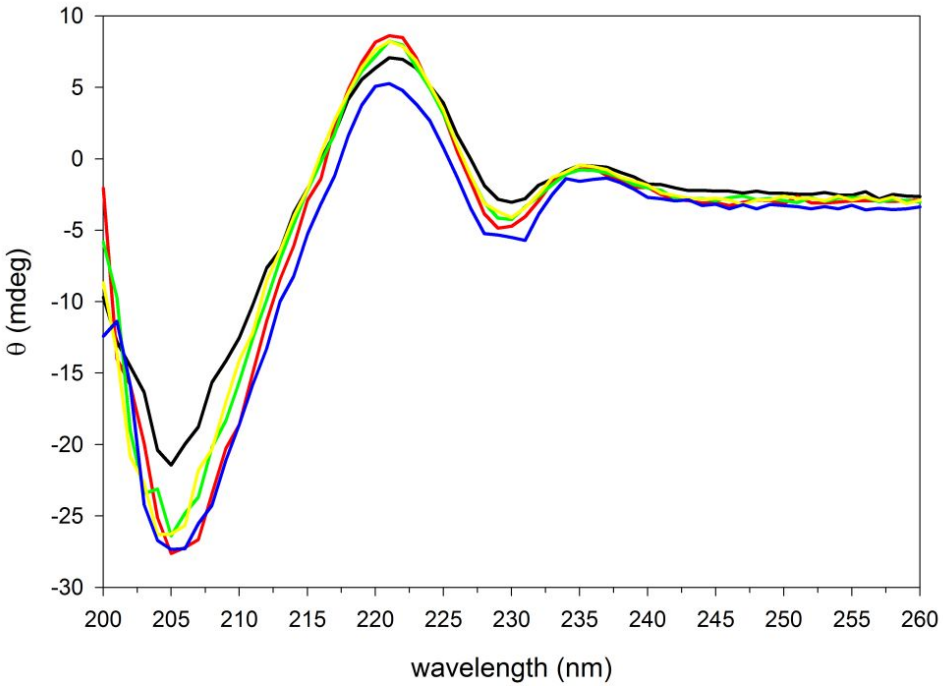

Figure 3.4 Far-UV (260-190 nm) Circular Dichroism Spectroscopy was used to determine whether the substitutions in the designed Fyn variants altered the secondary structure of the protein. The spectrum of each variant is represented in the following colors: black - WT Fyn, red - Fyn1, green - Fyn2, yellow - Fyn3, blue - Fyn5 (see text for construct nomenclature).

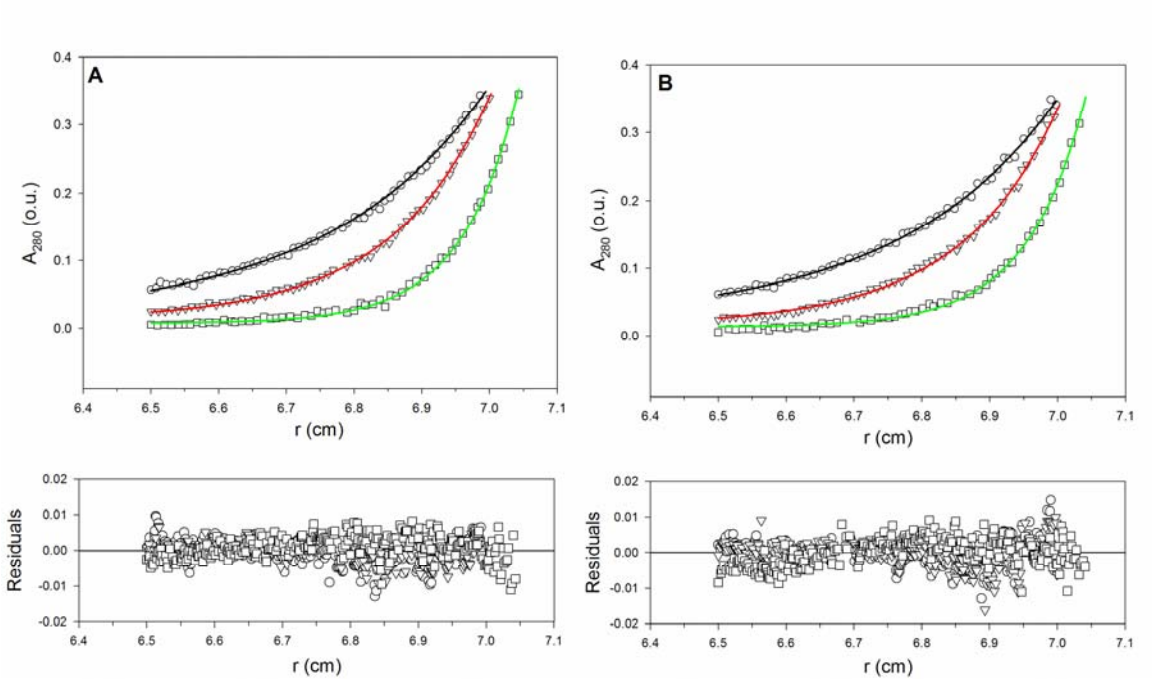

Figure 3.5 Results of Analytical Ultracentrifugation experiments for **A.** WT Fyn and **B.** Fyn5. The symbols represent the experimental data obtained at three speeds: circles - 22,000 rpm; inverted triangles - 28,000 rpm; squares - 37,000 rpm. The solid lines represent the global fit of the data to a single species model. The molecular weights measured by these experiments were 9.2 ± 0.5 kDa and 9.6 ± 0.5 kDa for WT Fyn and Fyn5, respectively, suggesting that the substitutions do not affect the oligomerization state of the protein.

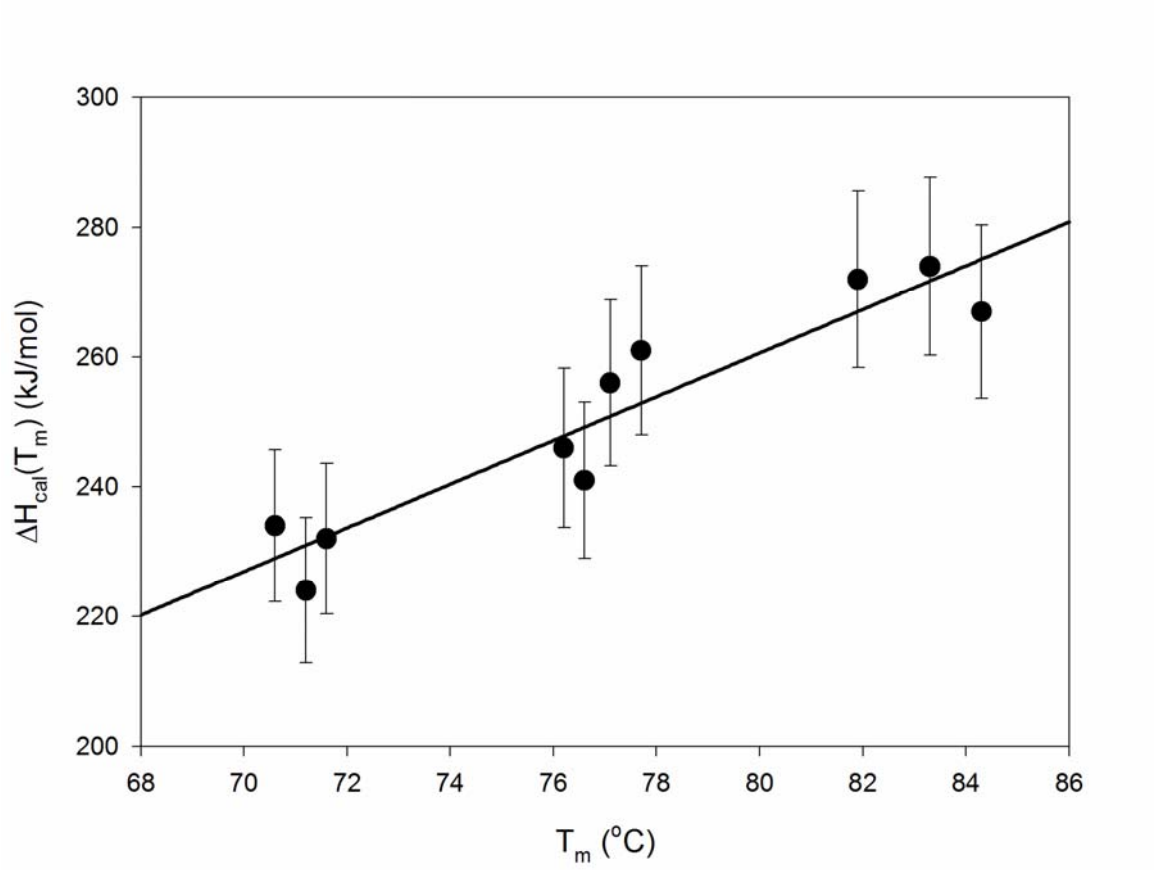

Figure 3.6 Dependence of the enthalpy of unfolding, $\Delta H(T_m)$, on the transition temperature, T_m, for the Fyn variants measured at pH 7.0. The error bars represent the estimated error of 5% for $\Delta H(T_m)$. The solid line is the linear regression of the data. The slope of this line corresponds to the heat capacity change upon unfolding, ΔC_P=3.4 ± 0.4 kJ/mol.

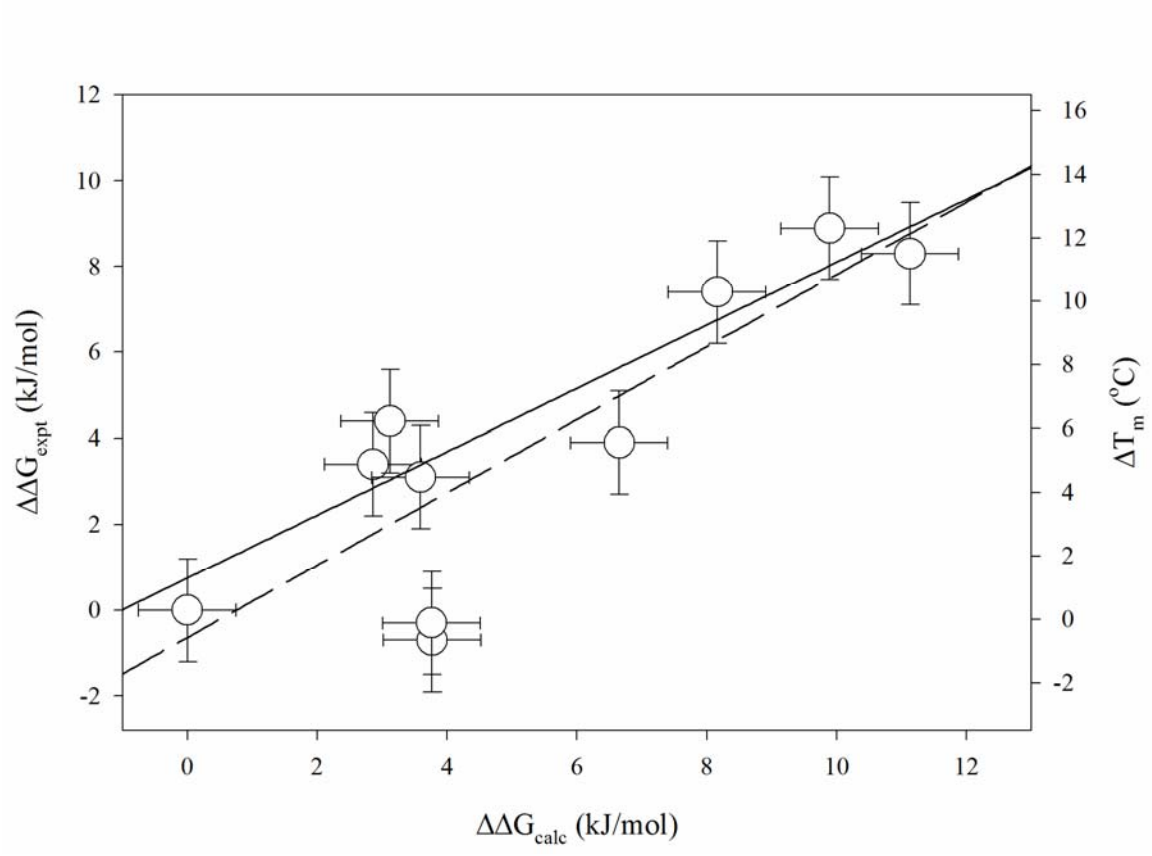

Figure 3.7 Comparison of experimentally measured changes in stability $\Delta\Delta G_{exp}$ or thermostability (ΔT_m) with those predicted by the TK-SA calculations, $\Delta\Delta G_{qq}$. The solid line represents the line of best fit disregarding the E11K and N30K substitutions and has a slope of 0.74 and a correlation coefficient of 0.88. The dashed line represents the line of best fit through all points and has a slope of 0.84 and a correlation coefficient of 0.71.

CHAPTER 4: DETERMINING THE IMPORTANCE OF RESIDUAL UNFOLDED STATE CHARGE-CHARGE INTERACTIONS FOR PROTEIN DESIGN STRATEGIES

4.1 Introduction

One aspect of the protein folding problem that remains poorly understood is the role of the unfolded state ensemble in protein stability. Besides contributing to our fundamental knowledge of how proteins fold into and maintain their three-dimensional structures, a comprehensive model of how residual unfolded state interactions contribute to the Gibbs free energy of unfolding (ΔG) could greatly improve the accuracy of computational design algorithms. One of the major assumptions of computational design methods is that there are no residual charge-charge interactions in the unfolded state. This assumption has not adversely affected the predictions of most methods, including the TK-SA algorithm, where it has been demonstrated that the relative changes in thermodynamic stability ($\Delta\Delta G = \Delta G_{DES} - \Delta G_{WT}$) of most proteins redesigned by the TK-SA approach were correctly predicted on a qualitative (sign of $\Delta\Delta G$) or semi-quantitative (relative rank order) level (*11, 15, 16, 97, 110, 187*).

There are, however, several reports of proteins for which such an assumption might not be valid. For a number of proteins, including the N-terminal ribosomal L9 protein (NTL9) (*88-90, 188, 189*), chymotrypsin inhibitor 2 (CI2) (*190*), barnase (*87*), ovomucoid third domain (OMTKY3) (*86*), RNaseA (*94*), RNase T1 (*94*), and hen egg white lysozyme (HEWL) (*138, 191, 192*), it was found that the extended linear chain or model compound representations of the unfolded state could not successfully predict the effects of pH on their thermodynamic stabilities (ΔG). However, when simple structural (*92*) or statistical (*93, 193-199*) models of the unfolded state were used, significant improvement in the agreement between the calculations and experimental data were observed. In fact, the existence of residual charge-charge interactions in the unfolded state of NTL9 was exploited to engineer a stable variant of this protein (*90*).

When it was observed that there was a set of CspB variants for which the TK-SA model did not correctly predict the effects of substitutions on the stability of the protein (*110*), we incorporated two different unfolded state models – the Gaussian polymer chain model developed by Zhou (*93*) and a structural model developed by Elcock (*92*) – to determine if residual unfolded state charge-charge interactions were responsible for the disagreement between the predicted and experimentally measured ΔG values. In the Gaussian polymer chain model, the unfolded state is represented by a Gaussian chain, where the distance between charges is distributed according to the probability function:

$$p(r) = 4\pi r^2 \left(\frac{3}{2\pi d^2} \right)^{3/2} \exp\left(-\frac{3r^2}{2d^d} \right) \quad \textbf{(4.1)}$$

where r is the distance between charges and d is the root-mean-squared distance, defined as $d = 7.5l^{1/2}+5.0$, where l is the number of peptide bonds separating the two residues (*93*). The charge-charge interaction energies between residues i and j are then calculated as:

$$W_{ij} = 332 \int_0^\infty p(r) \frac{\exp(-\kappa r)}{\varepsilon r} dr \quad \textbf{(4.2)}$$

where κ is a screening constant proportional to the ionic strength (I) of the solvent ($\kappa = I^{1/2}/3.04$ $Å^{-1}$ at room temperature; and ε is the dielectric constant of the solvent, taken to be 78.5 for water. When this model of the unfolded state was incorporated into the TK-SA algorithm, it was found that the discrepancy between the predicted stabilities of the CspB variants and their experimentally measured stabilities could not be explained by unfolded state interactions as described by the Gaussian polymer chain model (*110*).

The observation that the Gaussian model of the unfolded state did not significantly affect the TK-SA predictions of the relative stabilities of proteins raised the question of how well statistical models actually describe the unfolded state. One reason that the Gaussian model might not properly describe residual charge-charge interactions in the unfolded state is that the Gaussian

model assumes that the unfolded polypeptide chain behaves like an ideal polymer in a good solvent. The term "good solvent" refers to conditions where interactions between the polypeptide chain and the solvent are more favorable than interactions between the atoms comprising the polypeptide chain. While this might be an appropriate assumption when a protein is denatured using urea or guanidinium, it might not be appropriate for acid/base denaturation or thermal denaturation. Proteins fold because water is a "poor solvent" for the unfolded protein. In other words, the interactions between the polypeptide chain and the solvent are less favorable than the intrachain interactions. Therefore, the assumption that the unfolded state behaves as a chain in good solvent might not be an appropriate model for the unfolded state in aqueous environments (*200-202*). Furthermore, real polypeptide chains have constraints to the conformations they can assume due to the steric limitations imposed by the side chains, so the assumption that the unfolded state can be modeled by an ideal chain without such constraints might not be valid.

Structural models of the unfolded state, such as those described by Elcock (*92*), provide a possible solution to this problem by providing a structural basis for the contributions of charge-charge interactions in the unfolded state. This model of the unfolded state uses the native state structure of the protein as a starting point for the simulations. The van der Waals radii of each atom in the protein are artificially increased by 6 Å to essentially "explode" the protein molecule and remove most secondary structural elements and tertiary contacts(*92*). This process was performed in 1 Å increments, with an energy minimization of the new protein structure after each step. Since the 6 Å increase in van der Waals radii severely strains the structure of the protein, a final round of energy minimization is performed using the original van der Waals radii to allow bond lengths to return to appropriate values (*92*). However, this model was also unable to improve the TK-SA model predictions of the stabilities of the set of CspB variants (*110*). One possible explanation is that Elcock's model might also be inappropriate for describing the structure of the unfolded state ensemble. This model does not seem to remove turn structures,

and so therefore, might not provide a complete representation of the unfolded state, especially for proteins such as CspB which have primarily β-sheet secondary structural composition. Furthermore, only one unfolded structure is generated per starting structure from this approach. However, in reality, the unfolded state ensemble is comprised of many structures, and therefore, is likely to be much more amorphous.

It is possible that the inability of existing unfolded state models to improve the predictions of the TK-SA model is due to the fact that these models do not accurately describe the structures, and therefore the interactions in the unfolded state ensemble. To test this hypothesis, detailed structural representations of the unfolded state were created by generating libraries of 2,000 – 100,000 structures based on two different theoretical interpretations of the unfolded state – the random coil approach (RC) (*172*) and the excluded volume limit (EV) (*173*). Several different protein sequences of different lengths were used to characterize the behaviors of the libraries. The RC approach was used to generate unfolded state libraries of ubiquitin (16,000 structures), apomyoglobin (8,000 structures), and staphylococcal nuclease (5,000 structures). The EV limit was used to generate libraries of ubiquitin (2,000 structures & 100,000 structures), apomyoglobin (20,000 structures), and NTL9 (2,000 structures). The differences between the principles guiding the generation of an RC library versus EV limit library of unfolded state structures are discussed in Chapter 2.7. Briefly, in the RC approach, the backbone dihedral angles are assigned based on the probability of their occurrence outside regions of regular secondary structural elements, based on the Ramachandran plot. Once the entire backbone has been built, the side chains are added using homology modeling. The EV limit, on the other hand, starts with the entire polypeptide chain (including side chain atoms) in a random, sterically allowed conformation. The backbone and side chain dihedral angles are chosen such that they minimize an excluded volume energy function.

Once the libraries were generated, the polymeric behavior of these unfolded state ensembles were compared to each other and to the Gaussian chain model in order to determine whether the Gaussian model is accurately representing the distance distributions of "real" polymer chains. Since the interactions between charges are not considered in generating either the RC or the EV libraries of the unfolded state, molecular dynamics (MD) simulations were used to "turn on" these effects, and were run for each structure in the EV libraries of ubiquitin (UBQ) and NTL9. The polymeric behavior of the post-MD libraries were compared to their respective starting libraries and the Gaussian polymer chain model of the unfolded state. Finally, the charge-charge interaction energies for the EV library and the post-MD libraries were compared to those calculated by Zhou's Gaussian model (Eqs. 4.1 & 4.2) (*93*). The implications for the results of these calculations for protein design strategies are discussed.

4.2 Results & Discussion

4.2.1 Comparison of RC and EV structural libraries

One of the first questions we had regarding the RC and EV libraries was whether two fundamentally different approaches to generating structural representations of the unfolded state produced similar results in terms of pairwise distance distributions and charge-charge interaction energies. We also wanted to see how similar these parameters were to the Gaussian polymer chain (GPC) model of the unfolded state. Figure 4.1 shows the distance distributions for four pairs of charged residues in the RC and EV structural libraries of ubiquitin, as well as the GPC model. Interestingly, although the principles governing the generation of structural libraries by the RC or EV methods are very different, the resulting distance distributions for a given pair in the same protein are quite similar. The slight shift of the mean of the distribution to smaller distances for the RC library suggests that this method generates slightly more compact structures.

This is most likely due to the fact that the side chains are not modeled in great detail when building the backbone conformation of the protein, whereas in the EV method, the side chains are always present.

For pairs of residues that are close in sequence (Fig. 4.1A), the RC and EV distance distributions are clearly different from that of the GPC model. As the sequence separation between residues increases (Fig. 4.1B and Fig 4.1C), the distance distributions of the RC and EV structural libraries become more like the GPC distributions. At extremely large sequence separations (Fig. 4.1D), the distance distributions of the structural libraries behave like those of the GPC. This observation suggests that there is a characteristic length limit below which "real" unfolded state structures do not behave as ideal chains. In fact, for residues that are close in sequence, the identities of the residues in the pairs can have a significant effect on the shape and width of the distribution (Fig. 4.2A). However, after the pair of interest is separated by only 15 residues, the distance distributions were no longer dependent on the identities of the residues (Fig. 4.2B).

In order to determine whether the sequence of residues separating the charge-charge pair of interest would affect the distance distributions, the distance distributions for several K/E pairs in RC libraries of ubiquitin, apomyoglobin, and staphylococcal nuclease were also compared. Figure 4.3 demonstrates that the behavior of the distance distributions is largely independent of protein sequence, even for residues that are close together, confirming previous observations that the RC and EV libraries exhibit polymeric properties (*172, 173, 203*). These results also suggest that it should be possible to generalize the behavior of one protein sequence to describe the unfolded state ensembles of many proteins.

The observation that distance distributions for residues close in sequence in the RC and EV libraries are markedly different from those represented by the GPC highlights a potentially significant problem with using the GPC model to determine the effects of residual charge-charge

interactions in the unfolded state. Namely, the interaction energies between pairs of charges (E_{ij}) have the largest magnitude for pairs that are close together in sequence (Fig. 4.4), where the GPC does not accurately describe what is observed in the structural libraries. At separations where the EV and RC distance distributions begin to agree with those of the GPC, the interaction energies are negligible. It is possible that this discrepancy in distance distributions between GPC and the RC and EV models could have significant effects on the predicted magnitude of the charge-charge interactions in the unfolded state.

The observation that the EV and RC libraries had very similar properties in terms of their distance distributions implies that they should be similar in all other structural descriptors. For this reason, the remainder of this chapter will discuss only the differences between the EV library and the GPC model. In order to determine the effects of the different distance distributions on the unfolded state charge-charge interactions, the pairwise charge-charge interaction energies were calculated for the EV libraries and compared to those calculated using the GPC model. Figure 4.5 shows the total charge-charge interaction energies ($E_{int,unf}$) on a per-residue basis for each charged residue in ubiquitin calculated using the GPC and EV models of the unfolded state. Although the distance distributions for residues close in sequence are different for the GPC model versus the structural libraries, there seems to be very little effect on the $E_{int,unf}$ of each residue or of the entire protein. One explanation for this result is that the EV model does not account for the interactions between charges while building the unfolded state ensembles. It is possible that an unfolded state ensemble that includes a realistic representation of the interactions between charges in the structural libraries would yield better models for predicting the existence and effects of residual charge-charge interactions in the unfolded state.

4.2.2 Modeling charge-charge interactions using molecular dynamics (MD) simulations

Molecular dynamics (MD) simulations were used to "turn on" the interactions between charges. The size of an unfolded polypeptide chain precludes using explicit solvent models due to the limitations of currently available computational resources. For this reason, the MD simulations were run using the AMBER99SB force-field (*174*) and the generalized Born implicit solvent model, corrected for solvent accessibility (GB-SA) (*176*) available in the AMBER9 software package (*175*). Each step in a MD simulation represents the changes in atomic position (r_i) as a function of time by altering intramolecular interactions in a stepwise fashion. Trajectories are generated first by determining the force (F_i) on an atom based on the change in energy (E) between its current position and a position a small distance away. Knowing the force and mass (m_i) of an atom, it is possible to calculate the atomic acceleration (a_i):

$$F_i = -\frac{dE}{dr_i} = m_i a_i \qquad \textbf{(4.3)}$$

From the acceleration, it is then possible to obtain the atomic velocities (v_i) and positions at each step of the simulation:

$$\begin{aligned} v_i &= \int a_i dt \\ r_i &= \int v_i dt \end{aligned} \qquad \textbf{(4.4)}$$

The energy function $E(\boldsymbol{R})$ used in the AMBER99SB force field contains terms to account for bond energies; torsional energies, which consists of two terms (angles and dihedrals); van der Waals energies (Lennard-Jones 6-12 potential); and electrostatic energies (*175, 204*), and takes the form:

$$\begin{aligned} E(\boldsymbol{R}) = &\sum_{bonds} K_r \left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta \left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}\left(1 + \cos\left[n\phi - \gamma\right]\right) \\ &+ \sum_{i<j}^{atoms} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \sum_{i<j}^{atoms} \frac{q_i q_j}{\varepsilon R_{ij}} \end{aligned} \qquad \textbf{(4.5)}$$

where ***R*** is the set of Cartesian coordinates describing the positions of all atoms in the protein. The "bonds" term is comprised of K_r, which is an empirically determined bond force constant, r is the internal coordinates for a given bond, and r_{eq} is the empirically determined reference value for bonds. The "angles" term is comprised of the empricially determined bond angle force constant (K_θ), the bond angle (θ), and the bond angle reference value (θ_{eq}). The "dihedrals" term is comprised of V_n, which is a dihedral force constant; n, which is the dihedral periodicity; and γ, which is a phase of the dihedral angle (φ). The Lennard-Jones 6-12 potential is a function of the interatomic distances (R_{ij}) and the Lennard-Jones constants $A_{ij} = 4\epsilon\sigma^{12}$ and $B_{ij} = 4\epsilon\sigma^{6}$ which are functions of depth of the potential well (ϵ) and the finite distance where the potential is equal to zero (σ). The electrostatic energy term is a function of the charges on atoms i (q_i) and j (q_j); the dielectric constant (ε), taken to be 78.5, and the interatomic distances (R_{ij}).

To represent the solvation effects using the implicit GB-SA model (*176*), the following term is added to Eq. 4.5:

$$\sum_{ij}^{atoms} \frac{q_i q_j}{f^{gb}\left(R_{ij}\right)} + \sigma A \qquad \textbf{(4.6)}$$

where the polar contribution to solvation free energy is represented via the f^{gb} function, as previously described (*176*), and the second term represents the nonpolar contribution of the solvation free energy and is proportional to the solvent accessible surface area, A. To account for possible effects of long-range charge-charge interactions on the behavior of each structure, an essentially infinite cutoff distance (300 Å) was used in the electrostatic term of Eq. 4.1. The ionic strength of the solvent was set to 100 mM in the GB-SA term (Eq. 4.2) of all simulations.

We also tested the possible effects of pH on the behavior of the library by performing simulations under conditions that would model the extremes of pH. In the low-pH ("pH 2") simulations, the charges on all acidic residues, including the C-terminus, were neutralized. The neutral pH ("pH 7") simulations allowed all residues, except His, and including both termini to be

charged. All basic residues, including the N-terminus, were neutralized for the high-pH ("pH 14") simulations. MD simulations were performed on each structure in an EV library of 2,000 ubiquitin structures, until the population averaged radius of gyration ($<R_g>$) converged (300 ps, Fig 4.6A). To explore the possible effects of long-range charge-charge interactions on a protein believed to have residual structure in the unfolded state, MD simulations were also run on an EV library of 300 NTL9 structures until $<R_g>$ converged (300 ps, Fig. 4.6B). The calculations were run on the supercomputer available through the RPI Computational Center for Nanotechnology Innovations (CCNI). The overall simulation time for the 2,000 structure ubiquitin library was almost 1 μs and almost 100 ns for the 300 structure NTL9 library. These time scales are on the order of the time scales of formation of regular secondary structural elements such as loops (~ 50 ns), α-helices (~20 – 200 ps), and β-turns (~ 20 ns) (for a review, see (*205*)). Therefore, the simulations are long enough to detect whether residual structure due to charge-charge interactions can be formed in the unfolded state ensembles of these proteins.

Figure 4.6 shows how $<R_g>$ changes as a function of simulation time at pH 2, pH 7, and pH 14 for ubiquitin (Fig. 4.6A) and NTL9 (Fig. 4.6B). At the beginning of the simulations, $<R_g>$ decreases rapidly as a function of time as the unfolded state ensembles of both proteins respond to the folding conditions defined by the implicit solvent model. Eventually, the change in $<R_g>(T)$ begins to level off, such that after 200 ps, very little change in the population averaged radius of gyration is observed. The simulations were run for an additional 100 ps to ensure that $<R_g>$ was indeed converged. From intrinsic viscosity experiments on unfolded proteins, it is expected that the charge repulsions that are present at the extreme pH values will cause the unfolded state to become more expanded than the unfolded state at pH 7 (*206-210*). Although, the pH-dependent size difference is not very dramatic, this behavior is exactly what is observed for ubiquitin, which has a similar number of acidic and basic residues. Interestingly, the NTL9 unfolded state ensemble is most compact at pH 14 (Fig. 4.6B). This is most likely due to eliminating the charge

repulsions present at both pH 2 and pH 7 due to a series of Lys residues extremely close in sequence. Indeed, if the NTL9 structural ensemble is altered so that all charges are reversed (NTL9rev, Fig. 4.6B), the ensemble becomes more compact at pH 2 than at pH 7 or pH 14.

Contact maps for the post-MD structural libraries of ubiquitin and NTL9 were examined to determine whether the collapse of the unfolded state ensembles going from the good solvent conditions of the EV limit to the poor solvent conditions of an aqueous environment resulted in the formation of residual structure. Figure 4.7 compares the contact maps of the post-MD libraries with the EV libraries and native structures of each protein. To minimize noise in the contact maps of the unfolded state ensembles, the average distances between pairs of residues separated by less than four residues were not plotted. It can be seen from the contact maps that, although the native contacts of UBQ and NTL9 are very different, there is very little difference between proteins in the average pairwise distances of the EV and post-MD unfolded state ensembles. This observation suggests that the polymeric nature of the EV libraries is maintained in the collapsed ensembles present after 300 ps in an aqueous environment. Furthermore, there is no evidence of specific contacts being formed in the contact maps of the post-MD ensembles, demonstrating that the collapse of the ensembles upon exposure to poor solvent does not necessarily indicate the residual structure in the unfolded state.

4.2.3 Polymeric nature of the unfolded state under folding conditions

Although the collapse of the unfolded state ensemble in an aqueous environment does not appear to cause the formation of specific contacts, we wanted to explore how the polymeric properties of the post-MD ensembles compared to the EV limit ensembles. We compared the behavior between the EV limit libraries and the post-MD libraries of ubiquitin and NTL9 in terms of three structural descriptors – Kratky profiles, interchain scaling, and the correlation between solvent accessible surface area (ASA) and R_g (P(ASA,R_g)).

4.2.3.1 Kratky profiles

Figure 4.8 shows a comparison of the Kratky profiles for the ubiquitin and NTL9 chains in the EV limit, after the MD simulations, and in the native state. Kratky profiles, also known as scattering profiles, are powerful structural descriptors because they can be used to directly relate the results of simulations to the scattering measurements obtained with small-angle x-ray scattering (SAXS) experiments. These profiles essentially measure the density of the polypeptide chain across a specific length scale (*211*). In a scattering experiment, Kratky profiles plot scattering intensity I_s as a function of the scattering angle θ, which can be represented by the wave number, q ($q = \frac{4\pi n}{\lambda}\sin\left(\frac{\theta}{2}\right)$, where λ is the wavelength of incident light and n is the refractive index of the solution), in the form $q^2 I_s(q)$ vs. q. The unfolded state ensemble can be converted to scattering intensity by calculating the ensemble averaged form factor, $\langle P(q) \rangle$, which is directly proportional to $I_s(q)$:

$$K(q) = N_{res} q^2 \langle P(q) \rangle = N_{res} q^2 \left\langle \frac{2}{N_{at}(N_{at}-1)} \sum_{i=1}^{N_{at}} \sum_{j=i+1}^{N_{at}} \frac{\sin(qr_{ij})}{qr_{ij}} \right\rangle \qquad \textbf{(4.3)}$$

where N_{res} is the number of residues in the sequence; q are wave numbers in units of $Å^{-1}$; N_{at} is the number of atoms in the protein; and r_{ij} represents the pairwise distances between atoms. A natively folded protein will have large peaks at low to mid-q values ($0 \leq q \leq 0.5$), indicating densely packed, compact structures (*211*). The native state profiles of ubiquitin and NTL9 are slightly different in this regime, with ubiquitin having a slightly larger peak at low-q than NTL9, suggesting that ubiquitin is more tightly packed. The native state profiles also have small differences at high-q values ($q > 0.5$), indicating the local structural propensities of these proteins are different in the native state (*203*). If a polypeptide chain is loosely structured, one would expect the Kratky profile to have a flatter dependence of *K(q)* in the low- to mid-q range. Indeed,

this is what is observed for both ubiquitin and NTL9 in the EV limit. Interestingly, the EV limit profiles of ubiquitin and NTL9 are the same over the entire *q* range, providing further support to the observations discussed in the previous sections that the unfolded states of different proteins behave similarly in the EV limit. The small peak in the low *q* range of the Kratky profiles of ubiquitin and NTL9 after 300ps in an aqueous environment indicate that the polypeptide chain has collapsed and become more densely packed than in the EV limit, though not to the extent of the native state. The profile in the high *q* range also changed relative to the EV limit profiles, indicating a change in local structural propensity of the ensembles. However, the behavior of the UBQ and NTL9 post-MD libraries were still more similar to each other than the UBQ and NTL9 native state profiles. The observation that the post-MD profiles of UBQ and NTL9 are also remarkably similar to each other provides further evidence that the unfolded state ensembles of proteins in the initial stages of folding are also sequence independent, and therefore, little or no residual structure is present in the unfolded state ensembles of these proteins under folding conditions.

4.2.3.2 Interchain scaling

Another property of polypeptide chains in the EV limit is that the correlations of several properties are scale invariant (*173, 211*). For example, the ensemble averaged interchain distances between two residues, $\langle R_{ij} \rangle$, will scale like the average end-to-end distances of an ensemble of polypeptide chains with different lengths. In other words, the interchain distances should behave as $\left\langle R_{ij} \right\rangle \propto \left| i - j \right|^{\upsilon}$, where |*i*-*j*| is the number of residues separating residues *i* and *j*, and $\upsilon \approx 0.59$ for polypeptides in the EV limit. This correlation only holds for pairs that are separated by more than 7 residues, which is the length scale where the specific amino acids of residues *i* and *j* becomes important (*173, 203*). If the post-MD libraries still behave as

polypeptide chains in a good solvent, then one would expect a plot of $\ln(R_{ij})$ vs. $\ln(|i-j|)$ to have a slope of ≈ 0.59 (*173*). Figure 4.9 shows the interchain scaling behavior of the ubiquitin and NTL9 in the EV limit and after 300ps in folding conditions. It can be seen that the interchain scaling behavior of the polypeptide chains becomes more similar to what would be expected for a polymer in poor solvent ($\upsilon \approx 0.33$), which is consistent with the idea that water is a poor solvent for the unfolded state of proteins. Interestingly, the same interchain scaling behavior of the UBQ and NTL9 libraries that exists in the EV limit ensembles is retained under folding conditions, confirming the notion that the specific protein sequence is not important at larger separation distances.

4.2.3.3 Correlation between *ASA* and R_g

Figure 4.10 shows the correlation between solvent accessible surface area (*ASA*) and R_g for ubiquitin and NTL9 in the EV limit, after 300ps under folding conditions, and in the native state. For both proteins in the EV limit, there is very little correlation between ASA and R_g. The R_g distribution is broader than the ASA distribution over the population, demonstrating that the size fluctuations of unfolded proteins in the EV limit do not correlate with burial of surface area, which is consistent with the idea that the EV limit models polymers in a good solvent. At the other extreme, the native state of each protein shows a narrower distribution around R_g, such that small changes in R_g can result in fairly large changes in *ASA*. The narrower distribution around R_g demonstrates the limited conformational flexibility of proteins in the folded state compared with those in the unfolded state. After 300 ps under folding conditions, the collapse of the ubiquitin and NTL9 populations starts to shift the *ASA* vs. R_g correlations more toward those of the natively folded proteins. The size of the chains, as described by R_g decreases and the distribution becomes narrower than in the EV limit. However, the ASA distribution is much broader than either the EV limit or the native state, such that small changes in R_g cause large

changes in *ASA*. The changes in the three structural parameters described here corroborate the observations of the previous section that the polypeptide chain collapses after 300ps under folding conditions. Furthermore, the collapse causes the polypeptide chain to behave as a polymer in a poor solvent, which is consistent with the idea that water is a poor solvent for the unfolded state of proteins.

4.2.4 Effects of collapse on unfolded state charge-charge interactions

Since the unfolded state ensembles of ubiquitin and NTL9 are more collapsed under folding conditions than in the EV limit, and since their Kratky profiles (Fig. 4.8), interchain scaling (Fig. 4.9), and P(ASA,R_g) (Fig. 4.10) behaviors are dramatically different between the EV limit, the folding conditions, and the Gaussian polymer chain, we wanted to determine if there would be any dramatic differences between the unfolded state charge-charge interactions calculated using each model of the unfolded state. First we examined how the electrostatic contact maps for ubiquitin (Fig. 4.11) and NTL9 (Fig. 4.12) at pH 2, pH 7, and pH 14 changed due to the MD simulations. A comparison of the charge-charge contacts predicted by the GPC and EV limit will make it possible to determine if any specific charge-charge interactions contacts arise as a result of placing the unfolded polypeptide chain under folding conditions. The pH 2 and pH 14 contact maps are primarily made up of unfavorable charge-charge interactions for both the Gaussian chain and the post-MD chains, which is to be expected since only like charges are present under these conditions. The pH 7 contact maps contain both favorable and unfavorable interactions for both proteins. Interestingly, there are no significant differences in the pairwise charge-charge interactions among these three models of the unfolded state. Essentially all of the observed charge-charge interactions of a given residue are with its nearest neighbors (less than four residues apart). Furthermore, the consistent decrease in the magnitude of the interaction energy as the separation between residue pairs increases suggests that there are no specific

residual charge-charge interactions present in these unfolded state models. An examination of the total charge-charge interaction energies of each charged residue with every other residue in the protein ($E_{int,unf}$, Fig. 4.5) demonstrates that there is also very little difference in the contribution of each residue to the total charge-charge interactions in the unfolded state between the Gaussian model, the EV limit or the post-MD structural libraries. These results suggest that the charge-charge interactions in the unfolded state are too weak to be the force responsible for organizing residual structure. If any structure in the unfolded state does in fact exist, then residual charge-charge interactions are the consequence, not the cause, of such structure. The results further imply that statistical models, like the Gaussian polymer chain, are as good as more detailed structural models for calculating the charge-charge interactions in the unfolded state.

If the Gaussian polymer chain model is sufficient for predicting the effects of charge-charge interactions in the unfolded state, then why does it not significantly improve the predictions of the TK-SA model? One explanation is that in our design approach, we seek to determine the difference in stabilities between the designed variant and the wild-type protein ($\Delta\Delta G_{qq} = \Delta G_{VAR} - \Delta G_{WT}$) rather than the absolute stability of the designed variant (ΔG_{VAR}). The effect of substitutions on charge-charge interactions in the unfolded state is likely to be much smaller than the magnitude of the same contributions in the native sate of proteins. For example, the difference in $E_{int,Gauss}$ between sequences as different as UBQ and NTL9 is less than 2 kJ/mol, regardless of pH (Fig. 4.5), which is on the order of the error in experimental measurements. This observation suggests that changes in the unfolded state charge-charge interaction energies between a wild-type protein and a designed variant with only a few substitutions will be negligible. Therefore, while consideration of the unfolded state might be important for predictions of the absolute ΔG of a protein, for the purposes of protein design, where we are more interested in predicting $\Delta\Delta G$, it is not necessary to consider the unfolded state.

4.3 Future Directions

Although the results of this study suggest that detailed consideration of the unfolded state is not important for protein design approaches, there are still many characteristics of the unfolded state that can be further studied with the existing ensembles. One of the most interesting aspects to pursue would be the mechanism of collapse of the unfolded state in the folding process. Our short, 300 ps simulations provide an excellent starting point for studying these effects. It is unlikely that all structural descriptors of the unfolded state ensemble will respond to exposure to folding conditions on the same time scale. We can get an idea of how the polymer is changing by analyzing the full 300 ps trajectory of the ubiquitin and NTL9 MD simulations (which were saved in 1 – 5 ps snapshots) in terms of a variety of descriptors, such as end-to-end distance distributions, bond vector correlations, Kratky profiles, interchain scaling, and P(*ASA*, R_g), as a function of time. A large scale analysis such as this will make it possible to determine the rates at which different features of an unfolded polypeptide chain respond to folding conditions. In addition, it has been experimentally determined that the time scale of full hydrophobic collapse is on the order of 4 – 50 μs (*205*), so it might eventually be necessary to run longer simulations to get a full structural description of the mechanism of hydrophobic collapse, but further analysis of the simulations performed here will allow us to see how structural descriptors begin to change once an unfolded polypeptide chain is exposed to folding conditions. MD simulations of this nature are essential tools for understanding mechanisms of protein folding by identifying states in the folding pathway that are not easily accessible by current experimental techniques.

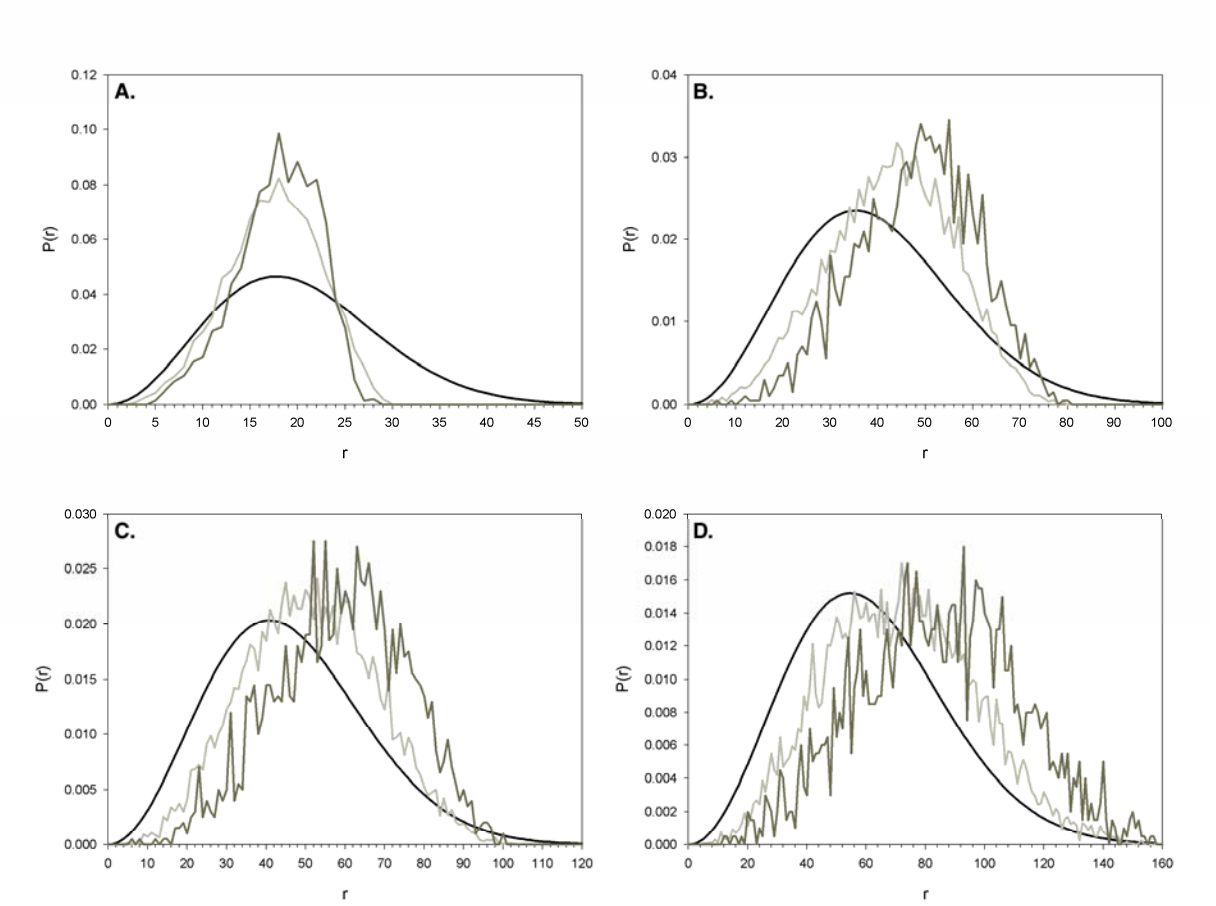

Figure 4.1 Comparison of distance distributions of three different representations of the unfolded state of ubiquitin. In all panels of the figure, the black line represents the Gaussian polymer chain (GPC) model, light grey lines represent 16,000 structure RC library, and dark grey lines are the 2,000 structure EV library. **A.** K6-K11 (| i – j | = 5) **B.** K6-D32 (| i – j | = 26) **C.** K6-R42 (| i – j | = 36) **D.** K6-R74 (| i – j | = 68). For residue pairs close in sequence, the distance distributions are not accurately described by the Gaussian polymer chain model. However, as the sequence separation between the pairs increases, the Gaussian model can more accurately describe the distance distributions of real polymer chains.

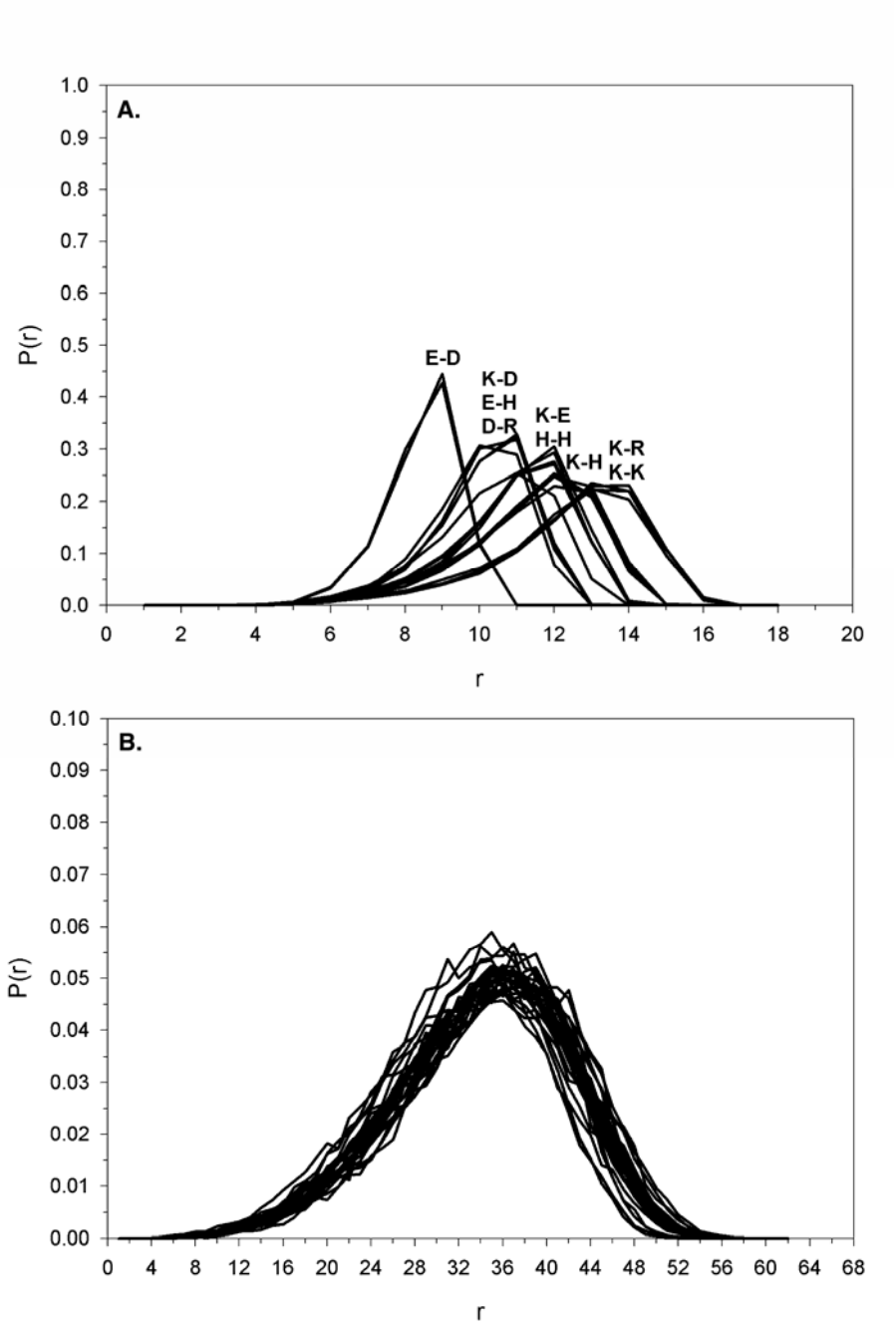

Figure 4.2 Effect of sequence separation on distance distributions for several types of interacting pairs in ubiquitin and apomyoglobin. **A.** When residues are very close in sequence (| i – j | = 1), the identity of the interacting pair significantly effects the mean and shape of the distance distribution. **B.** When residues are further apart (| i – j | = 15), the identity of the pair becomes less important.

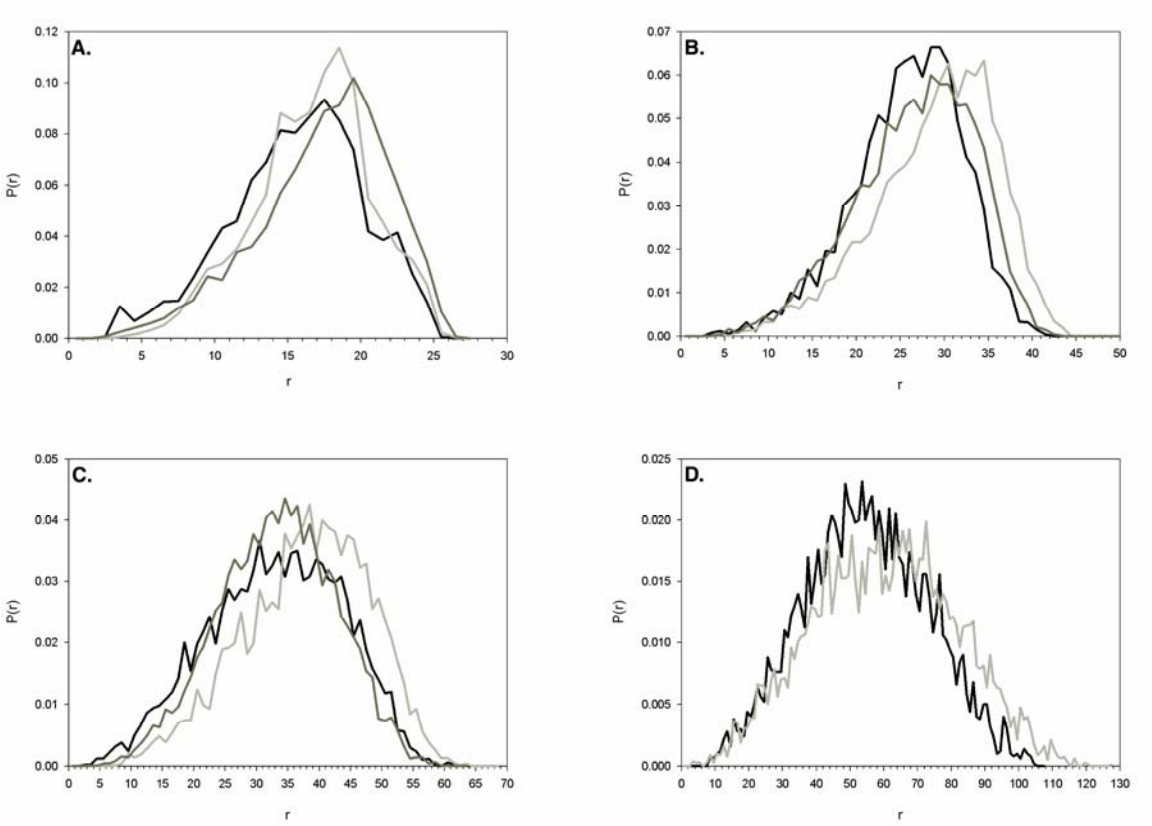

Figure 4.3 Comparison of distance distributions for K/E pairs in three different proteins. In all panels of the figure, the black lines represent staphylococcal nuclease, the light grey lines represent apomyoglobin, and dark grey lines represent ubiquitin. **A.** | i – j | = 5; **B.** | i – j | = 11; **C.** | i – j | = 18; **D.** | i – j | = 43. At all ranges of sequence separation, the distance distributions for K/E pairs are remarkably similar, regardless of the specific sequence of the protein.

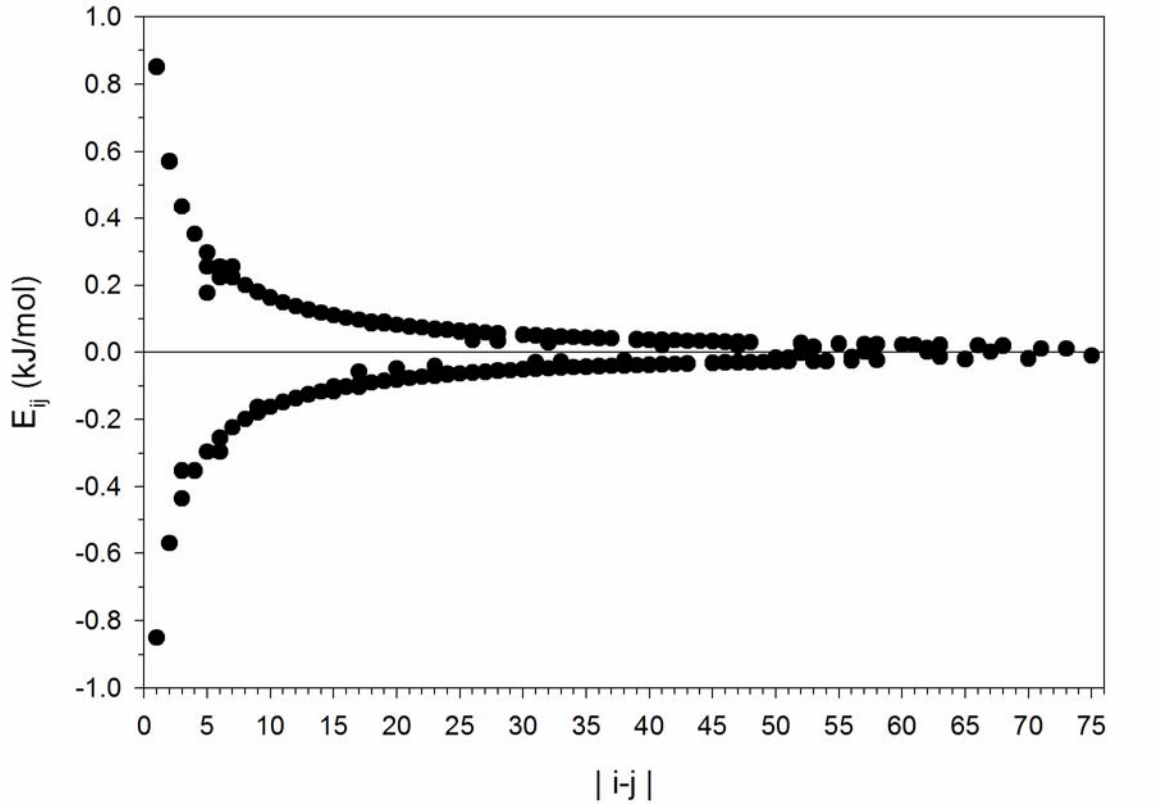

Figure 4.4 Energy of ubiquitin unfolded state charge-charge interactions as a function of sequence separation, calculated using the Gaussian polymer chain model of the unfolded state (*93*). The contribution of unfolded state interactions is largest for residues that are close in sequence and becomes negligible for pairs that are further than 20-25 residues apart.

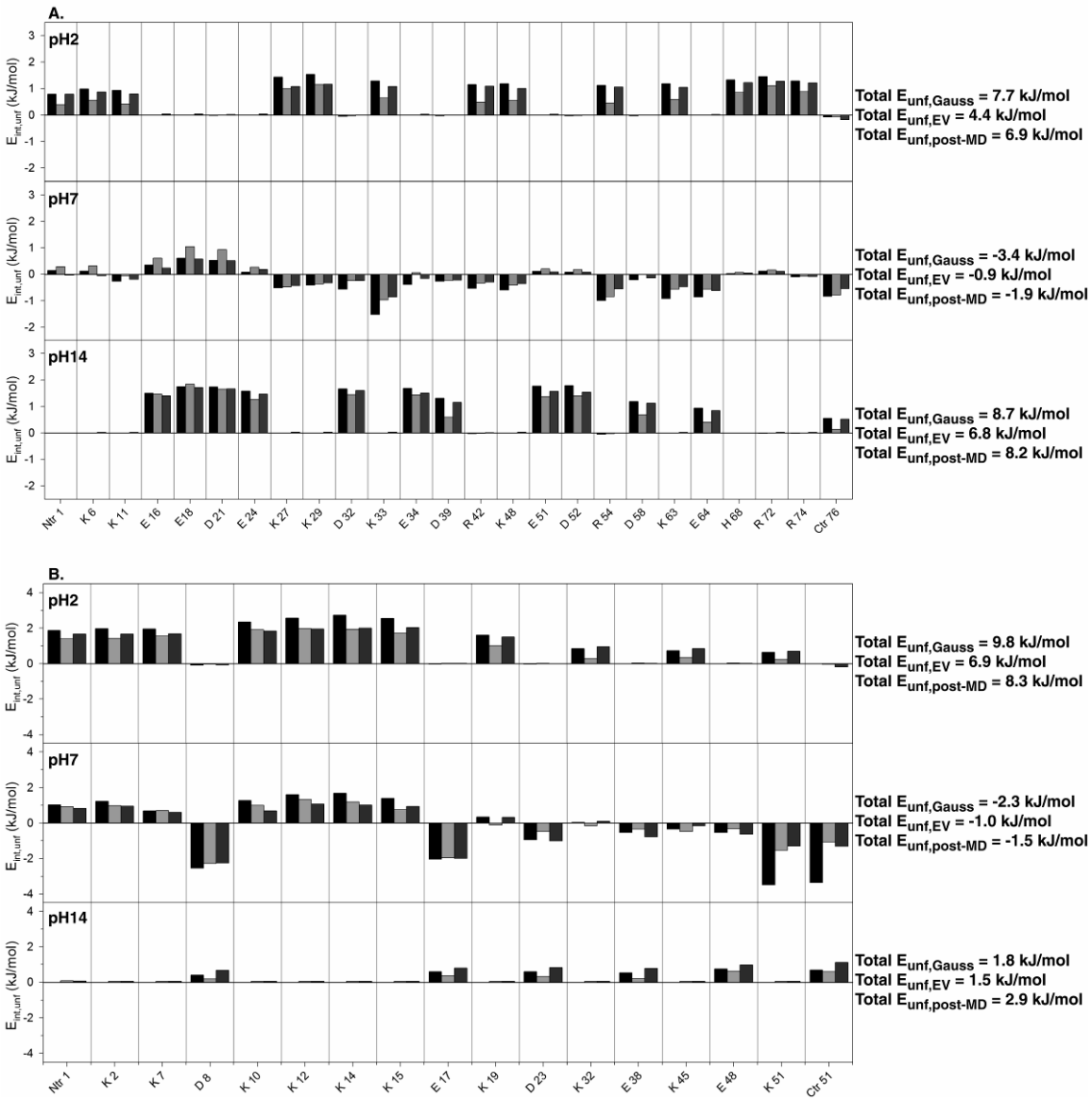

Figure 4.5 Total unfolded state charge-charge interaction energies per residue for **A.** ubiquitin and **B.** NTL9 at pH 2, pH 7, and pH 14. Each bar represents the sum of the interaction energy of residue *X* with every other residue in the protein. Black bars represent the Gaussian model. Light grey bars are the energies calculated using the EV library, and are averaged over 2,000 structures for both ubiquitin and NTL9. Dark grey bars represent the energies calculated on the final structural libraries after 300ps of MD simulation, and are averaged over 2,000 structures for ubiquitin or 300 structures for NTL9. Positive values of $E_{int,unf}$ indicate unfavorable interactions, while negative values represent favorable interactions.

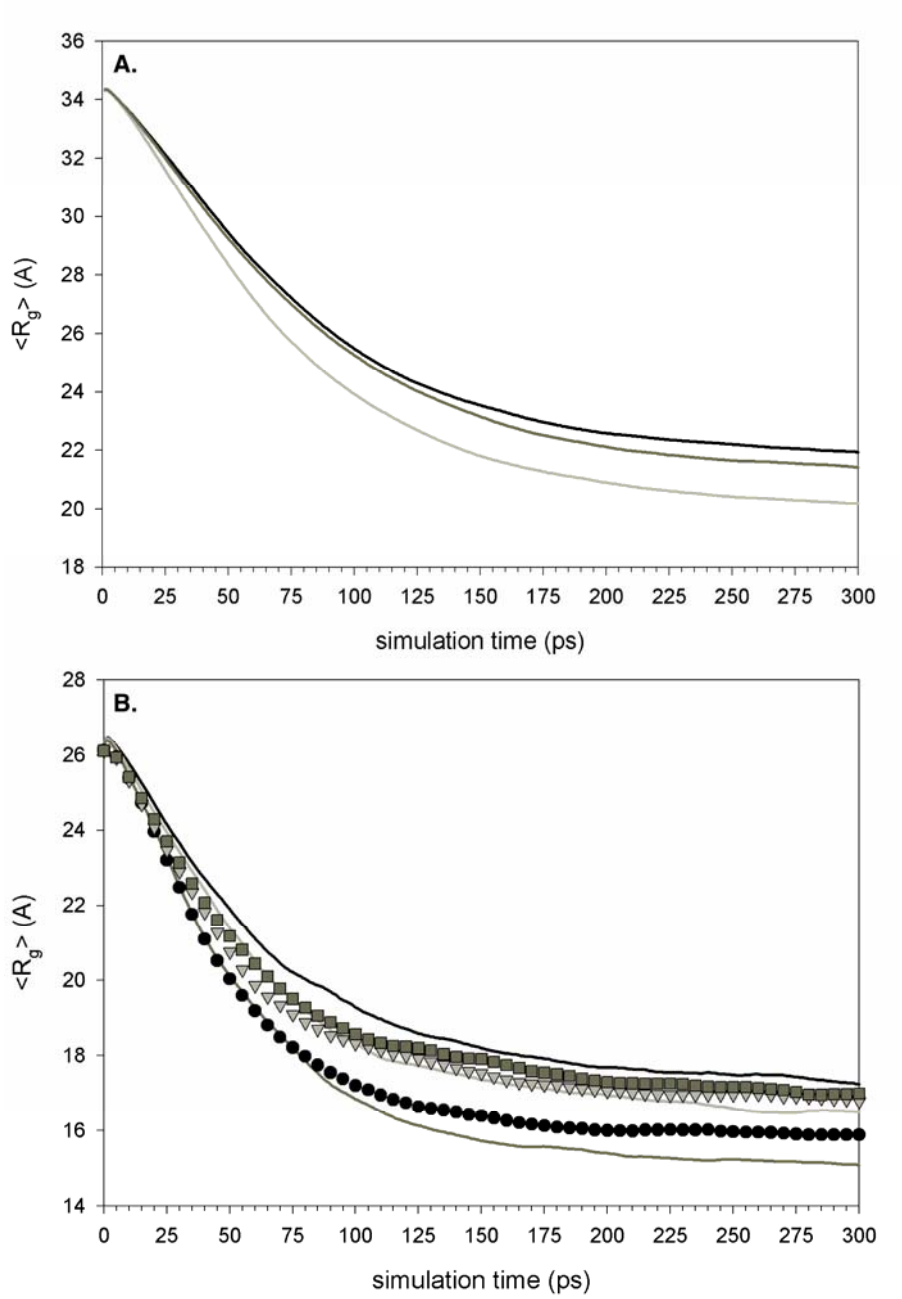

Figure 4.6 Population averaged radius of gyration (<R_g>) as a function of simulation time. **A.** Ubiquitin simulations: black line – pH2; light grey line – pH7; dark grey line – pH14. **B.** NTL9: black line – pH2; light grey line – pH7; dark grey line – pH14; black circles – NTL9rev pH2; light grey triangles – NTL9rev pH7; dark grey squares – NTL9rev pH14. The <R_g> changes rapidly within the first few picoseconds (ps) of simulation. After 200ps, very little change is observed in the average R_g of the population. Furthermore, the pH-dependent behavior seems to correlate with the relative numbers of acidic and basic residues in each protein.

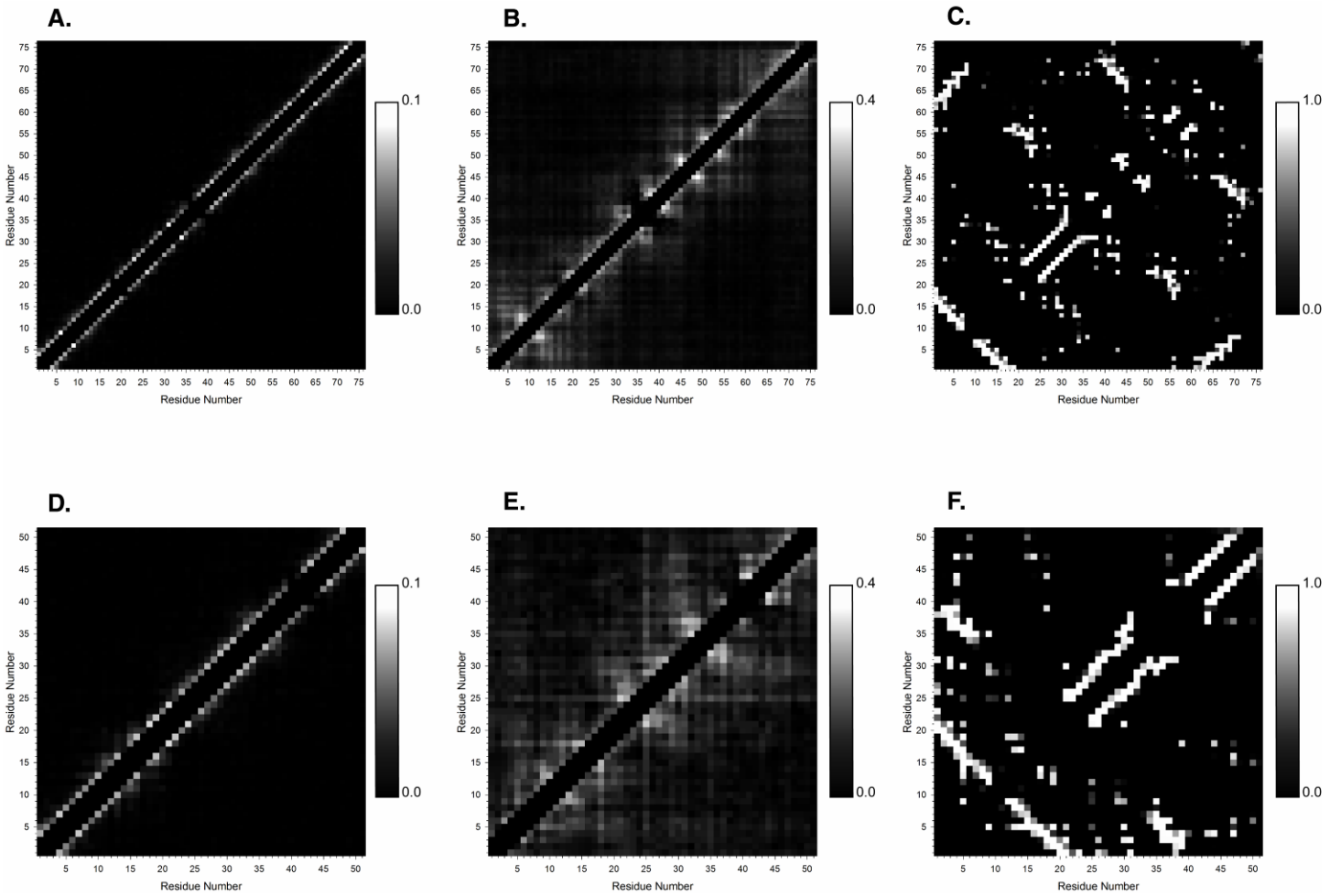

Figure 4.7 Contact maps for ubiquitin (**A**-**C**) and NTL9 (**D**-**F**) in the EV limit (**A** & **D**), after 300 ps MD simulation (**B** & **E**), and in the native state (**C** & **F**). The color of each box represents the probability that the residues are within 3.5 Å of one another. Darker boxes represent lower probabilities of contact than whiter boxes. Although more contacts are made in the post-MD population than in the EV limit, there is no evidence of the ordered contacts observed in the native state of either protein.

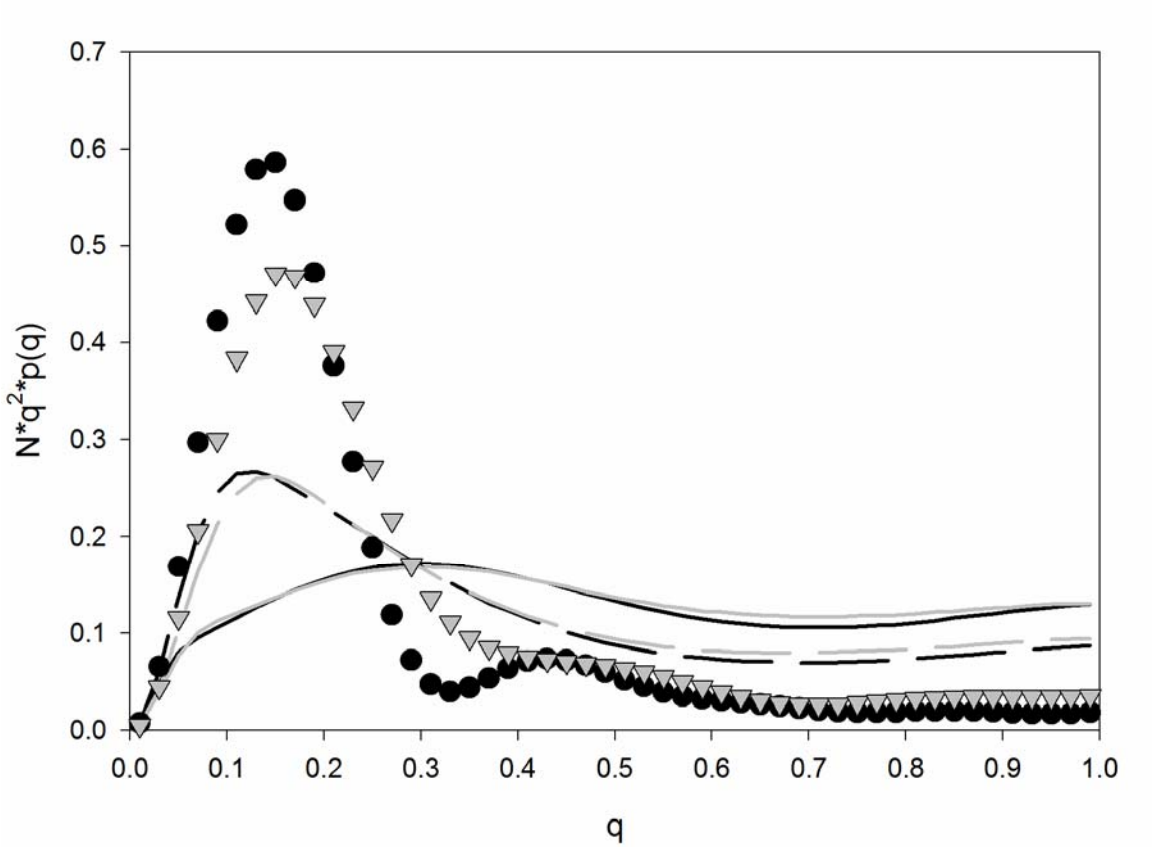

Figure 4.8 Kratky profiles for ubiquitin (black lines and circles) and NTL9 (light grey lines and triangles) in the EV limit (solid lines), after 300 ps MD simulation (dashed lines), and in the native state (symbols). The small peak at low q values of the post-MD populations of both ubiquitin and NTL9 is indicative of a structure that is more collapsed and densely packed than the chains in the EV limit, but not to the extent of the native proteins.

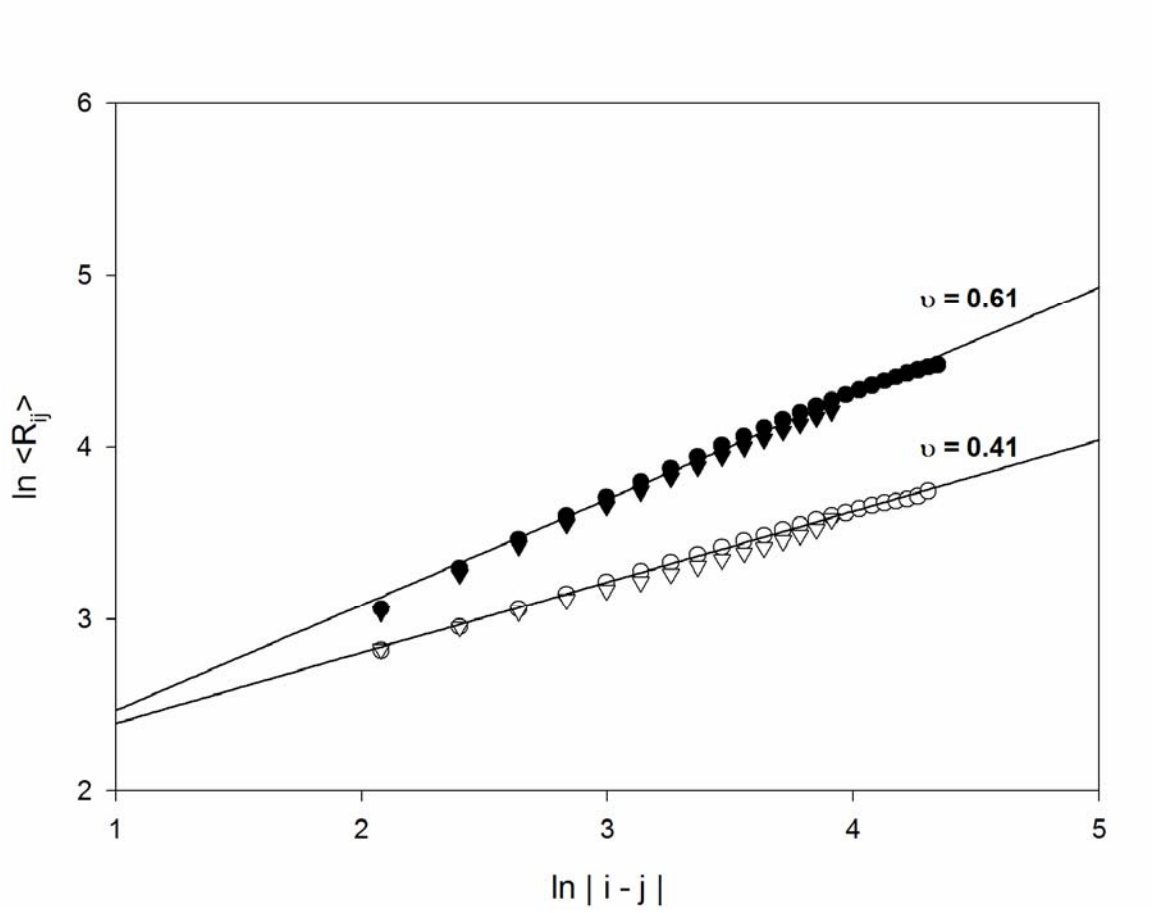

Figure 4.9 Log-log plot of pairwise interchain distances (R_{ij}) as a function of their sequence separation (| i – j|) for ubiquitin (●) and NTL9 (▼) in the EV limit (●,▼) and after 300 ps MD simulation (○,▽). The EV limit models a polymer in a good solvent, so the slope of this line (υ) is expected to be around 0.59. After 300ps of MD simulation in an aqueous environment, the chains start to adopt scaling indicative of a polymer in a poor solvent, which would have an expected scaling of $\upsilon \approx 0.33$.

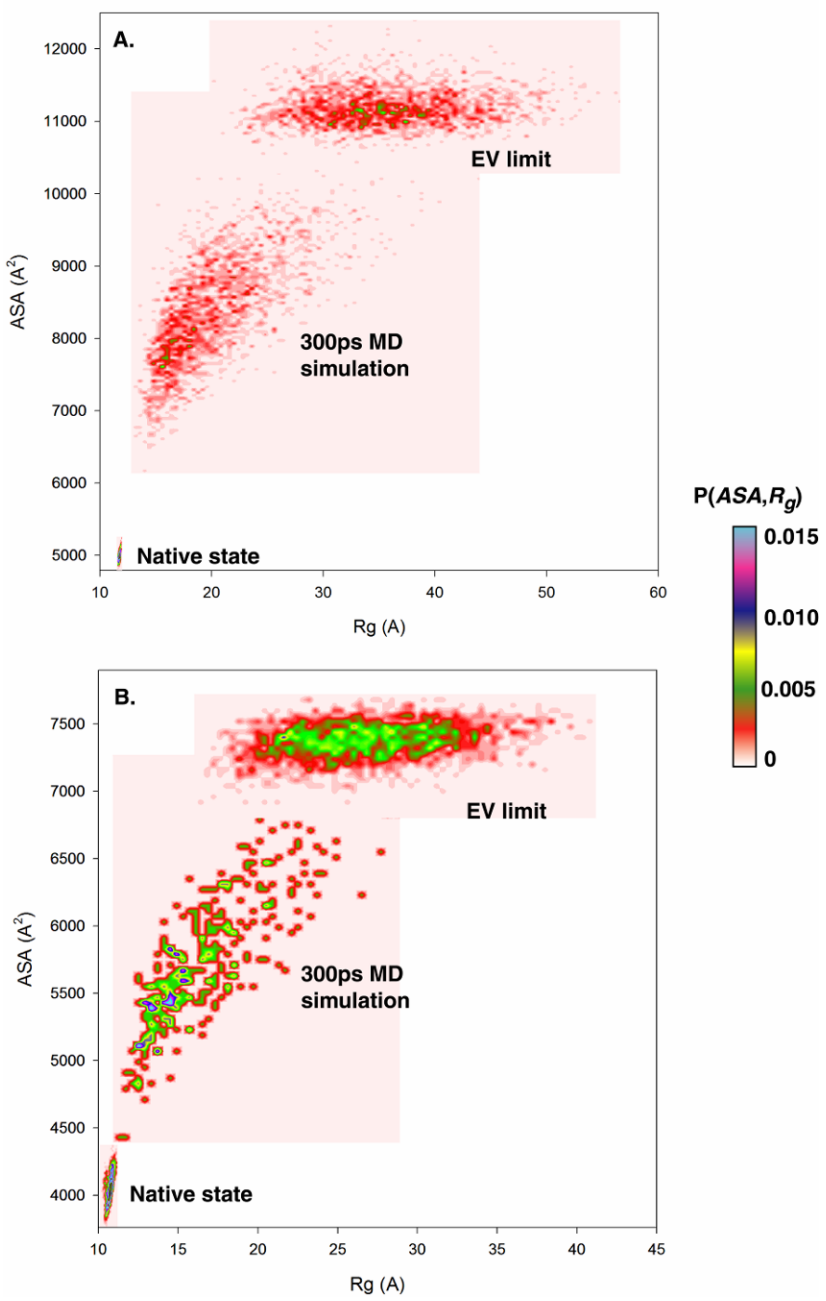

Figure 4.10 Contour plots of solvent accessible surface area (*ASA*) vs. radius of gyration (R_g) for **A.** ubiquitin and **B.** NTL9 in the EV limit, after 300 ps MD simulation, and in the native state. In both panels, the EV limit libraries display a broad distribution around R_g, such that the unfolded polypeptide chains maintain a similar *ASA* regardless of the size of the chain. In contrast, in the native states of ubiquitin and NTL9, small fluctuations in R_g will lead to relatively large changes in ASA. After 300 ps in an aqueous environment, a correlation between *ASA* and R_g begins to develop as the chain collapses toward the R_g of the native state.

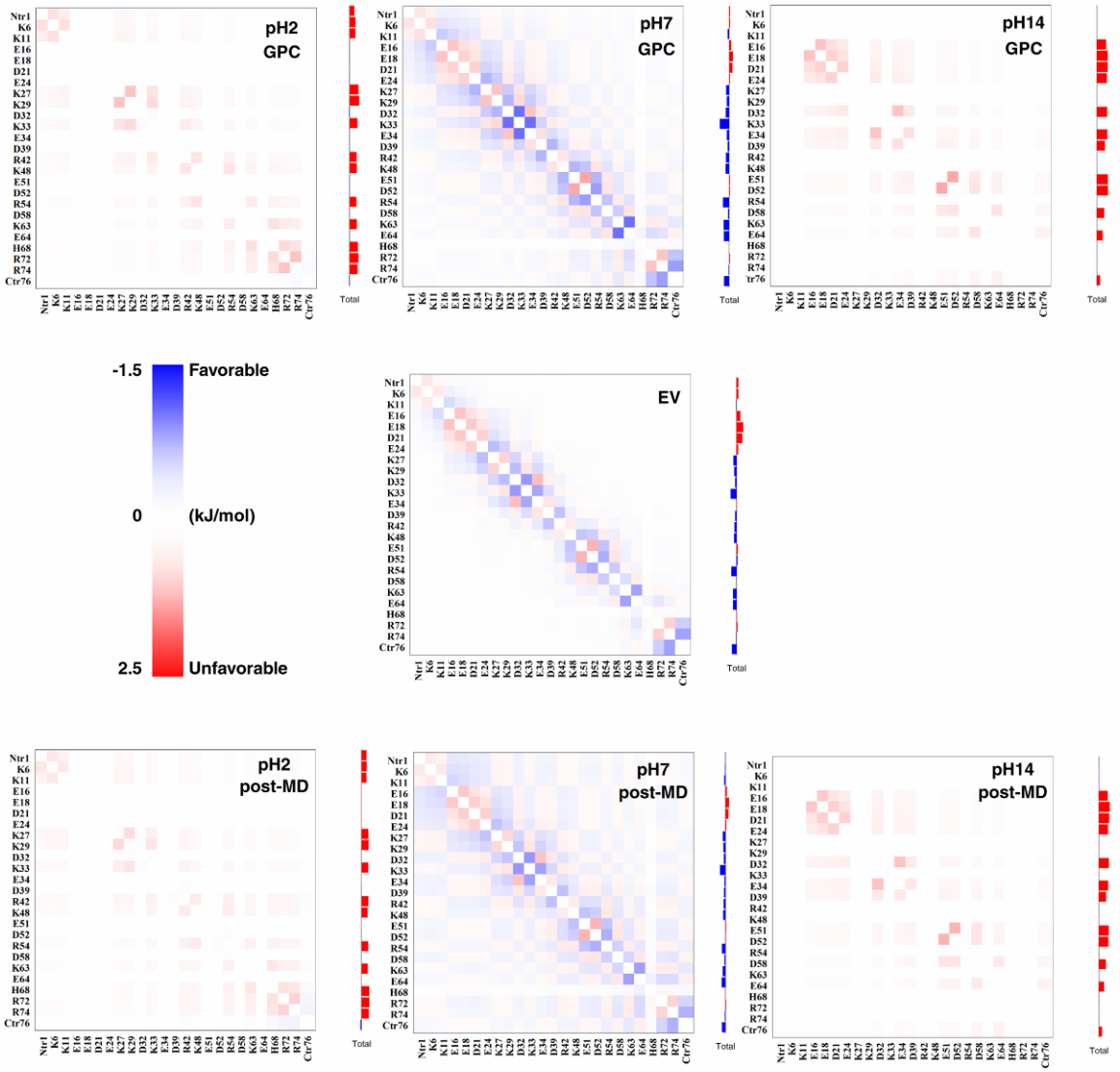

Figure 4.11 pH-dependent electrostatic "contact" maps for three unfolded state models of ubiquitin. The squares represent the pairwise charge-charge interactions in the unfolded state. Red squares indicate unfavorable interactions, while blue squares are favorable interactions. The bar graph to the side of each contact map represents the total interaction energy of a given residue with every other charged residue in the protein. There are no significant differences between interaction energies predicted by the Gaussian model (GPC), the EV limit, or the post-MD structures at any pH. There are also no indications of significant residual charge-charge interactions for any residue.

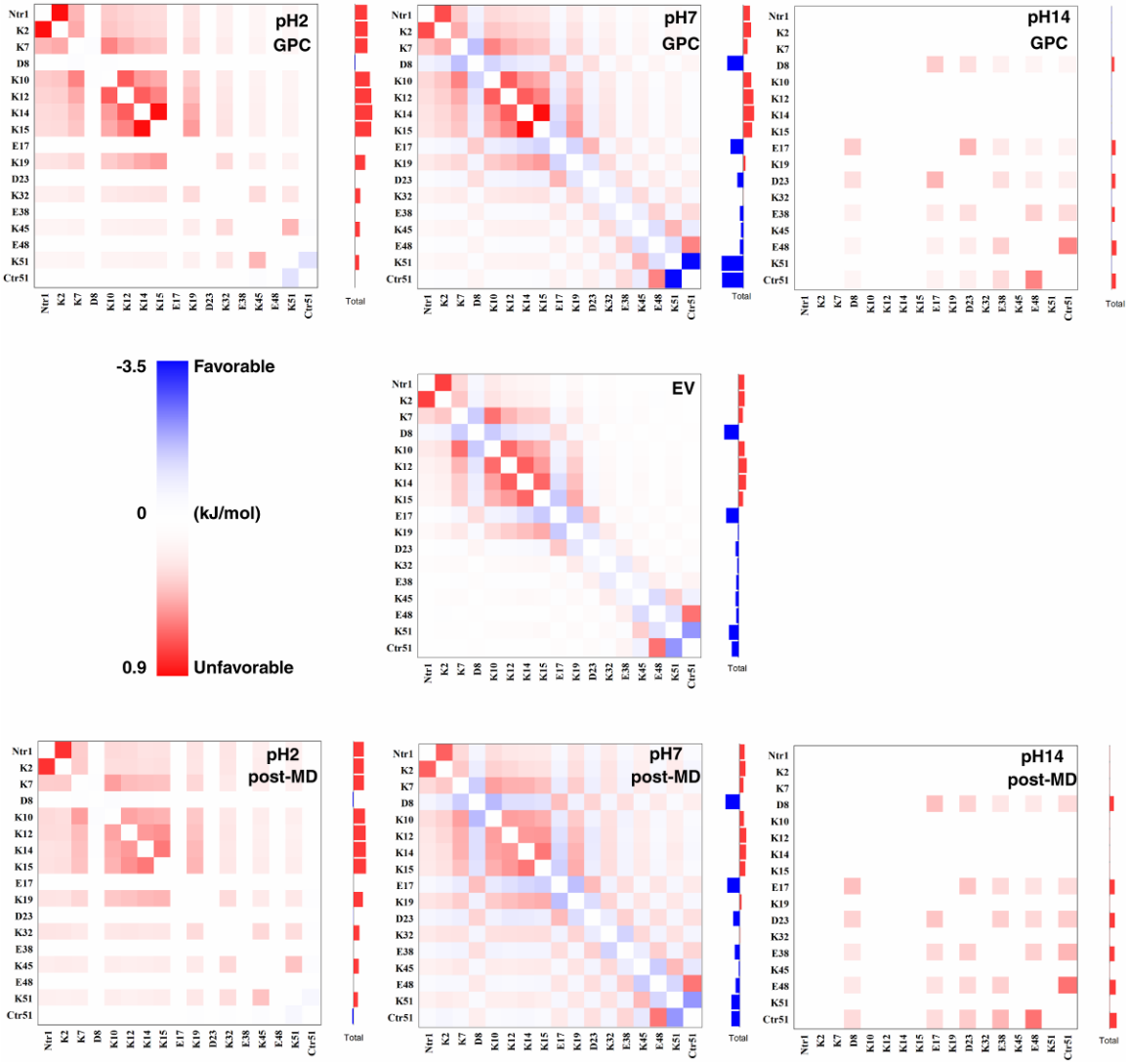

Figure 4.12 pH-dependent electrostatic "contact" maps for three unfolded state models of NTL9. The squares represent the pairwise charge-charge interactions in the unfolded state. Red squares indicate unfavorable interactions, while blue squares are favorable interactions. The bar graph to the side of each contact map represents the total interaction energy of a given residue with every other charged residue in the protein. There are no significant differences between interaction energies predicted by the Gaussian model (GPC), the EV limit, or the post-MD structures at any pH. There are also no indications of significant residual charge-charge interactions for any residue, except for the positively charged Lys51 and the negatively charged carboxyl group that exists on this C-terminal residue.

CHAPTER 5: EFFECTS OF PROTEIN STABILIZATION THROUGH THE RATIONAL DESIGN OF SURFACE CHARGE-CHARGE INTERACTIONS ON THE KINETICS OF PROTEIN FOLDING AND UNFOLDING REACTIONS

5.1 Introduction

In order to gain a more complete understanding of the mechanisms by which the rational design of surface charge-charge interactions stabilizes proteins, the effects of stabilization on the folding and unfolding kinetics of the protein need to be studied. If it is possible to predict the effect that optimizing surface charge-charge interactions has on the kinetics of folding and unfolding, then it might be possible to incorporate selection for kinetic stability into the TK-SA model. The term "kinetic stability" is typically used to describe proteins that have extremely slow unfolding rates (*212-215*). These proteins are not necessarily thermostable (high T_m) or thermodynamically stable (high ΔG), but they unfold slowly enough that this process is rarely observed. Understanding the forces that drive the kinetic stabilization is of great importance in the pharmaceutical industry, because kinetic stabilization of proteins will not only extend the shelf lives of protein-based therapeutics or vaccines, but could also increase the *in vivo* half-lives of such drugs, leading to more effective treatments.

By studying how the folding and unfolding kinetics are affected for proteins that have been stabilized by the optimization of surface charge-charge interactions, we can begin to understand the intramolecular interactions that are important for defining kinetic stability. Interactions in the protein core are known to be important for folding, and it has been demonstrated how dramatically substitutions in the core can affect the stability and structure of a protein (*3, 6-8, 17, 62, 147, 216*). It is also believed that the first step in protein folding is the hydrophobic collapse of the protein core, and that the interactions between surface residues can

begin to occur only after the core is formed. Conversely, for a protein to unfold, the surface interactions would need to be disrupted before the core can fully unfold. For a protein that undergoes two-state unfolding, the thermodynamic stability, defined by the Gibbs free energy of unfolding (ΔG), at a given temperature, T, can be represented by:

$$\Delta G(T) = -RT \ln(K_{eq}) \quad \textbf{(5.1)}$$

where R is the gas constant and K_{eq} is the equilibrium constant, which can be represented by the folding (k_f) and unfolding (k_u) rates, such that Eq. 5.1 becomes:

$$\Delta G(T) = -RT \ln\left(\frac{k_u}{k_f}\right) \quad \textbf{(5.2)}$$

Substitutions that increase the stability (ΔG) of a protein can either significantly decrease the unfolding rate, resulting in kinetically stable proteins, or increase the folding rate. It is also possible for a combination of these two mechanisms to occur.

One of the fundamental assumptions in the interpretation of folding and unfolding kinetics is that the substitutions do not significantly perturb the unfolded state structure of the protein. Therefore, all substitutions are assumed to exert their effects either in the native state, the transition state, or both. Figure 5.1A shows how a designed protein can be stabilized such that the unfolding rate is decreased ($k_{u,DES} < k_{u,WT}$), with no effect on the folding rate ($k_{f,DES} = k_{f,WT}$). This mechanism would occur if substitutions did not perturb the interactions in either the unfolded state or transition state ensembles ($\Delta G_{\ddagger,DES\rightarrow U} = \Delta G_{\ddagger,WT\rightarrow U}$). The stabilizing interactions are only present in the native state and provide a favorable contribution to the free energy of the system. The energy barrier between the native state and the transition state increases due to these favorable interactions ($\Delta G_{N,DES\rightarrow\ddagger,DES} > \Delta G_{N,WT\rightarrow\ddagger,WT}$), resulting in a much slower unfolding rate of the designed protein relative to the wild-type.

Alternatively, a protein can be stabilized in such a way that the folding rate is increased ($k_{f,DES} > k_{f,WT}$) with no effect on the unfolding rate of the protein ($k_{u,DES} = k_{u,WT}$) (Fig. 5.1B). In this case, the stabilizing interactions are present in both the transition state and the native state of the protein, while the unfolded state ensemble is not affected. By having the stabilizing interactions present in the transition state, the energy barrier between the unfolded state ensemble and the transition state ensemble is decreased ($\Delta G_{‡,DES \rightarrow U} < \Delta G_{‡,WT \rightarrow U}$), resulting in a faster folding rate. The free energy of the native state of the designed protein is decreased by the same amount ($\Delta G_{N,DES \rightarrow ‡,DES} = \Delta G_{N,WT \rightarrow ‡,WT}$), relative to the wild-type, so the unfolding rates remain unchanged.

The primary method for experimentally determining whether interactions are present in the transition state ensemble of a protein is Φ-value analysis. The Φ-value of a protein is usually determined by substituting the position of interest to alanine. The folding and unfolding kinetics of the Ala variant are measured and compared to that of the wild-type:

$$\Phi = \frac{\Delta G_{WT}^{‡\rightarrow \mathrm{U}} - \Delta G_{Ala}^{‡\rightarrow \mathrm{U}}}{\Delta G_{WT}^{N\rightarrow U} - \Delta G_{Ala}^{N\rightarrow U}} = \frac{-RT \ln\left(\frac{k_f^{Ala}}{k_f^{WT}}\right)}{-RT\left[\ln\left(\frac{k_u^{WT}}{k_u^{Ala}}\right) + \ln\left(\frac{k_f^{Ala}}{k_f^{WT}}\right)\right]} \qquad \textbf{(5.3)}$$

where $\Delta G_{WT}^{N\rightarrow U}$ and $\Delta G_{Ala}^{N\rightarrow U}$ are the energy differences between the native and unfolded states of the wild-type protein and the Ala variant, respectively; $\Delta G_{WT}^{‡\rightarrow \mathrm{U}}$ and $\Delta G_{Ala}^{‡\rightarrow \mathrm{U}}$ are the energy differences between the transition states (‡) and unfolded states of the wild-type and Ala variant, respectively; k_f^{WT} and k_f^{Ala} are the folding rates of the wild-type and alanine variant; and k_u^{WT} and k_u^{Ala} are the unfolding rates of the wild-type and alanine variant, respectively. Ideally, Φ-values should have a value of either 0 or 1 (*217*). A Φ-value of 0 indicates that the residue does not participate in native-like interactions in the transition state, which describes the kinetic mechanism shown in Fig. 5.1A. A Φ-value of 1 means that native-like interactions are

present in the transition state, and would indicate a kinetic mechanism like that shown in Fig. 5.1B. In practice, however, fractional Φ-values between 0 and 1 are often observed (*182, 218-225*). The standard interpretation of such Φ-values is: $0.7 < \Phi \leq 1$ indicates strong native-like interactions are present in the transition state ensemble; $0.2 < \Phi \leq 0.7$ indicates weak native-like interactions in the transition state; and $0 < \Phi \leq 0.2$ indicates that the residue does not participate in native-like interactions in the transition state. Φ-values less than zero and greater than one have also been shown to occur (*220*), and are typically taken to mean that strong non-native contacts are formed in the transition state ensemble.

If the first step in folding is driven by interactions between residues in the protein core, then one would expect that substitutions in the protein core would be more likely to affect folding rates than unfolding rates. Conversely, if the interactions between surface residues are not involved in the first steps of folding, and are only present after the protein core has formed, then substitutions on the surface should be more likely to affect the unfolding rates. In other words, a protein that is stabilized through the optimization surface charge-charge interactions should have a kinetic mechanism similar to that shown in Fig. 5.1A. This hypothesis was supported by several studies on the folding kinetics of CspB variants from mesophilic (CspB-Bs), thermophilic (CspB-Bc) and hyperthermophilic (CspB-Tm) organisms, which differ primarily in their surface charge distributions (*15*). CspB-Bs has the least favorable distribution of surface charges, while CspB-Tm has the most favorable distribution. It was found that the more stable CspB-Bc and CspB-Tm variants, have unfolding rates that are 20-fold and 220-fold slower than CspB-Bs, while the folding rates were both similar to that of CspB-Bs (*66*). This result suggests that not only do optimized surface charge-charge interactions define the unfolding rate of the CspB variants, but also that these interactions are not present in the CspB folding transition state.

To examine whether a CspB variant that was engineered to have optimized surface charge-charge interactions had similar behavior to the naturally occurring CspB variants, the

folding and unfolding kinetics of CspB-TB were measured. CspB-TB has the same core residues of CspB-Bs, but the surface charge distribution of CspB-Tm (*15*). It was found that CspB-TB did have the same kinetic mechanism of stabilization as the naturally occurring proteins – its unfolding rate was found to be 50-fold slower than the mesophilic CspB-Bs, while the folding rate was similar to that of CspB-Bs (*226*). This observation provides further support to the hypothesis that charge-charge interactions are not present in the folding transition state. These results also raised the question of whether slower unfolding rates is the general kinetic mechanism for proteins that are stabilized by the rational design of surface charge-charge interactions. To address this question the folding and unfolding kinetics were characterized for the wild-type and designed variants of the Fyn SH3 domain (Fyn) (Chapter 3, (*97*)), procarboxypeptidase (Pc) (*16*), and tenascin (Ten) (*16*). These three model systems provide an excellent starting point to address this question because all three wild-type proteins have been very well characterized kinetically (*177, 180, 182, 220, 222-224, 227-230*).

5.2 Results & Discussion

The effects of the substitutions made in each of the Fyn variants to their thermodynamic stabilities has been previously discussed (Chapter 3, (*97*)). Except for two of the single substitution variants, all designed proteins had significantly increased thermostabilities (T_m) and thermodynamic stabilities (ΔG(25 °C)) compared to Fyn-WT (Table 5.1). The designed variants of Pc and Ten also have significantly increased T_m and ΔG relative to their wild-type proteins (*16*). In order to directly compare the results of the kinetics experiments described here to previous equilibrium experiments, the thermal denaturation of Pc-WT, Pc-GA1, Pc-GA2, Ten-WT, and Ten-GA1 was measured using CD spectroscopy (Fig. 5.2). The changes in mean residue ellipticity as a function of temperature were monitored at 222 nm for the Pc variants and

230 nm for the Ten variants. The shifts in the midpoint of the transitions of the designed variants of Pc (Fig. 5.2A) and Ten (Fig. 5.2B) to higher temperatures, indicate that the T_m of the designed variants were higher than the respective wild-type proteins (Table 5.1), and the results are in good agreement to those obtained previously (*16*).

Equilibrium urea denaturation experiments were performed to corroborate the ΔG values obtained from extrapolation of the thermal denaturation data. We were unable to perform these experiments for the Fyn variants because none of them are unfolded at saturating urea concentrations. The urea denaturation of the Pc and Ten variants were measured using CD spectroscopy at 25 °C and 37 °C, respectively. It has previously been shown that tenascin unfolds extremely slowly at room temperature (*177*), so the higher temperature was used to ensure that the reaction reached equilibrium within the time constraints of the instrumentation. The results of the urea denaturation of all Pc variants are shown in Fig. 5.3A. Both Pc-GA1 and Pc-GA2 show increased stability relative to the wild-type, although the magnitude of this increase was slightly different between the urea denaturation and thermal denaturation experiments (Table 5.1). The small discrepancy between these experiments could be due to errors associated with defining the unfolded state baseline of the thermal denaturation data, which would cause large errors in the extrapolation of the data to obtain ΔG(25 °C). Importantly, the equilibrium denaturation experiments demonstrate that the designed Pc variants are more stable than their respective wild-type proteins at room temperature. The results of the urea denaturation of the Ten variants are shown in Fig. 5.3B. The fit of the data to a two-state model of unfolding show that Ten-GA1 is more stable than Ten-WT at 37 °C. Furthermore, the results of the urea denaturation of the Ten variants is in excellent agreement with the thermal denaturation results (Table 5.1).

The refolding kinetics of the Fyn, Pc-WT, Pc-GA1, and Ten variants were measured by fluorescence stopped-flow at 25 °C (Fyn & Pc) and 37 °C (Ten). The Fyn variants did not unfold completely in saturating concentrations of urea at 25 °C, so the unfolding rates of these proteins

were calculated by solving Eq. 5.2 for k_u, using the ΔG values obtained from DSC experiments. The Ten variants unfolded extremely slowly, so the kinetics of their unfolding reactions were measured by monitoring changes in the intrinsic tryptophan fluorescence of manually mixed solutions, as described in Chapter 2.9. The Pc-GA2 variant had a very small change in fluorescence upon unfolding, so the folding and unfolding kinetics of all Pc variants were also measured using CD stopped-flow.

The results of the kinetics experiments for the Fyn variants are given in Table 5.2. Interestingly, it appears the primary kinetic mechanism of stabilization for the designed variants Fyn2 and Fyn5 is an increase in the folding rates, while the Fyn3 variant appears to be stabilized primarily through a 7-fold slower unfolding rate than Fyn-WT. To explore the sources of the different kinetic mechanisms of stabilization for the designed Fyn variants, the folding kinetics of five single variants, each containing one of the substitutions in Fyn5, were characterized. Analysis of the single variants indicated that the two substitutions that did not significantly affect thermodynamic stability had very different effects on the folding and unfolding kinetics. The remaining three substitutions had significant effects on both thermodynamic stability and either k_f or k_u (Table 5.2).

The E11K substitution appears to have little effect on either the stability or the folding and unfolding rates of Fyn. The observation that the folding rates of Fyn-E11K and Fyn-WT are similar suggests that E11 does not participate in native-like charge-charge interactions in the transition state ensemble of Fyn. Since the folding rates and thermodynamic stabilities of Fyn-E11K and Fyn-WT are similar, the unfolding rates of these two proteins must also be similar (Eq. 5.2). The N30K substitution also has very little effect on the stability of Fyn ($\Delta\Delta G$ = -1 kJ/mol at 25 °C). The slight decrease in stability of Fyn-N30K relative to Fyn-WT seems to be primarily due to a 3-fold faster unfolding rate. The similar folding rate of Fyn-N30K relative to Fyn-WT suggests that this residue does not participate in native-like interactions in the transition state

(Fig. 5.1A). Therefore, the destabilizing interactions of N30K must only be present in the native state, which decreases $\Delta G_{N\rightarrow\ddagger}$, resulting in faster unfolding rates. The remaining single variants: Fyn-D16K, Fyn-H21K, and Fyn-E46K all have significant effects on the stability of Fyn. For D16K and H21K, the change in stability appears to be primarily due to 3.6-fold and 2.4-fold slower unfolding rates, respectively. The folding rates of both of these variants are similar to Fyn-WT (Table 5.2), suggesting that stabilizing interactions are not present in the transition state ensemble. Rather, it appears that the decreased unfolding rates of Fyn-D16K and Fyn-H21K are most likely due to the alleviation of the electrostatic repulsion caused by the E15-D16-D17 sequence present in Fyn-WT. The E46K substitution provides an example where stabilization can occur through a combination of changes in k_f and k_u. The folding rate of Fyn-E46K is 3-fold faster than Fyn-WT, and the unfolding rate is 2-fold slower. The significant increase in the folding rate of Fyn-E46K could be due to favorable interactions with E24 upon substitution, and suggests that the effects of these favorable interactions are also present in the transition state ensemble.

The thermal denaturation of the Fyn2 variant (E11K/E46K) was discussed in Chapter 3 and showed that the effects of these substitutions are additive ($\Delta\Delta G_{Fyn2\text{-}WT} = \Delta\Delta G_{E11K\text{-}WT} + \Delta\Delta G_{E46K\text{-}WT}$). The additivity of thermodynamic stabilities can mean that the folding and unfolding rates are also additive, but it is also possible for the folding and unfolding rates change in opposite directions. Interestingly, Fyn2 has a folding rate similar to Fyn-E46K and an unfolding rate similar to Fyn-E11K and Fyn-WT, suggesting that the effects of these substitutions on the folding kinetics are additive. Therefore, the increased folding rate of Fyn2 relative to Fyn-WT must be due to the presence of native-like interactions at position E46 in the transition state. The stabilization of the Fyn3 variant (E11K/D16K/E46K) has also been shown to be additive in terms of ΔG (see Chapter 3). If this additivity is also present in the folding kinetics of Fyn3, then we would expect $k_{f,Fyn3}$ to be 3-fold faster than $k_{f,WT}$ because the E46K substitution is the only one that

affects k_f. In addition, $k_{u,Fyn3}$ should be approximately 7-fold slower than $k_{u,WT}$ due to the combination of the E46K (2-fold decrease) and D16K substitutions (3.6-fold decrease). The results of the kinetics experiments shown in Table 5.2 show that the effects of these three substitutions are indeed additive. The increased thermodynamic stability of Fyn3 appears to be due to an unfolding rate that is 7-fold slower than Fyn-WT and a folding rate that is 2-fold faster. The five substitutions in the Fyn5 variant (E11K/D16K/H21K/N30K/E46K) have also been shown to be additive in terms of ΔG (see Chapter 3). In contrast with the Fyn2 and Fyn3 variants, however, the Fyn5 variant provides an example of how substitutions can be additive in terms of ΔG, yet seem to have a synergistic effect in terms of the folding and unfolding kinetics. Based on the principles of additivity, one would expect the folding rate of Fyn5 to be approximately 3-fold faster than Fyn-WT because the E46K substitution is the only one that has any effect on the folding rate of Fyn. Moreover, the unfolding rate of Fyn5 should be approximately 15-fold slower than Fyn-WT due to the D16K (3.6-fold decrease in k_u), H21K (2-fold decrease in k_u), and E46K (2-fold decrease in k_u) substitutions. Instead, we observe that the folding rate is 8.5-fold faster and the unfolding rate is 2-fold slower than Fyn-WT, suggesting that these substitutions have synergistic effects in terms of the folding and unfolding kinetics. Importantly, the observation that Fyn2, Fyn3, and Fyn5 all had increased folding rates relative to Fyn-WT argues that it is possible for long-range native-like charge-charge interactions to be present in the transition state ensemble of proteins. The effects of rationally designed surface charge-charge interactions on the folding/unfolding kinetics of Fyn do not support the hypothesis, based on the CspB data, that optimization of surface charge-charge interactions will result in slower rates of unfolding. In order to determine the source of the different kinetic behavior between the Fyn and CspB variants, the folding and unfolding kinetics of designed variants of Pc and Ten were also characterized.

Figure 5.4A compares the results of the kinetics experiments performed on the Pc variants using both CD and fluorescence spectroscopy. The natural log of the observed folding/unfolding rate (ln k_{obs}) is plotted as a function of urea concentration, which results in a U-shaped plot known as a chevron plot. The CD and fluorescence data were fit to a two-state model:

$$\ln k_{obs} = \ln\left(k_{f,H_2O}\exp\left(-m_f\left[Urea\right]\right)+k_{u,H_2O}\exp\left(-m_u\left[Urea\right]\right)\right) \qquad \textbf{(5.4)}$$

where $k_{f,H2O}$ and $k_{u,H2O}$ are the folding and unfolding rates at 0 M urea, respectively; and m_f and m_u are the slopes of the folding and unfolding arms of the chevron, respectively. From Fig. 5.4A, it can be seen that the data obtained for Pc-WT and Pc-GA1 using fluorescence (open symbols) are in good agreement with the CD data (filled symbols). Indeed, the fits of the data to a two-state model of unfolding give results for k_f and k_u that are in good agreement between the fluorescence and CD kinetics experiments (Table 5.2). Furthermore, the thermodynamic stabilities of the Pc variants based on the folding and unfolding rates (ΔG_{kin}) are within experimental error of the stabilities measured by equilibrium urea denaturation (ΔG_{eq}) (Table 5.2).

Although Pc-GA1 is more thermostable than Pc-WT (Table 5.1), the equilibrium urea denaturation experiments suggested that these two variants would have similar stabilities. This is precisely what was observed with the kinetics experiments, where the folding and unfolding rates of Pc-GA1 were both increased 4-fold relative to Pc-WT. Since the ratio of k_u and k_f are the same for Pc-WT and Pc-GA1, then according to Eq. 5.2, the thermodynamic stabilities will also be similar. The equilibrium unfolding of Pc-GA2 showed that this protein is both more thermostable and more thermodynamically stable than Pc-WT. The kinetic mechanism of this stabilization appears to be more similar to that of Fyn5 than CspB-TB; i.e. optimization of surface charge-charge interactions results in changes in the folding rate rather than the unfolding rate. The folding rate of Pc-GA2 was increased 11-fold, relative to Pc-WT (Table 5.2), providing another

example of the presence of native-like long-range charge-charge interactions in the transition state ensemble of proteins.

In order to determine if any of the positions subjected to substitution in the Pc-GA1 and Pc-GA2 variants were structured in the transition state, we examined Φ-values measured by alanine-scanning mutagenesis that have been reported in the literature (*220*). Typically, high Φ-values (0.71 – 1) indicate native-like structure in the transition state, medium values (0.21 – 0.7) indicate weak native-like interactions, and low values (0 – 0.2) indicate no structure in the transition state ensemble (*220, 227*). Only four of the eleven positions selected for substitution in Pc-GA1 and Pc-GA2 had been characterized, but Φ-values for the neighboring residues were available for three other positions. From the results presented in table 5.3, it can be seen that one position in Pc-GA1 (S65) and three positions in Pc-GA2 (Q23, Q60, and S65) are structured in the transition state ensemble. The remaining positions that were measured appear to be unstructured in the transition state ensemble. However, it is not always appropriate to infer the behavior of a given residue based solely on the Φ-value of one neighbor. For example, position 51 in Pc-WT is not structured in the transition state ensemble, while position 52 has some native-like structure (*220*). Although, the Φ-value analysis for the substitutions selected by TK-SA is incomplete, the results presented here suggest that the dramatic increase in the folding rate of Pc-GA2 could be explained by some of the substitutions being made at positions known to participate in native-like interactions in the transition state ensemble.

The folding and unfolding kinetics were also measured for the wild-type and one designed variant of tenascin. Based on the equilibrium denaturation experiments, Ten-GA1 is expected to be both more thermostable and more thermodynamically stable than Ten-WT. Figure 5.4B shows the results of the fluorescence kinetics experiments for the Ten variants. The symbols represent the experimental data and the solid lines are the fits of the data to a two-state model of unfolding. The fitted kinetic parameters are given in Table 5.2, where it can be seen

that both the folding and unfolding rates of the Ten variants are orders of magnitude slower than the Fyn and Pc variants. Based on the values for k_f and k_u, Ten-GA1 is more stable than Ten-WT by 7.3 kJ/mol, which is in good agreement with the $\Delta\Delta G$ value obtained from the equilibrium unfolding experiments. Furthermore, the stabilization of Ten-GA1 has the same kinetic mechanism as the designed Fyn and Pc variants: Ten-GA1 has a folding rate that is 50-fold faster than that of Ten-WT, while their unfolding rates are very similar.

To determine whether this dramatic increase in k_f is due to selecting positions for substitution known to be structured in the transition state ensemble, the previously reported Φ-values measured by alanine scanning-mutagenesis (*223*) were examined (Table 5.4). Although, only one of the exact positions (L29) was characterized in this study, both neighboring residues of the remaining three positions (Q7, D49, and T89) were also characterized. If both neighbors of a given residue participate in native-like interactions in the transition state, then it is assumed that the residue of interest must also participate in these interactions, and vice versa. From the results shown in Table 5.4, it appears that only one residue, D49, has native-like structure in the transition state. However, when the four substitutions are made simultaneously, it appears that the net effect is the presence of long-range native-like charge-charge interactions in the transition state ensemble of Ten.

The results of the experiments on Fyn, Pc, and Ten raise the question: why does the rational design of surface charge-charge interactions result in increased folding rates, when the CspB studies suggested that optimization of these interactions should result in slower unfolding rates? One potential answer to this question is that the substitutions affect the transition state ensembles of these four proteins differently. If Fyn, Pc, and Ten all had substitutions at positions known to be structured in the transition state, then perhaps the CspB substitutions occur at positions that do not participate in native-like interactions in the transition state. In order to test this hypothesis, we examined the Φ-value analyses of the CspB-Bs and CspB-Bc variants

measured by alanine-scanning mutagenesis (*225, 231*). Out of the 15 substitutions that were made to create CspB-TB, 11 had been characterized in these studies. Table 5.5 shows that six of these residues (E3, N10, E12, E19, E42, and S48) have medium to high Φ-values, indicating that they are structured in the transition state, whereas only one or two of the positions in Fyn, Pc, or Ten had native-like structure. This result does not support the hypothesis that the substitutions in CspB should occur at positions that are unstructured in the transition state.

Furthermore, the CspB Φ-value analysis conflicts with the mechanisms proposed by Figure 5.1, where substitutions at positions that have native-like structure in the transition state should only affect the folding rates. However, the conflicting results could be due to the nature of extrapolating results from alanine-scanning Φ-value analysis to the effects of charge-charge substitutions. Alanine substitutions perturb a variety of intramolecular interactions such as hydrogen bonding, secondary structure propensity, hydrophobic interactions and packing interactions. It is possible that perturbation of all of these forces simultaneously could make it appear that a particular residue is not structured in the transition state. However, if only one type of interaction was perturbed (i.e. charge-charge), it might be more evident that a given residue participates in native-like charge-charge interactions, even though the residue is not structured. For example, the Φ-value analysis based on alanine-scanning mutations in Fyn suggested that N30 is structured (*227*), while E46 is not (*227, 229*). Yet when the substitutions are to a lysine rather than alanine, the E46 position does have native-like interactions and N30 does not. It might be necessary, then, to determine the presence of long-range charge-charge interactions in the transition state ensembles by performing an analysis on the CspB, Pc, and Ten variants using charged residues, as was done here with Fyn.

Another potential explanation for the different kinetic mechanisms of CspB and the other proteins is that the CspB-TB variant was designed based on the sequences of naturally occurring stable proteins, whereas the Fyn, Pc, and Ten designs were not. It is possible that the

evolutionary determinants of protein stability are different than what can currently be modeled and selected for computationally. One can imagine that if life originated in a hot environment and then organisms had to adapt to a gradually cooling environment, the folding rate might be less important to adaptation than the unfolding rate. Such a mechanism would be especially important for regulatory proteins, which must be degraded when they are no longer needed. For a protein like CspB, which is activated in response to cell stress caused by cold temperatures, adaptation to a cooling environment would make it necessary to degrade this protein quickly, so evolution would favor faster unfolding rates. The result is that the ancestral thermostable protein has a much slower unfolding rate than its more modern mesophilic counterpart. Since it is not currently clear how one could computationally model evolutionary determinants to stability, it is possible that this is the source of the different kinetic mechanisms. To test this hypothesis, one would need to characterize a sequence of CspB that was selected for by the TK-SA model and compare its folding/unfolding kinetics to that of the CspB-TB variant, which was designed based on naturally occurring sequences. If the folding mechanism of the TK-SA designed CspB variant was the same as the CspB-TB variant, then it is possible that evolution favors different kinetic mechanisms to stabilize regulatory proteins versus non-regulatory proteins, such as Fyn, Pc, and Ten. However, if the folding mechanisms between the TK-SA designed CspB and CspB-TB were different, then it would suggest that the computational model is not capturing evolutionary pressures to stability.

It is also possible that the fundamental assumption in Φ-value analysis – namely, that the substitutions do not affect the unfolded state ensemble – is flawed (*189*). Although the data presented in Chapter 4 supports the conclusion that it is not necessary to consider the presence of residual charge-charge interactions in the unfolded state to predict changes in thermodynamic stability ($\Delta\Delta G$), it is possible that the unfolded state is extremely important for predicting the effects of substitutions on the level of kinetic stability. For example, the D16K substitution in

Fyn seems to stabilize the protein primarily by introducing favorable charge-charge interactions into the -E15-D16-D17- sequence. Since this residue engages in favorable interactions with neighboring residues, it would be incorrect to assume that these effects are not also present in the unfolded state ensemble. In order to see whether incorporating the unfolded state effects into the model might be important for determining kinetically stabilizing substitutions, we examined the total energy of unfolded state charge-charge interactions using the Gaussian chain model ($E_{unf,Gauss}$) for the wild-type and optimized variants of CspB, Fyn, Pc, and Ten. Table 5.6 shows that the CspB variants with optimized native state charge-charge interactions result in *unfavorable* unfolded state charge-charge interactions. On the other hand, the designed Fyn, Pc, and Ten variants all have more *favorable* unfolded state charge-charge interactions.

Figure 5.5 demonstrates how incorporating the unfolded state effects into the unfolding reaction scheme can affect the interpretation of kinetic experiments. Only two extreme points are modeled for clarity, although in practice, a combination of these mechanisms is also likely to occur. The first mechanism represents what could happen with the CspB variants (red lines), the energy of the unfolded state ensemble of the designed protein ($U_{DES} > U_{WT}$) is increased relative to wild-type, while the energy of the native state is decreased ($N_{DES} < N_{WT}$). This results in a larger thermodynamic stability for the designed variant relative to the wild-type. In order to maintain similar folding rates ($k_{f,DES} = k_{f,WT}$), the transition state must also be destabilized relative to the wild-type protein ($TS_{DES} > TS_{WT}$). The destabilization of the transition state would also increase the energy barrier between the native state and the transition state, resulting in slower rates of unfolding for the designed variant ($k_{u,DES} < k_{u,WT}$).

The second mechanism represents a possible explanation for the data from the Fyn, Pc, and Ten variants (green lines). In this case, the substitutions decrease the free energy of both the native ($N_{DES} < N_{WT}$) and unfolded states ($U_{DES} < U_{WT}$). However, the magnitude of the decrease in the free energy of the unfolded state must be smaller than that of the native state, since the

difference between N and U must be larger for the more stable designed variant than the wild-type protein. In order to maintain similar unfolding rates between the wild-type and designed variant ($k_{u,DES} = k_{u,WT}$), the free energy of the transition state must also be decreased by the same amount as the native state. This will ultimately result in a smaller energy barrier between the unfolded state and the transition state, resulting in faster rates of folding for the designed variant ($k_{f,DES} > k_{f,WT}$). In terms of k_f and k_u, both of these mechanisms have very similar results to those described in Fig. 5.1, but the underlying cause of the changes is vastly different.

The qualitative correlation between the changes unfolded state charge-charge interaction energies and the changes in k_f and k_u suggests that these energies could be important in our design approach, not for selecting thermodynamically stable protein sequences, but for selecting kinetically stable proteins. This hypothesis can be tested in two ways. First, we can characterize the folding and unfolding kinetics of the existing designed variants of ubiquitin, acylphosphatase, and U1A. Based on the unfolded state charge-charge interaction energies ($E_{unf,Gauss,WT}$ = -3.37 kJ/mol; $E_{unf,Gauss,GA1}$ = -3.48 kJ/mol; $E_{unf,Gauss,GA2}$ = -1.47 kJ/mol; $E_{unf,Gauss,GA3}$ = -1.52 kJ/mol), we would hypothesize that Ubq-GA1 should fold faster than Ubq-WT because the unfolded state energy becomes more favorable. On the other hand, Ubq-GA2 and Ubq-GA3 should unfold more slowly because their unfolded state energies are less favorable than Ubq-WT. The designed variants of both acylphosphatase (Acp-GA1) and U1A (U1A-GA1) would also be expected to fold faster than their respective wild-type proteins because the unfolded state charge-charge interaction energies of Acp-GA1 ($E_{unf,Gauss,GA1}$ = -7.41 kJ/mol) and U1A-GA1 ($E_{unf,Gauss,GA1}$ = -3.60 kJ/mol) are more favorable than that of Acp-WT ($E_{unf,Gauss,WT}$ = -6.43 kJ/mol) or U1A-WT ($E_{unf,Gauss,WT}$ = -2.75 kJ/mol), respectively. The second way to test the hypothesis is to perform negative design of the unfolded state by using the TK-SA model to select sequences of Fyn, Pc, Ten, Acp, and U1A that are identified by the genetic algorithm to be more stable than the respective wild-type proteins, but that have unfavorable unfolded state charge-charge interaction

energies. If the results of these experiments support the hypothesis that the unfolded state energies are important for determining the direction in which k_f and k_u will change, then we will be able to add another layer of selection – for kinetic stability – into our design approach.

5.3 Concluding Remarks

The results presented in this chapter challenge a long-standing notion that long-range charge-charge interactions are not present in the transition state of proteins. We have shown three examples of model systems with different sizes and secondary structural content that all have long-range native-like charge-charge interactions present in their transition state ensembles. However, several issues need to be addressed before we will be able to incorporate selection for kinetic stability into our design algorithm. The ability to rationally design kinetically stable proteins will be of particular importance to the pharmaceutical industry, where the extended shelf-life or *in vivo* half-life that should accompany kinetically stable proteins can overcome many of the current limitations to protein-based therapeutics.

As mentioned in the previous section, it is possible that the lack of correlation between changes in k_f and k_u and the Φ-values of the positions subjected to substitution could be due to probing the presence of native-like interactions in the transition state ensemble with alanine. The alanine substitutions disrupt not only electrostatic interactions, but can also perturb hydrophobic interactions, secondary structure propensity, and packing interactions. By making the charge-charge substitutions, we can probe only the effects due to altering the charge at a given position. In order to get a clearer picture of the role of charges in the transition state ensemble of proteins, it will be necessary to characterize the Φ-values of single variants of all substitutions made in Pc-GA2, Ten-GA1, and CspB-TB.

To address the question of whether or not our computational model is implicitly accounting for evolutionary determinants of stability, it would be very interesting to redesign Fyn, Pc, and Ten based on naturally occurring sequences of these proteins, and see if the results are similar to those obtained with CspB. Conversely, it would also be important to computationally design CspB using only the TK-SA model, and see if the kinetic mechanism of stabilization is the same as for Fyn, Pc, and Ten. The results of such experiments might begin to give us some insight into the differences between proteins that are stabilized evolutionarily versus those stabilized computationally.

The experiments mentioned above might also help us test the hypothesis about the importance of the unfolded state in kinetic stability. It is possible that stable proteins engineered from naturally occurring stable sequences could bias the selection toward sequences that simultaneously stabilize the native state and destabilize the unfolded state, resulting in proteins that are both thermodynamically and kinetically stable. Conversely, our design approach considers only the native state, which is why none of the three proteins tested as part of this chapter were found to be kinetically stable. By measuring the folding and unfolding kinetics of the other proteins that were rationally designed using the TK-SA model, we can gain a better understanding of the roles of the transition and unfolded states in protein folding.

Table 5.1 Thermostabilities of proteins redesigned by TK-SA approach

Protein	Substitutions	ΔT_m (°C) [a]	$\Delta\Delta G_{Thermal}$ (kJ/mol)[b]	$\Delta\Delta G_{Urea}$ (kJ/mol)[c]	Ref.
Fyn	WT	-	-	-	(*97*)
	E11K	-1.0	0.1	n.d.[d]	
	D16K	5.5	3.9	n.d.[d]	
	H21K	5.0	2.8	n.d.[d]	
	N30K	-0.4	-1.1	n.d.[d]	
	E46K	6.1	4.7	n.d.[d]	(*97*)
Fyn2	E11K/E46K	4.6	2.3	n.d.[d]	(*97*)
Fyn3	E11K/D16K/E46K	10.3	6.7	n.d.[d]	(*97*)
Fyn5	E11K/D16K/H21K/N30K/E46K	11.7	7.1	n.d.[d]	(*97*)
Pc	WT	-	-	-	(*16*)
Pc-GA1	Q2E/H42E/S65K/M67K/D70K	3.9	4.1	0.9 (FL) 1.0 (CD)	(*16*)
Pc-GA2	Q19E/Q23K/K32E/E39K/Q60K/ S65K/E69K	9.8	10.7	7.3 (CD)	(*16*)
Ten	WT	-	-	-	(*16*)
Ten-GA1	Q7K/L29K/D49K/T89K	10.0	8.8	8.5 (FL) 6.5 (CD)	(*16*)

a $\Delta T_m = T_{m,\ mut} - T_{m,\ WT}$, measured by DSC for all Fyn and Ten variants and by CD for all PC variants.

b $\Delta\Delta G_{Thermal} = \Delta G_{mut} - \Delta G_{WT}$, calculated by extrapolating data from thermal denaturation experiments to 25 °C for Fyn and PC and to 37 °C for Ten.

c $\Delta\Delta G_{Urea} = \Delta G_{mut} - \Delta G_{WT}$, measured by equilibrium urea denaturation experiments at 25 °C for PC and at 37 °C for Ten.

d Urea denaturation experiments were not performed for the Fyn variants because they do not unfold at saturating urea concentrations at 25 °C.

(FL) urea denaturation followed by fluorescence spectroscopy

(CD) – urea denaturation measured by CD spectroscopy.

Table 5.2 Kinetic parameters of Fyn, Pc, and Ten variants

Protein	k_f (s^{-1})	m_f (kJ/mol M)	k_u (s^{-1})[a]	m_u (kJ/mol M)	ΔG_{kin} (kJ/mol)	ΔG_{eq} (kJ/mol)
Fyn-WT	76 [b]	3.22 [b]	0.029	-	-	19.5 [c]
Fyn-E11K	84 [b]	3.39 [b]	0.031	-	-	19.6 [c]
Fyn-D16K	102 [b]	3.13 [b]	0.008	-	-	23.4 [c]
Fyn-H21K	97 [b]	3.43 [b]	0.012	-	-	22.3 [c]
Fyn-N30K	105 [b]	3.22 [b]	0.062	-	-	18.4 [c]
Fyn-E46K	232 [b]	3.43 [b]	0.013	-	-	24.2 [c]
Fyn2	218 [b]	3.64 [b]	0.033	-	-	21.8 [c]
Fyn3	155 [b]	3.01 [b]	0.004	-	-	26.2 [c]
Fyn5	648 [b]	3.72 [b]	0.014	-	-	26.6 [c]
Pc-WT	668 [b] 1056 [d]	2.79 [b] 3.06 [d]	0.198 [b] 0.396 [d]	1.76 [b] 1.06 [d]	20.2 [b] 19.5 [d]	16.0 [b] 16.0 [d]
Pc-GA1	2246 [b] 2295 [d]	3.03 [e]	0.817 [b] 1.036 [d]	1.35 [e]	19.6 [b] 19.1 [d]	16.9 [b] 17.0 [d]
Pc-GA2	7581 [d]	3.07 [d]	0.643 [d]	0.87 [d]	23.2 [d]	23.3 [d]
Ten-WT	0.401 [b]	3.84 [b]	0.001 [b]	0.85 [b]	15.4 [b]	16.3 [b] 12.7 [d]
Ten-GA1	5.29 [b]	4.25 [b]	0.00078 [b]	1.04 [b]	22.7 [b]	24.9 [b] 19.2 [d]

a Unfolding of Fyn variants was calculated from the folding rates and the ΔG values measured by DSC.

b Parameters measured by fluorescence spectroscopy.

c Parameters measured by DSC.

d Parameters measured by CD spectroscopy.

e Pc-GA1 CD and fluorescence stopped-flow data were fit globally to minimize errors from the noise in the CD experiment.

Table 5.3 Φ-value analysis of selected Pc positions based on alanine-scanning data*

Position	k_f (s^{-1})	k_u (s^{-1})	ΔG_{kin} (kJ/mol)	Φ [a]	Pc variant
WT [#]	757	0.48	18.4	-	-
Q2	-	-	-	n.d. [b]	GA1
Q19 (G)	488	1.43	14.2	0.22 ± 0.01 (pos. 20) [c]	GA2
Q23 (V)	403	0.27	18.0	9.3 ± 31.0	GA2
K32 (G)	963	1.23	16.3	0.00 ± 0.10 (pos. 31) [c]	GA2
E39 (L)	518	11.0	9.6	0.11 ± 0.05	GA2
H42	330 - 403	n.d. [d]	8.7	~ 0 (pos. 41) [c]	GA1
Q60 (G)	757	0.84	4.3	0.48 ± 0.07	GA2
S65	192	1.63	11.7	0.43 ± 0.02	GA1 GA2
M67	-	-	-	n.d. [b]	GA1
E69	-	-	-	n.d. [b]	GA2
D70 (V)	446	3.6	12.1	0.21 ± 0.03 (pos. 71) [c]	GA1

* unless otherwise noted.

The wild-type Pc variant studied by (*220*) was not His-tagged, so the previously measured folding and unfolding rates differ slightly from our measurements. The data are included here as the reference state for these mutations.

a Data from Ref. (*220*), measured at 25 °C.

b Neither the exact position, nor a neighboring position was available for analysis.

c Although the status of a neighboring position does not always indicate the behavior of a given residue in the transition state ensemble (see (*220*)), when it was available, this information was included as a reference.

d Unfolding rate was too fast to be accurately determined.

Table 5.4 Φ-value analysis of selected Ten positions based on alanine-scanning data*

Position	k_f (s^{-1}) [a]	k_u (s^{-1}) [b]	$\Delta\Delta G_{kin}$ (kJ/mol) [a]	Φ [a]
WT [#]	5.4	1.7×10^{-5}	-	-
Q7	4.6 4.3	1.9×10^{-5} 5.5×10^{-5}	-11.4 -0.8	0.04 ± 0.01 (pos. 5) 0.10 ± 0.01 (pos. 8)
L29	4	1.0×10^{-4}	-5.8	0.13 ± 0.03
D49	0.47 0.65	1.9×10^{-4} 1.7×10^{-4}	-9.2 -12.5	0.67 ± 0.09 (pos. 48) 0.42 ± 0.02 (pos. 50)
T89	2.02 2.96	1.9×10^{-5} 4.8×10^{-5}	-22.7 -14.4	0.11 ± 0.01 (pos. 88) 0.11 ± 0.01 (pos. 90)

* unless otherwise noted.

a Φ-values, k_f and $\Delta\Delta G$ values from refolding in urea at 25 °C (*223*). The $\Delta\Delta G$ values ($\Delta G_{VAR} - \Delta G_{WT}$) are based on equilibrium denaturation experiments. When the exact positions were not available, neighboring positions are listed as a reference. The analysis given in the text is based on the assumption that when both neighboring residues are (un)structured in the transition state, then the residue in question is also (un)structured.

b k_u obtained from GuSCN denaturation at 25 °C (*224*).

These experiments were performed at a different temperature than ours. As a result, the folding and unfolding rates are slightly different from what we measured. The data are given here as a reference state for the Φ-values.

Table 5.5 Φ-value analysis of selected CspB positions from alanine-scanning data*

Position	k_f (s^{-1}) **	k_u (s^{-1}) **	$\Delta\Delta G_{kin}$ (kJ/mol) [a]	Φ [b]	Ref.
Bs#	1090	1.60			(*225*)
Bs-WT	253	9938	-	-	(*231*)
Bc	2253	462			(*231*)
L3# (E)	477	1.77	-4.5	0.44	(*225*)
E3 (R)	1434	1538	10.3	0.48	(*231*)
N10#	259	2.28	-4.3	0.79	(*225*)
E12#	1230	1.71	0.2	1.50	(*225*)
E19#	1190	1.50	0.4	0.50	(*225*)
V20	-	-	-	n.d.	
G35	-	-	-	n.d.	
G36	-	-	-	n.d.	
E42#	257 862	22.4 3.22	-9.7 -2.2	0.31 (pos. 41) 0.27 (pos. 45)	(*225*)
S48	-	-	-	0.21-0.7 [c]	(*231*)
N55	-	-	-	n.d.	
N62# (G)	636 519	49.9 157	-9.5 -12.1	0.14 (pos. 60) 0.09 (pos. 63)	(*225*)
V64 (T) [d]	2041 [d]	642 [d]	0.3 [d]	0.23 [d]	(*231*)
K65	-	-	-	n.d.	(*231*)
E66 (L)	488	1050	8.3	0.23	(*231*)
L66 (E)	1575 [d]	1997 [d]	0.7 [d]	0.23 [d]	
67A [d]	2520 [d]	512 [d]	0 [d]	0 [d]	(*231*)

* unless otherwise noted.

** The folding and unfolding rates were measured at 70°C in Ref. (*231*) and 15°C in Ref. (*225*).

a $\Delta\Delta G = \Delta G_{VAR} - \Delta G_{WT}$

b Φ-value analysis of positions in CspB-Bs that are substituted in CspB-TB. When the exact residue had not been characterized, the neighboring residues are shown for reference. As in Table 5.3, if data for both neighboring residues are available, and both are (un)structured in the transition state, then it is assumed that the intervening residue is also (un)structured in the transition state.

This CspB-Bs variant is pseudo-wild-type; it has a leucine at position 3

c Graphical data was published for this residue in Ref. (*231*), where it was only indicated that the Φ-value fell in the medium range, defined as $0.21 \leq \Phi \leq 0.70$. A Φ-value in this range generally indicates that the residue is weakly structured in the transition state.

d CspB-Bc is the reference state for this position.

Table 5.6 Correlation of folding/unfolding rates with unfolded state charge-charge interaction energies

Protein	$E_{unf,Gauss}$ **(kJ/mol)**	k_f **(s^{-1})**	k_u **(s^{-1})**
Fyn-WT	0.64	76	0.029
Fyn5	-5.49	648	0.014
Pc-WT	-1.27	668	0.198
Pc-GA1	-4.01	2246	0.817
Pc-GA2	-4.16	7581	0.643
Ten-WT	3.30	0.401	0.001
Ten-GA1	-1.83	5.29	0.0008
CspB-Bs	1.11	689	9.93
CspB-Bc	1.65	1370	0.64
CspB-Tm	3.69	565	0.018
CspB-TB	3.39	154	0.2

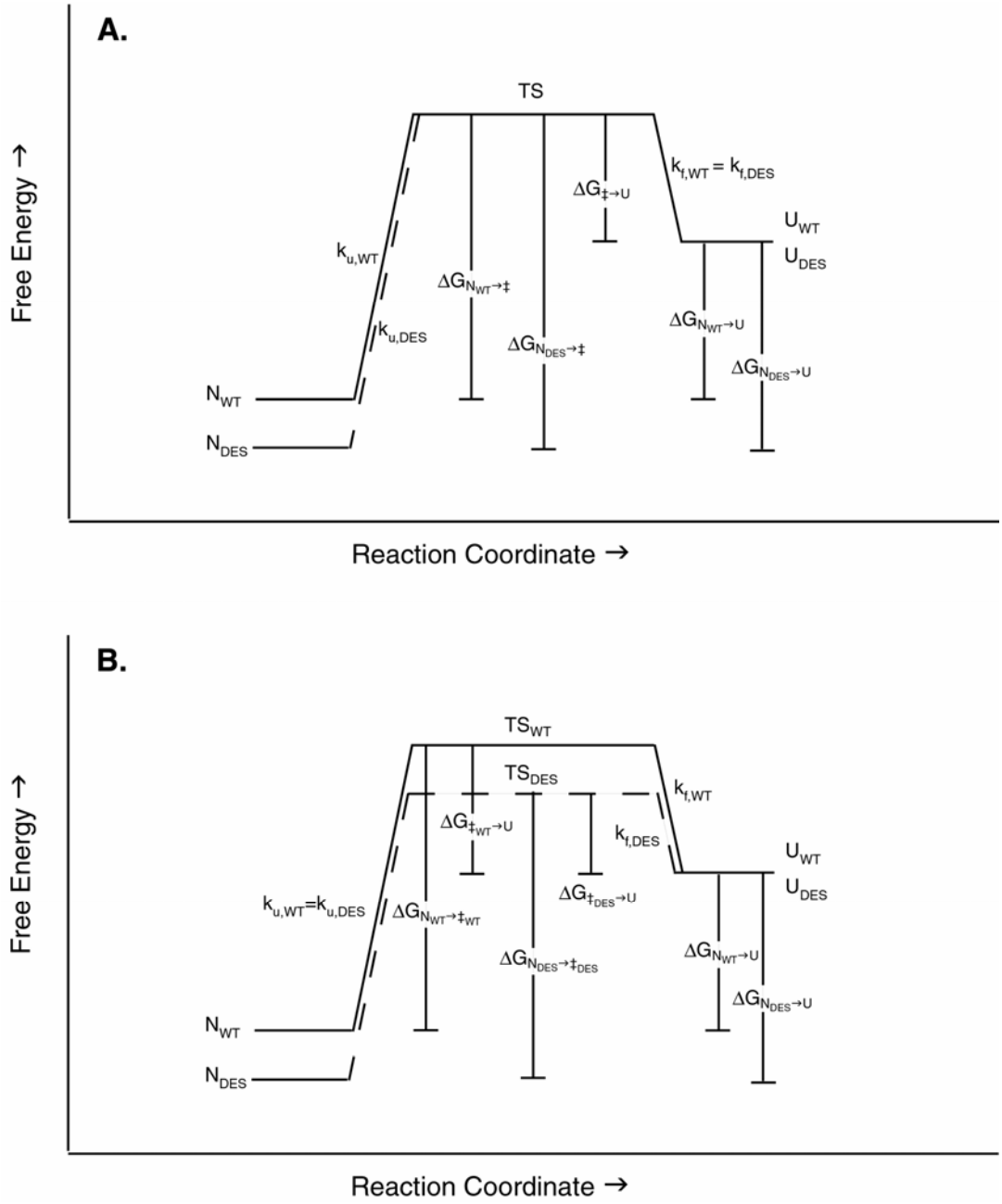

Figure 5.1 Two possible kinetic mechanisms of stabilization. **A.** The substitutions in the designed protein (DES) stabilize the native state, relative to the wild-type (WT), resulting in a slower unfolding rate($\Delta G_{N \to \ddagger, DES} > \Delta G_{N \to \ddagger, WT}$). **B.** The folding transition state of the designed protein (TS_{DES}) is stabilized by the same amount as the native state, relative to WT ($\Delta G_{\ddagger, DES \to U, DES} < \Delta G_{\ddagger, WT \to U, WT}$), resulting in a faster folding rate. For the purposes of this illustration, the designed proteins depicted in both **A** and **B** are stabilized by the same amount.

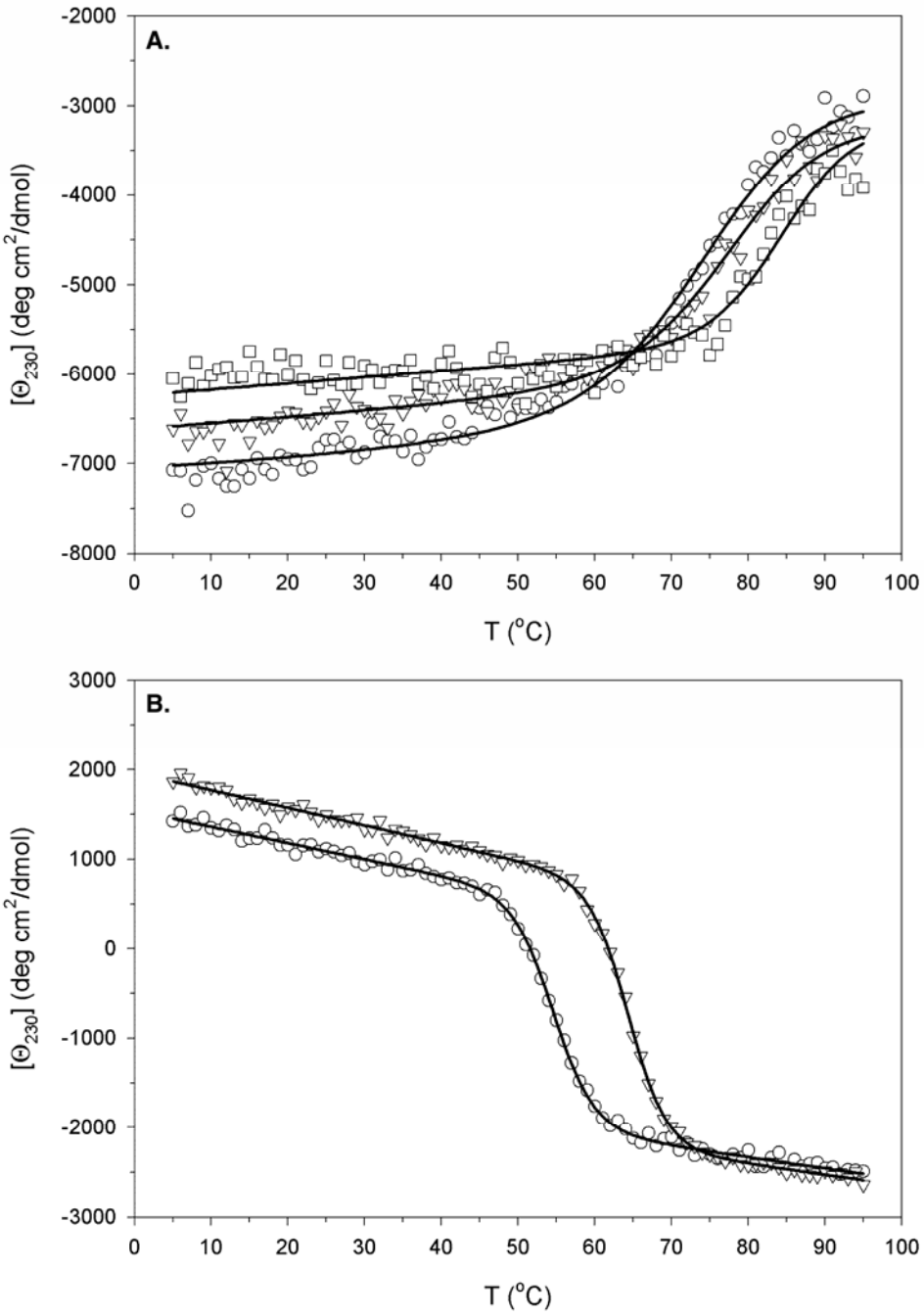

Figure 5.2 Thermal denaturation of procarboxypeptidase (Pc) and tenascin (Ten) variants, monitored by CD spectroscopy. **A.** The unfolding transitions of Pc-WT (○), Pc-GA1 (□), and Pc-GA2 (□) were monitored at a wavelength of 222 nm. **B.** The unfolding transitions of Ten-WT (○) and Ten-GA1 (□) were monitored at a wavelength 230 nm. In both **A** and **B**, the symbols represent the experimental data, and the solid lines represent the fits of the data to a two state model of unfolding.

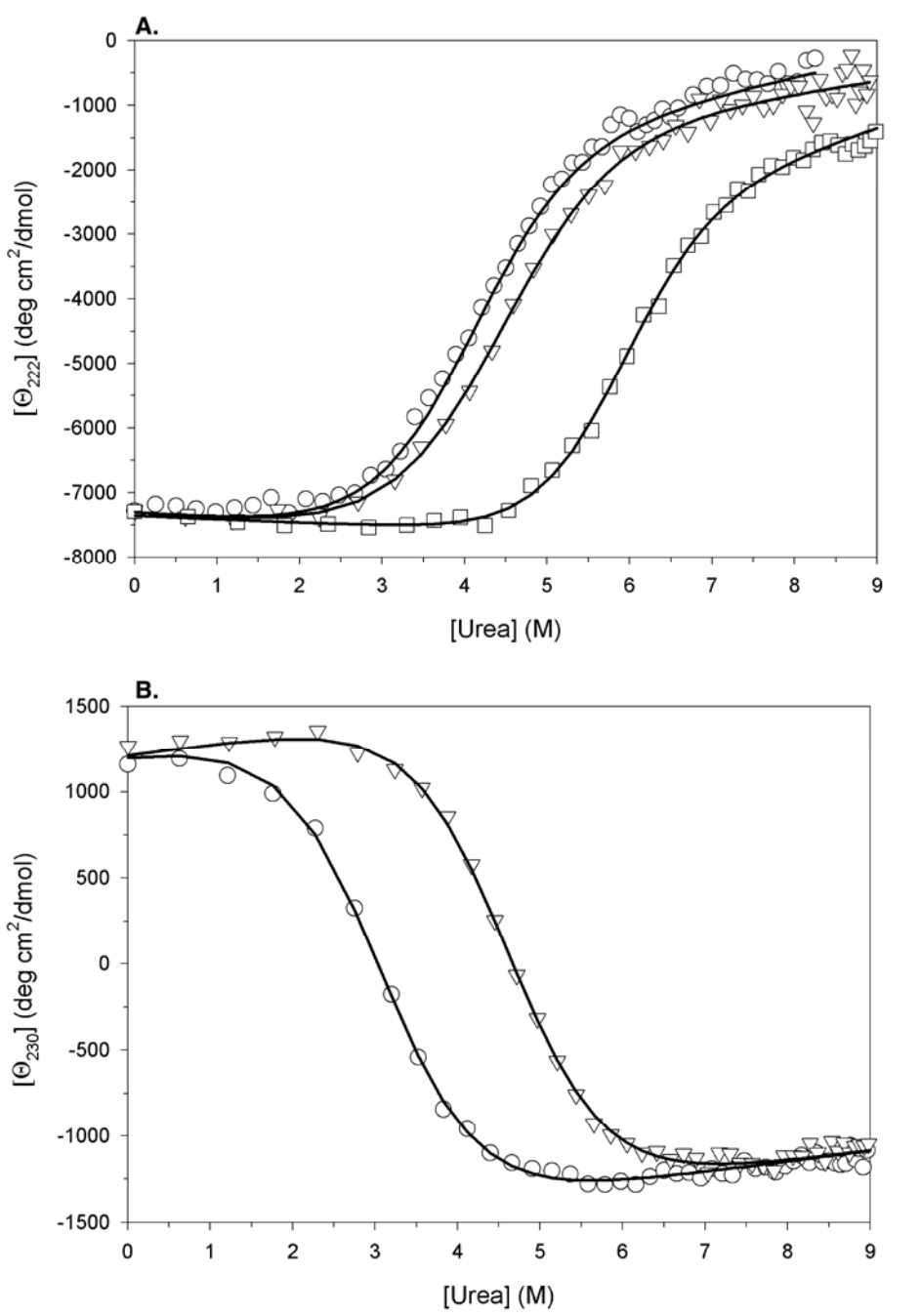

Figure 5.3 Urea denaturation of Pc and Ten variants, monitored by CD spectroscopy. **A.** The unfolding transitions of Pc-WT (○), Pc-GA1 (□), and Pc-GA2 (□) were monitored at a wavelength of 222 nm. **B.** The unfolding transitions of Ten-WT (○) and Ten-GA1 (□) were monitored at a wavelength 230 nm. In both **A** and **B**, the symbols represent the experimental data, and the solid lines represent the fits of the data to a two state model of unfolding.

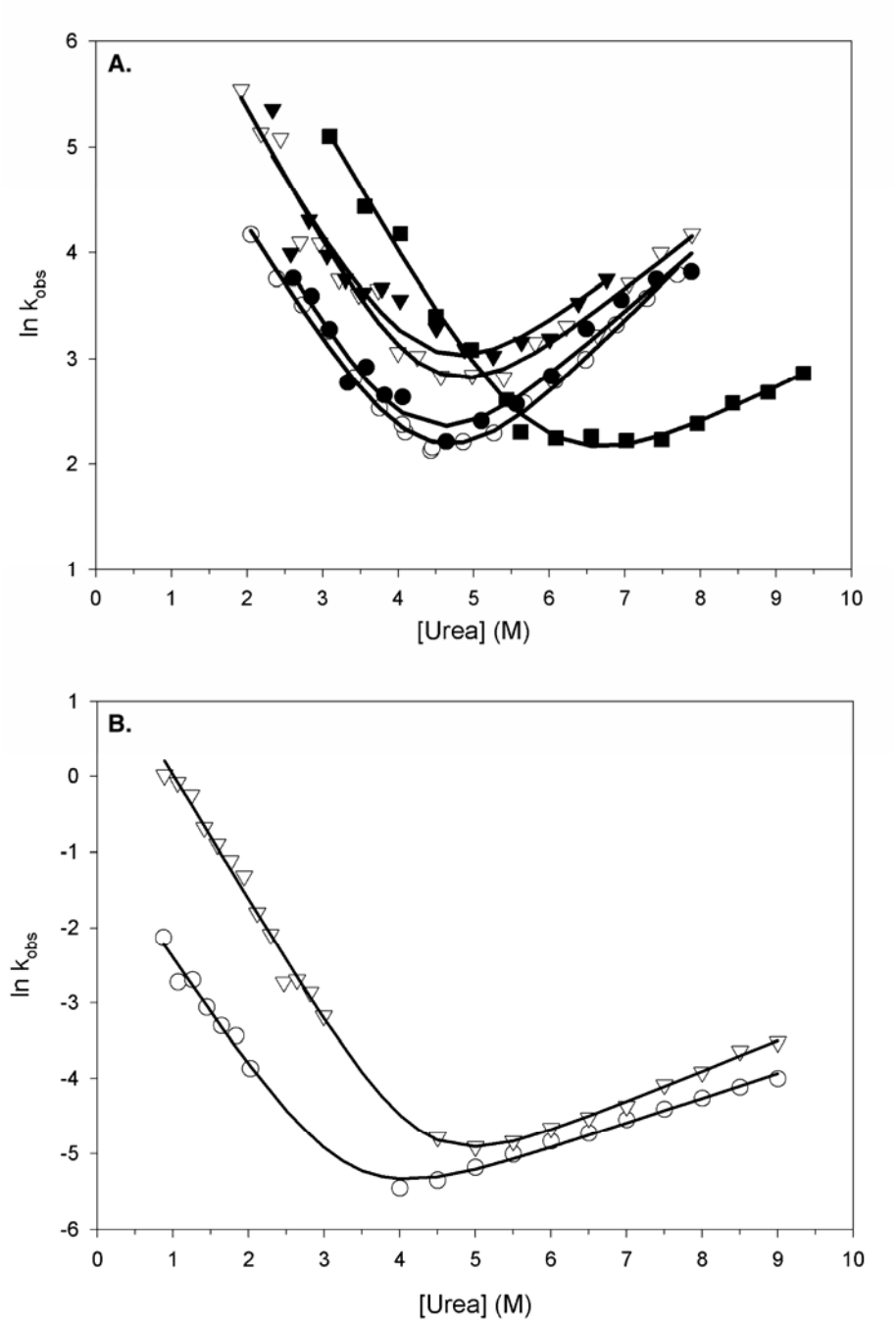

Figure 5.4 Chevron plots for Pc (**A**) and Ten (**B**) variants. The open symbols represent the data from fluorescence experiments, the closed symbols represent data from CD experiments, and the solid lines are the fits of the data to a two-state model of unfolding. **A.** Pc-WT (○,●), Pc-GA1 (□, ▼), and Pc-GA2 (■). **B.** Ten-WT (○), Ten-GA1 (□).

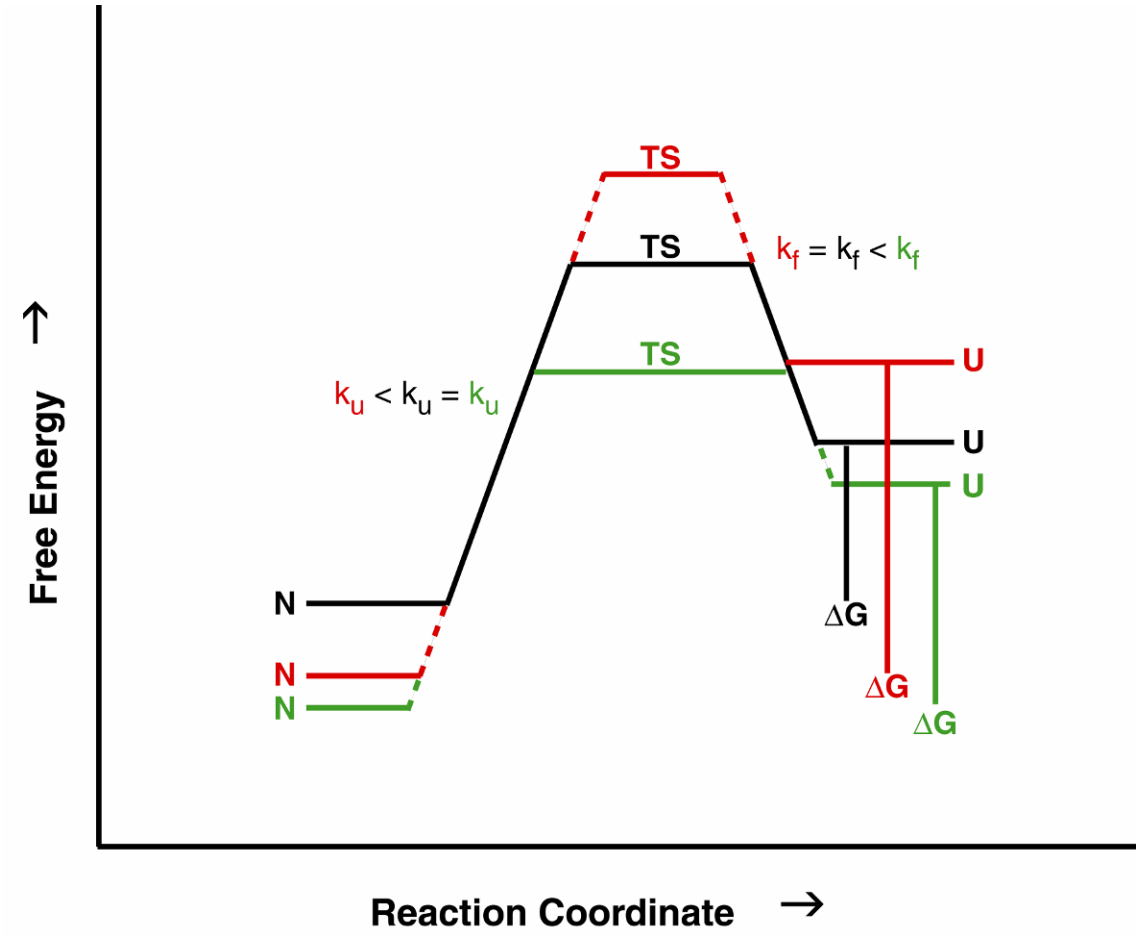

Figure 5.5 Two kinetic models of unfolding if substitutions can affect denatured state ensemble. In both models, the thermodynamic stability is increased relative to the reference state (black lines), but the effects on the folding and unfolding rates are very different. The red model provides a schematic explanation of how the optimization of CspB affected the folding and unfolding kinetics due to substitutions forming favorable interactions in the native state *and* unfavorable interactions in the unfolded state. The green model provides a diagram to explain how the optimization of the Fyn, Pc, and Ten variants could affect folding and unfolding kinetics due to the increase in favorable interactions in both the native and unfolded states. See the text for a detailed explanation of the models.

CHAPTER 6: RATIONAL DESIGN OF SURFACE CHARGES PROTECTS PROTEINS FROM AGGREGATION UPON THERMAL DENATURATION

6.1 Introduction

It is well established that both short- and long-range surface charge-charge interactions can contribute favorably and significantly to protein stability (*11-13, 15, 16, 41, 68, 97, 103, 119, 187, 232*). Indeed, the increased presence of charged residues on the protein surface seems to be one of the primary differences between thermophilic proteins and their mesophilic counterparts (*48, 72, 233, 234*). The preceding chapters of this thesis have discussed in great detail how the surface charge-charge interactions can be modulated to change both the thermodynamic and kinetic stabilities of many different proteins.

The solubility of a protein is another aspect of protein chemistry where charged residues on the surface of the protein are important (*233, 235-237*). This empirical knowledge has been used to modulate the solubility of proteins in several different ways (*236, 238-243*). Short peptide tags comprised primarily of Lys or Arg residues, have been used to improve the solubility of hydrophobic proteins and peptides by as much as 6-fold (*241, 242*). Several groups also used structural-based engineering approaches to increase the solubility of proteins (*237-240, 243*), in one case making a membrane protein soluble in water (*240*). These observations suggested it might also be possible to decrease the propensity for aggregation upon denaturation by increasing the net charge of a protein, in a process that has come to be called “supercharging”. Indeed, increasing the number of charged residues on the protein surface was shown to be a viable approach to decrease aggregation propensity by Lawrence, et al (*243*), who engineered “supercharged” variants of three different proteins known to be prone to aggregation: streptavidin, glutathione-S-transferase (GST), and green fluorescent protein (GFP). The supercharged variants of each protein were created by substituting a number of surface positions

with either basic residues (net charge at least +30) or acidic residues (net charge at least -25). Supercharging appeared to have small effects on the function of these proteins, as the streptavidin variants were still able to form tetrameric structures and interact with biotin, although the binding capacity was decreased (*243*). The GST variants were also able to dimerize after supercharging and to retain function, with similar K_M and k_{cat} values as the wild-type protein. The supercharged variants of both streptavidin and GST showed no detectable aggregation upon heating. Furthermore, the recovery of some enzymatic activity upon cooling of the supercharged GST variants indicated that the unfolding was reversible (*243*).

The supercharged GFP variants were also significantly less prone to aggregation upon thermal denaturation (*243*). However, when the GuHCl-induced denaturation of wild-type GFP and its supercharged variants were characterized, it was discovered that the supercharged variants were significantly less stable than the wild-type protein (*243*). It is possible that this was due to the engineering procedure, which selected residues for substitution based solely on the apparent solvent accessibility of the side chain in a crystal structure. The contribution of charges at these positions to the Gibbs free energy of unfolding (ΔG) was not considered. One way to ameliorate this issue is to use a rational design approach to select for supercharged variants that contain substitutions at positions that do not have significant contributions to the stability. This should result in a protein that is more soluble without sacrificing stability. The ability to prevent protein aggregation has exciting implications for the biotechnology field, although it will be the most useful if aggregation can be prevented without losing stability. This chapter discusses the rational design of a supercharged variant of ubiquitin (Fig. 6.1A), which is predicted to have improved solubility, relative to the wild-type molecule, without a significant loss of stability.

6.2 Results & Discussion

6.2.1 Design of supercharged ubiquitin

The TK-SA model (*11, 15, 16, 67, 97, 110, 187*) was used to rationally design a supercharged variant of ubiquitin (Ubq-SC). In the original TKSA-GA method, the selection of sequences using the genetic algorithm (GA) is based solely on an increase in favorable charge-charge interaction energies. To adapt this algorithm for rationally designing a supercharged protein with increased solubility (SCTKSA-GA), the selection in based on two criteria: the energy of charge-charge interactions should be at least as favorable as in the WT, and the net charge of the protein at a given pH must be larger than a preset value. Figure 6.2 shows the plot of the total energy of charge-charge interactions (ΔG_{qq}) versus the net charge of the protein at pH 7.5. Each symbol represents one of the sequences identified by SCTKSA-GA as satisfying both criteria. From Fig. 6.2, it is clear that there are many sequences that have both more favorable ΔG_{qq} than that of wild-type ubiquitin (Ubq-WT, -12.4 kJ/mol) and a net charge that is larger than the ~0 net charge of Ubq-WT. From this figure, it is also evident that the number of sequences identified as having a much higher net charge than Ubq-WT, but more favorable ΔG_{qq}, decreases with increasing net charge. In fact, at a net charge of ~ +11.5 there are only a handful sequences that have more favorable ΔG_{qq} than Ubq-WT.

Figure 6.1B shows a sequence alignment of the sequences that were identified to have a more favorable ΔG_{qq} than Ubq-WT at a net charge of ~ +11.5. The number of substitutions (15 – 17) in each sequence is also indicated in this figure. From Figure 6.1B, it can be seen that most of the sequences have substitutions at the same positions, so as the very first test of the idea that we should be able to make ubiquitin less prone to irreversible aggregation by increasing the net

charge, we selected sequence #5, which contained 15 amino acid substitutions. For the remainder of this chapter, this ubiquitin variant will be referred to as supercharged ubiquitin (Ubq-SC).

The Ubq-SC variant selected for experimental characterization contains four Glu and 11 Lys substitutions on the protein surface, resulting in a net charge of +12 at pH 7.0 (Fig. 6.1A & B). In contrast, wild-type ubiquitin has no net charge at pH 7.0. An examination of ΔG_{qq} on a per residue basis (Fig. 6.1C) indicates that these substitutions are often more unfavorable than the wild-type residue. However, there are a few positions where the substitution increases the favorable charge-charge interaction energies, which results in a similar total energy of charge-charge interactions for Ubq-WT and Ubq-SC. Based on the results of these calculations, we predict that Ubq-SC should be more soluble and less prone to aggregation upon unfolding than Ubq-WT, without adversely affecting the stability.

6.2.2 Experimental characterization of supercharged ubiquitin

Due to the large number of substitutions on the protein surface, there is a possibility that Ubq-SC could have an altered oligomeric state, relative to Ubq-WT. Analytical ultracentrifugation (AUC) was performed to determine whether Ubq-SC retained the monomeric behavior of Ubq-WT (Fig 6.3). The sedimentation equilibrium experiments were performed at three different speeds, and the data for each variant were globally fit to a single species model. The molecular masses measured by AUC (WT – M_{AUC} = 7.5kDa, SC – M_{AUC} = 9.9kDa) are similar, within experimental error, to those expected based on amino acid composition (WT – M_{TH} = 8.4kDa, SC – M_{TH} = 8.9kDa), indicating that both proteins are indeed monomeric.

Another possible effect of making such a large number of substitutions as in Ubq-SC, is that these substitutions can alter the structure of the protein. Far-UV circular dichrosim (CD) spectroscopy was used to obtain low resolution information about the secondary structural content of Ubq-WT and Ubq-SC. The CD spectra of the wild-type and supercharged ubiquitin

variants are shown in Figure 6.4. From this figure, it can be seen that the far-UV CD spectra of both ubiquitin variants are quite similar, suggesting that the secondary structure is not affected by the large number of substitutions in the supercharged variant.

Thermal denaturation of the ubiquitin variants were characterized using DSC at pH 5.0 and pH 7.0 in order to characterize both the stability and aggregation propensity of Ubq-SC relative to Ubq-WT. The heat capacity profiles of Ubq-WT at pH 5.0 and pH 7.0 both show an interesting behavior (Fig. 6.5). The temperature dependence of the partial molar heat capacity indicates that wild-type ubiquitin is soluble up to 70 °C. However, as soon as the thermally induced unfolding transition starts, wild-type ubiquitin begins to aggregate. Aggregation in DSC profiles is manifested as a sharp decrease in the heat capacity due to a dramatic release of heat upon the formation of aggregates. It appears that Ubq-WT is slightly less prone to aggregation at pH 5.0 (Fig. 6.5A) than at pH 7.0 (Fig. 6.5B) because the aggregation occurs later in the unfolding transition. In contrast, the Ubq-SC variant, does not aggregate upon thermal denaturation at either pH. The reversibility of the unfolding of Ubq-SC was tested by rescanning the sample after cooling to 5 °C. It can be seen from Fig. 6.5 that most of the original signal is recovered in the rescanned sample, indicating that the thermal denaturation of Ubq-SC is also highly reversible. The dramatic increase in reversibility that is observed in the Ubq-SC variant is consistent with the prediction that supercharging should increase the solubility of ubiquitin. The increase in solubility appears to have a more pronounced effect on the unfolded state, as the mechanism of the thermally induced irreversibility of wild-type ubiquitin is related to the aggregation of the unfolded state.

Figure 6.5 reveals another interesting and important feature of the stability of the supercharged ubiquitin variant: the T_m of Ubq-SC is lower than the aggregation temperature of Ubq-WT. Since the aggregation appears to occur before Ubq-WT is fully unfolded, this observation would suggest that Ubq-SC is destabilized relative to wild-type ubiquitin. This result

contrasts with our predictions based on the TK-SA model. The calculated charge-charge interaction energies of Ubq-WT and Ubq-SC are very similar (Fig. 6.1B), suggesting that these variants should have similar stabilities. Since this is the first time when relative changes in the stability of a designed variant were not correctly predicted by the TK-SA model, we performed further studies on the thermal unfolding of Ubq-SC.

In order to determine the thermodynamic mechanism that causes the destabilization of Ubq-SC, DSC experiments were performed at pH 3.5, pH 3.75, and pH 4.5 (Fig. 6.6). The destabilization of Ubq-SC is even more evident under acidic pH conditions, where the heat capacity profiles indicate that Ubq-SC is not fully folded, even at low temperatures. In contrast, wild-type ubiquitin is fully folded at low temperatures, with a high T_m under these conditions (*244*). Figure 6.7 shows the dependence of the calorimetric enthalpies of unfolding (ΔH_{cal}) as a function of transition temperature (T_m) for Ubq-WT (data from (*244*)) and Ubq-SC. The slope of the $\Delta H_{cal}(T_m)$ function represents the change in heat capacity upon unfolding (ΔC_P). From Fig. 6.7, it can be seen that the ΔC_P for Ubq-SC is similar to that of Ubq-WT, within experimental error. This is to be expected because these two proteins differ only in the residues on the protein surface, while the residues that are buried in the core of the native states of Ubq-WT and Ubq-SC are the same. Since the exposure of buried residues is what largely defines the values of ΔC_P, the slopes of the $\Delta H_{cal}(T_m)$ versus T_m plots shown in Figure 6.7 should be similar.

Interestingly, the enthalpy of unfolding of Ubq-SC is significantly (over 50 kJ/mol) lower than that of Ubq-WT (Fig. 6.7). A similar dramatic decrease in the enthalpy of unfolding has been previously observed for the CpsB-Bs and CspB-TB pair (*15*), which also differ primarily in the number of charged residues on the protein surface. In particular, the magnitude of the change in the net charge going from CspB-Bs (-5) to CspB-TB (+3), is remarkably similar to the magnitude of the relative change in net charge going from Ubq-WT (~0) to Ubq-SC (+11). However, the substitutions in CspB-TB resulted in an increase in thermostability (T_m) such that

the dramatic decrease in enthalpy resulted in thermodynamic stabilities ($\Delta G(25°C)$) that were similar for CspB-Bs and CspB-TB. In the case of Ubq-SC, however, the large change in the net charge of the protein resulted in a decreased thermostability, which ultimately manifests as a lower thermodynamic stability at room temperature ($\Delta G(25°C)_{WT}$ = 30.5 kJ/mol vs. $\Delta G(25°C)_{SC}$ = 5.2 kJ/mol, based on pH 3.5 data). Based on the these results, the destabilization of Ubq-SC appears to be primarily enthalpic in nature ($\Delta H_{cal,SC} < \Delta H_{cal,WT}$), with very little affect on ΔC_P.

The observation that Ubq-SC is destabilized relative to Ubq-WT contradicts our predictions that the substitutions should have little effect on the stability of the protein (Fig. 6.1 & Fig. 6.2). One explanation for this behavior stems from the fact that there are 15 substitutions in Ubq-SC, relative to Ubq-WT, with four residues substituted to Glu and 11 to Lys. Homology modeling was used to generate structures of Ubq-SC for calculating charge-charge interaction energies using the TK-SA model. It is possible that homology modeling fails to correctly predict the positions of the surface residues due to the large number of substitutions in Ubq-SC. Although the far-UV CD spectra of Ubq-WT and Ubq-SC suggest that the structures of these two proteins are similar, it is important to mention that far-UV CD only probes the secondary structural composition of the protein. It is possible for two proteins to have similar secondary structures, as measured by far-UV CD, while having different tertiary contacts. The TK-SA model calculates the charge-charge interaction energies based on the distances between charged residues, and for long-range interactions, these distances will be primarily determined by the tertiary structure of the protein, rather than the secondary structure. Therefore, it is possible that such a large number of substitutions alter the structure of the protein in a way that is not possible to detect by far-UV CD spectroscopy. If the structure of Ubq-SC is indeed different from the structure generated with homology modeling, then the TK-SA model might not correctly predict how the substitutions will affect the stability of Ubq-SC.

Another explanation for the decreased stability of Ubq-SC is that the overall net charge of Ubq-SC is +12, compared to a net charge of zero for Ubq-WT. It is likely that this high positive charge density results in charge-charge repulsions that decrease the stability of the protein. In other words, after a certain number of substitutions, the environment of the protein surface was such that the addition of more charged residues to the protein surface created only unfavorable interactions. Perhaps it will be necessary to supercharge proteins in such a way that the total number of charged residues is increased, but the net excess charge density is lower. If the overall net charge is reduced, then the repulsive interactions due to having large numbers of like-charges could also be decreased. This could make it possible to make a protein more soluble without sacrificing stability (*236*).

It is also possible that increasing the solubility of any protein will inherently make it less stable. The basis for this hypothesis is that one of the major forces contributing to protein stability is the hydrophobic effect. In other words, proteins fold in aqueous environments due to the significant entropic penalty that is associated with water organizing itself around hydrophobic components of proteins, combined with the favorable energetics of the hydrophobic parts of proteins interacting with each other. It is possible that the increased number of surface charges will also have effects on the energetic of the unfolded state ensemble. If the charges are spaced along the sequence in such a way that there is less hydrophobic surface area exposed in the unfolded state, then the unfavorable entropy of the unfolded state due reorganizing bulk water will be decreased. In this case, solubilizing proteins by increasing the number of charges on the surface could decrease the effective strength of the hydrophobic effect as a driving force for protein folding and stability. Coupling the weakened hydrophobic effect with the repulsive charge-charge interactions that are likely to occur in the native state of a highly charged protein, could explain why all supercharged protein variants characterized to date are destabilized relative to their wild-type counterparts (*243*).

6.3 Concluding Remarks

The results presented in this chapter highlight the complex relationship between protein stability and solubility. Understanding the mechanisms driving both phenomena is important for gaining a full understanding of how proteins fold and interact with their environments. In this very first attempt, we were able successful in designing a soluble protein. We were, however, unable to do so without adverse effects on the stability. Nevertheless, these results help dispel a common belief that proteins aggregate because they are not stable, as they clearly demonstrate that the thermodynamic stability of a protein does not define its aggregation propensity and vice versa. Rather, the physico-chemical forces that govern the intramolecular interactions within a protein and define its stability likely have a different hierarchy of importance than those that govern protein-solvent interactions and define protein solubility.

To gain a better understanding of the forces that dictate stability and solubility and the relationship between the two, it will be necessary to study this problem in greater detail. For example, we have decreased the aggregation propensity of Ubq-WT by supercharging with basic residues, but it has been shown that supercharging proteins with acidic residues is also an effective way to increase their solubility (*243*). It is possible that rationally designing solubility with negatively charged residues could decrease the error of our predictions because the side chains of negatively charged residues are shorter with fewer degrees of freedom than the side chains of basic residues. If the error in our stability predictions is primarily caused by incorrect predictions of the relative positions of the side chains in such a heavily charged protein, then using shorter, less flexible side chains could alleviate this problem. Furthermore, supercharging proteins with negatively charged residues provides a way to increase solubility under a wider variety of conditions than would be possible if only positively charged residues were used.

To address the question of whether or not homology modeling is sufficient for predicting the structure of a protein with a large number of substitutions to charged residues, it will be important to use higher resolution methods, such as NMR spectroscopy to solve the structure of Ubq-SC. If increasing the net charge on a protein really does have a dramatic effect on its tertiary structure which is not accurately predicted with homology modeling, then a high resolution structure of Ubq-SC would allow us to have a better template for understanding how increasing the net charge on a protein affects its structure. A high resolution structure could also provide a better template for understanding the effects of supercharging on the structures of other proteins, which would help improve the accuracy of our predictions.

If we can determine how to decrease the aggregation propensity of proteins without affecting their stabilities or functions, then supercharging could also provide a new method for studying the mechanisms and kinetics of aggregation. For example, by making a series of single, double, triple, or higher order variants of Ubq-WT based on the substitutions in Ubq-SC, we might be able to determine which residues are important in the aggregation pathway of ubiquitin versus those that are important for stability. By performing a similar analysis on several other proteins, it might be possible to develop a general set of rules for selecting residues that contribute to solubility and not stability. If similar studies were also performed on proteins or peptides that aggregate with regular structures, Aβ peptide, it might also be possible to determine what dictates aggregation in the form of fibril formation versus the "amorphous" aggregates of proteins like ubiquitin. A comprehensive understanding of the forces dictating protein aggregation is likely to be necessary for developing the most effective treatments to diseases caused by this behavior.

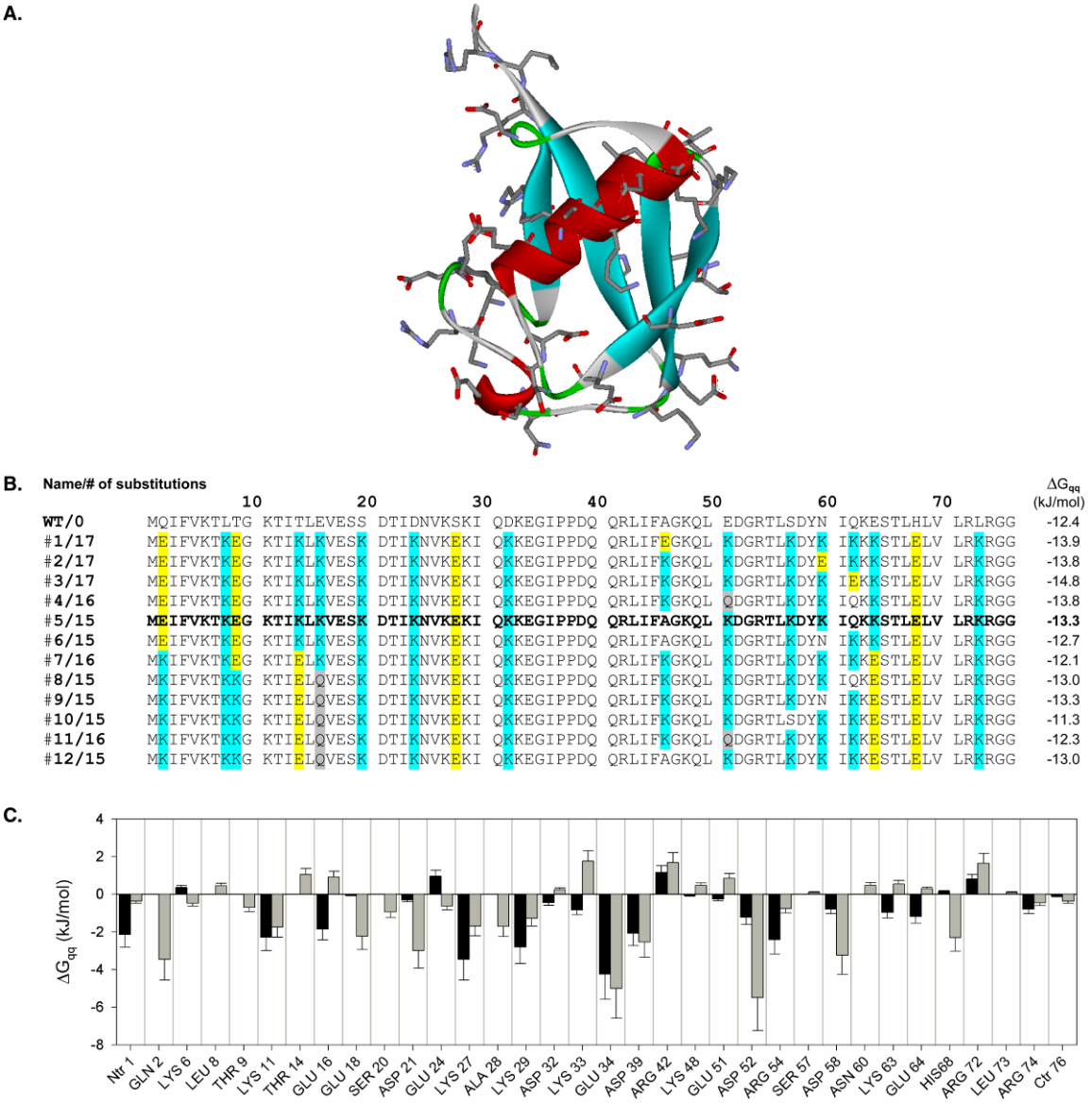

Figure 6.1 Results of SCTKSA-GA predictions for supercharged ubiquitin. **A.** Cartoon representation of ubiquitin (PDB code: 1UBQ). The side chains that were subjected to substitution are shown as line representations. **B.** Sequence alignment of wild-type ubiquitin and 12 candidate sequences for supercharged ubiquitin. All sequences have 15 to 17 substitutions, resulting in net charges between 11.5 and 11.7 at neutral pH. Substitutions to acidic residues (Glu) are represented by yellow squares; substitutions to basic residues (Lys) are highlighted by light blue squares, and substitutions to neutral polar residues (Gln) are indicated by grey squares. All 12 candidate sequences had similar charge-charge interaction energies (ΔG_{qq}) to each other, and to the wild-type, so as the first test of the model, sequence #5 (bold text) was selected for further characterization. **C.** The energy of charge-charge interactions on a per residue basis. The black bars represent Ubq-WT, and the grey bars represent Ubq-SC. The error bars are calculated from an average of 11 structures. Favorable contributions to the energy of charge-charge interactions are denoted by negative values of ΔG_{qq}.

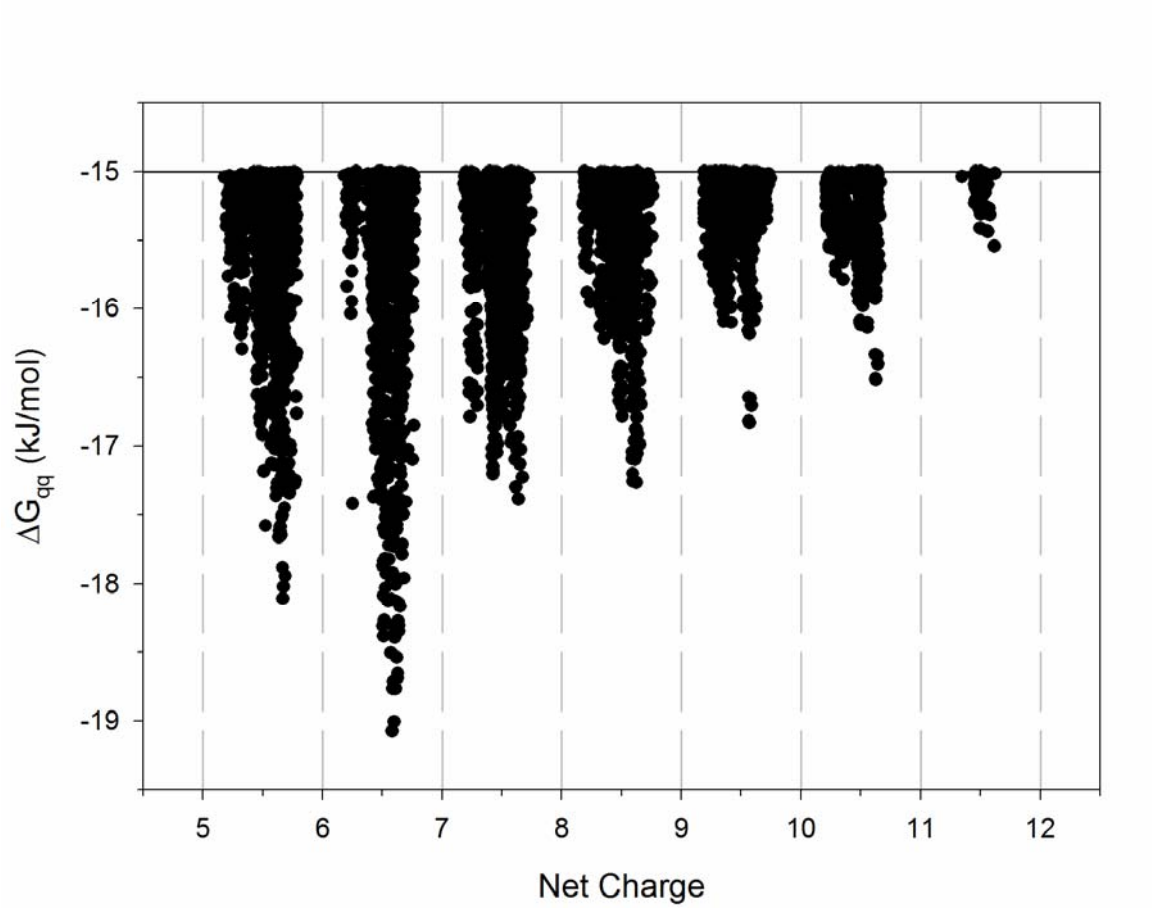

Figure 6.2 Total energy of charge-charge interactions (ΔG_{qq}) as a function of the net charge of the sequence. The number of sequences that are identified by the SCTKSA-GA algorithm as satisfying both selection criteria decreases as the net charge of the sequence increases. Furthermore, as the net charge of a sequence increases, the ΔG_{qq} seems to become less favorable. However, all sequences represented here have more favorable ΔG_{qq} than Ubq-WT.

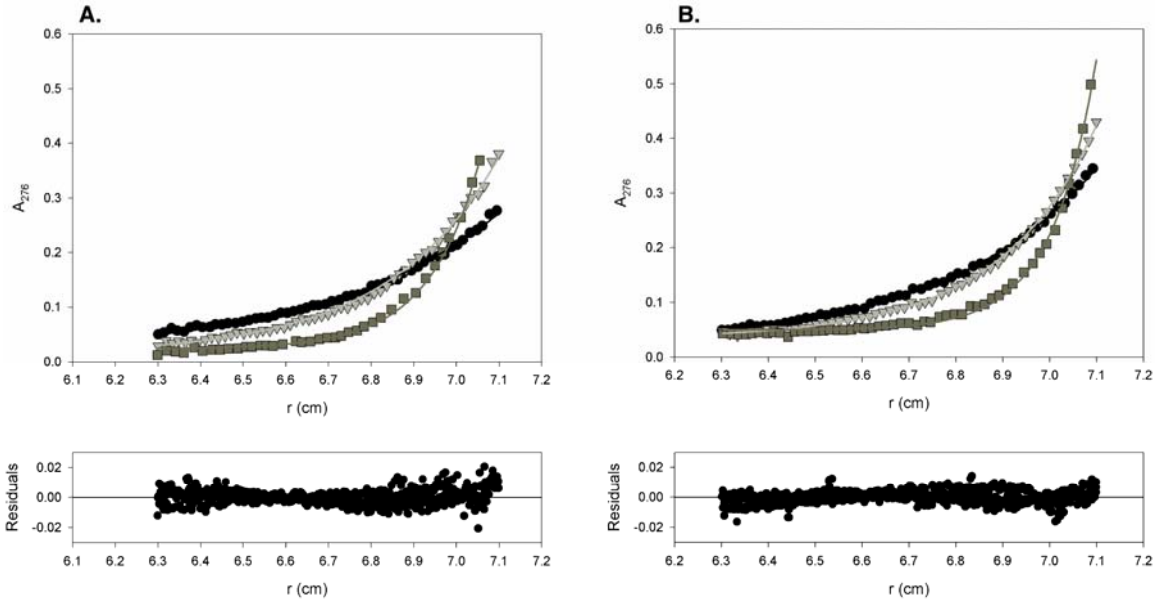

Figure 6.3 Analytical Ultracentrifugation profiles for **A.** Ubq-WT and **B.** Ubq-SC. The data were collected at three speeds: 20,000 rpm (●), 25,000 rpm (▼) and 37,000 rpm (■) and globally fit to a single species model. The resulting molecular weights suggest that both Ubq-WT and Ubq-SC are monomeric.

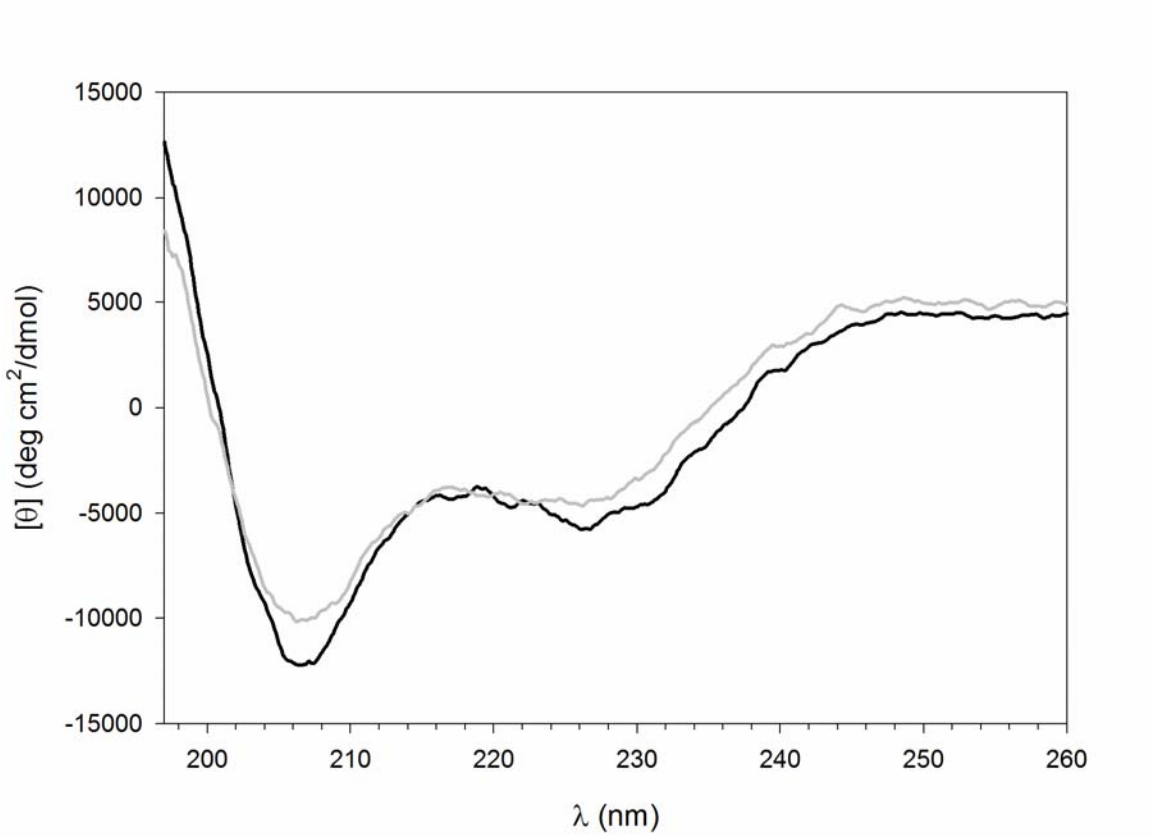

Figure 6.4 Far-UV CD spectra of Ubq-WT (black line) and Ubq-SC (grey line). The CD spectra were measured at 25 °C in a 1 mm barrel cuvette, with protein concentrations of 0.05 mg/mL. The spectrum of Ubq-SC is similar to that of Ubq-WT, suggesting that the large number of substitutions in Ubq-SC do not perturb the secondary structural content of the protein.

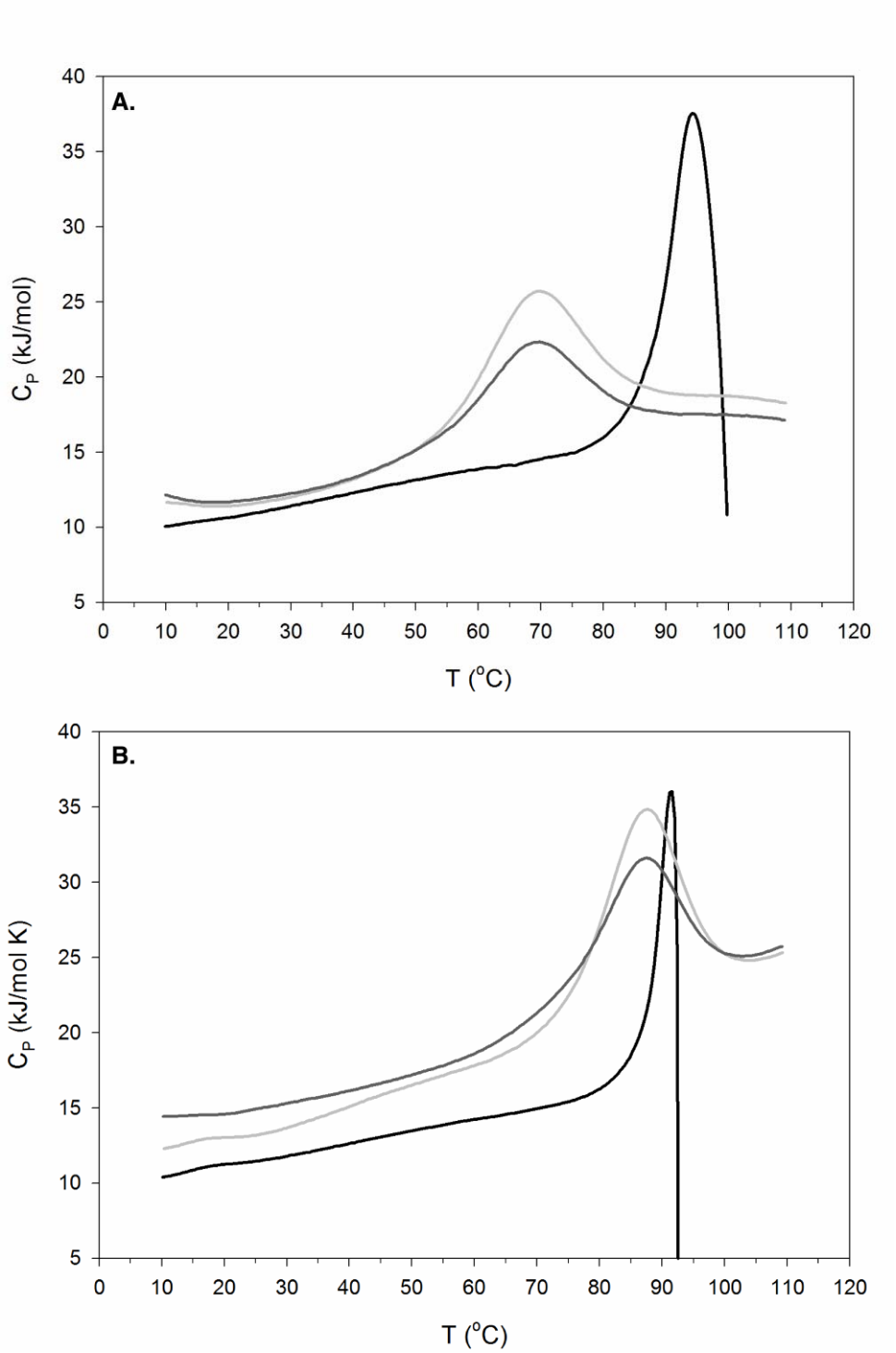

Figure 6.5 Heat capacity profiles measured by DSC for the thermal unfolding of Ubq-WT (black lines) and Ubq-SC (grey lines) at **A.** pH 5.0 and **B.** pH 7.0. Ubq-WT undergoes irreversible unfolding under both experimental conditions, whereas, Ubq-SC unfolds reversibly (dark grey lines).

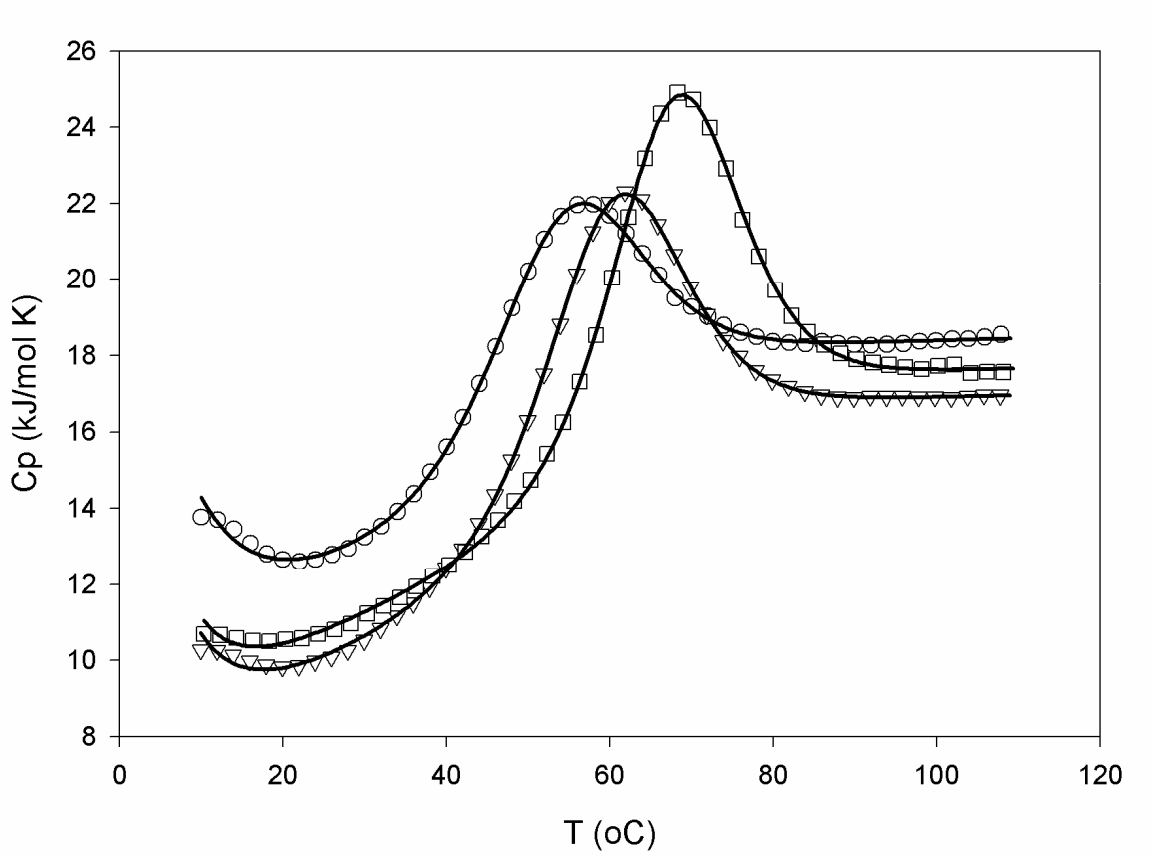

Figure 6.6 pH-dependent heat capacity profiles for Ubq-SC at pH 3.5 (○), pH 3.75 (▽), and pH 4.5 (□). The symbols represent the experimental data, shown every 5°C for clarity. The solid lines represent the fits of the data to a two-state model of unfolding. Ubq-SC shows evidence of cold denaturation under acidic conditions.

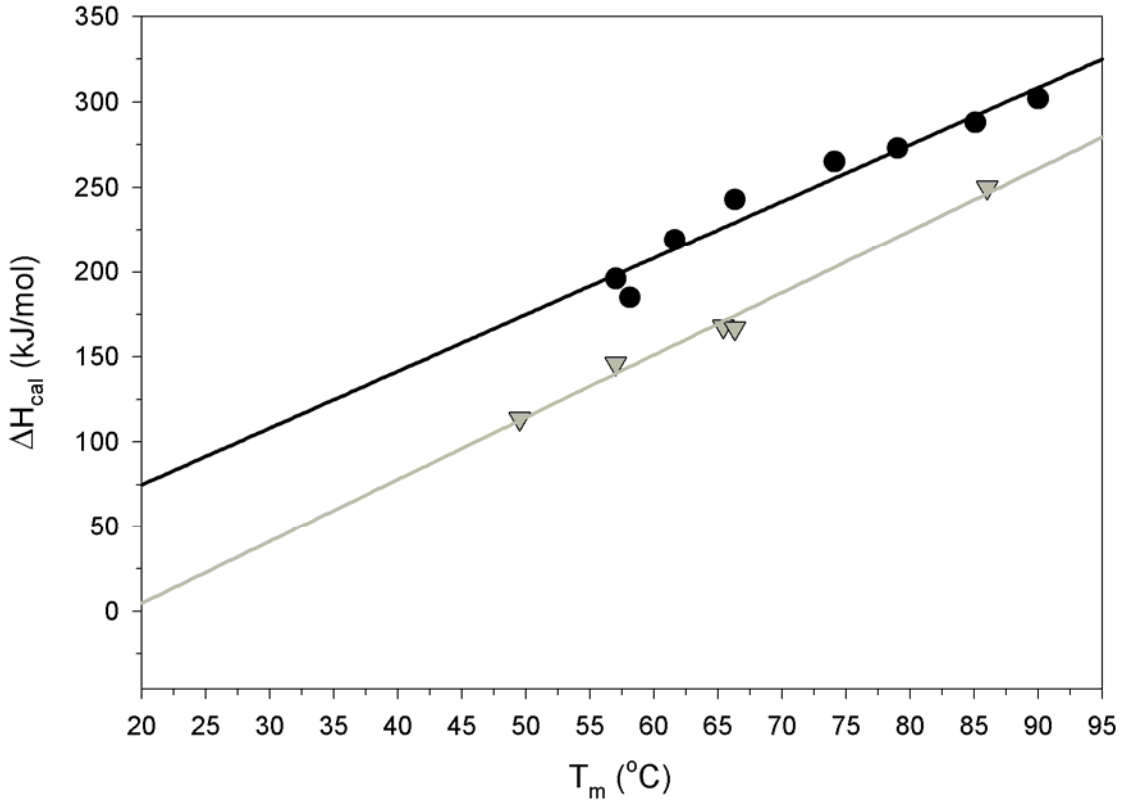

Figure 6.7 The calorimetric enthalpies of unfolding (ΔH_{cal}) as a function of the transition temperature (T_m) of Ubq-WT (●, taken from ref. (*244*)) and Ubq-SC (▼). The slopes of the lines represent the ΔC_P of unfolding of the proteins, and are similar, within the error of the experiment (3.3 ± 0.3 kJ/mol K for Ubq-WT and 3.6 ± 0.2 kJ/mol K for Ubq-SC). This suggests that the destabilization of Ubq-SC is enthalpic in nature.

CHAPTER 7: THERMODYNAMIC CHARACTERIZATION OF GLOBULAR PROTEINS USING PRESSURE PERTURBATION CALORIMETRY

7.1 Introduction

Understanding the forces that govern protein stability and solubility is the focus of this thesis. The work discussed in the previous chapters has addressed the questions of how surface charges play a role in both the stability and solubility of proteins in solution; how protein stabilization affects the kinetics of folding and unfolding reactions; and the role of the unfolded state in protein stability. However, there are still some aspects regarding the thermodynamic characterization of protein stability which have yet to be extensively studied, such as the response of proteins to pressure and the volumetric changes that occur upon unfolding. A more complete understanding of how intramolecular interactions govern protein stability can only be obtained once these areas are as well studied as other biophysical responses, such as thermal denaturation.

The study of the transfer free energies of model compounds from liquid hydrocarbons to water has been very successful in helping to understand the relative contributions of intramolecular interactions, such as the hydrophobic effect, to thermal and chemical denaturation (i.e. urea- or GuHCl-induced) of proteins (*1, 166, 245-256*). Unfortunately, these studies failed to explain the volumetric changes that should occur upon isothermal pressure-induced denaturation (*257*). Based on model compound data, the solvation of polar groups and the transfer of nonpolar groups from a hydrophobic to an aqueous environment were both expected to contribute negatively to the changes in the specific volume of a protein upon unfolding (*1, 258-260*). The change in the intrinsic void volume (volume of cavities) of proteins was also expected to have a negative contribution to the volumetric changes upon unfolding. As a result of these measurements, it was believed that the unfolding of proteins should be accompanied by a large decrease in their specific volumes. However, in most cases, only small decreases, or even small

increases, in the partial specific volume of proteins upon unfolding were observed (*179, 258, 261-263*).

Chalikian and Breslauer (*261*) were the first to try to resolve this issue by introducing the concept of thermal volume, such that the specific volume of a protein ($\bar{v}_p$) is actually made up of three components:

$$\bar{v}_p = v_{\text{int}} + \Delta v_{hyd} + v_t \qquad \textbf{(7.1)}$$

where v_{int} is the intrinsic volume of the protein, which is a sum of the van der Waals volumes of all atoms in the protein and the internal cavities; Δv_{hyd} is the volume change in the solvent due to the hydration of the solvent accessible surface of the protein; and v_t is the thermal volume that results from the thermally induced molecular vibrations of the protein and solvent. The effect of the thermal volume is to expand the solvent away from the surface of the protein, such that solvent-free volume element forms around the protein. It is possible then, that the negative contribution of Δv_{hyd} for hydrophobic residues measured in the model compound studies is simply a reflection of the lower thermal volume of water compared to nonpolar solvent (*261*). In addition, the protein interior is most likely denser and more heterogeneous than a nonpolar solvent, providing another explanation for why the model compound studies were unable to accurately describe the volumetric changes in proteins upon unfolding (*261, 262*). Chalikian and Breslauer (*261*) also demonstrated how the thermal volume of the solvent can compensate for the negative changes in v_{int} and Δv_{hyd} in such a way that the overall $\Delta\bar{v}_p$ is only slightly negative. Furthermore, if these three contributions to specific volume respond differently to changes in temperature or pressure, then it is possible that the protein could react such that $\Delta\bar{v}_p$ can also be positive (*261*). Indeed, such behavior has been observed for a number of proteins (*137, 179, 261, 264*).

The idea that $\bar{v}_p$ can change with response to pressure, as well as temperature, led to the development of methods to study various aspects of proteins at high pressure, which makes it possible to stabilize conformational states that are not usually populated enough to be studied under standard conditions (i.e. atmospheric pressure, ~ 14.7 psi). These methods have been successfully used to study the structures of kinetic intermediates and protein aggregation pathways (for a review, see (*265*)). However, one of the aspects of the pressure-volume relationship that is difficult to measure using the high-pressure techniques of densitometry, FITR, or SAXS was the thermal expansivity coefficient, α(*T*). This parameter is the temperature derivative of the *V*(*T*) function, so methods that measure only the volume of a protein will inherently have large errors in α(*T*).

Pressure perturbation calorimetry (PPC) is a relatively new experimental method that overcomes the problems associated with indirect measurements of α. In a PPC experiment, α is measured directly as the difference between the heats produced by a calorimetric cell containing dilute protein solution and that of a cell containing only buffer as they are subjected to rapid changes in pressure (~ 80 psi) under isothermal conditions. By performing PPC at a series of different temperatures, it is possible to measure α as a function of temperature (*179*). A few PPC experiments on several different proteins have demonstrated how this valuable biophysical technique can be used, not only to measure α and ΔV, but also to give information about the interactions between the solvent and proteins in their native and unfolded states.

The first PPC studies focused on developing a framework for understanding pressure-induced protein denaturation. Studying the pressure responses of small molecules (*179, 266*), single amino acids (*179, 266*), and tripeptides (*266*) in water showed how sensitive α(*T*) is to the hydrophobicity of the solute. For example, the polar amino acids tend to have a large, positive value of α at lower temperatures, which decreases as a function of *T*, eventually leveling off at higher temperatures. In contrast, the hydrophobic amino acids tend to have large, negative values

of α at low *T*, which increase as a function of temperature, and also level off at high *T* (< 100 ºC). The studies on single amino acids and tripeptides provided another explanation for how model compound studies could fail to predict the volumetric changes upon unfolding. The α(*T*) profiles for a single amino acid, *X*, and its G-*X*-G tripeptide, were remarkably different, and these differences appear to be due to the glycine residues effectively separating the charges of the N- and C- termini of -*X*- (*266*). Since α is the temperature derivative of *V*, the dramatic differences between the α(*T*) profile of a single amino acid and a tripeptide, demonstrate how the extrapolation of model compound data to a full-length protein could fail to predict the magnitude (and/or sign) of ΔV.

PPC experiments have also been performed on several model protein systems (*179, 263, 266, 267*). In addition to measuring the volumetric changes of the proteins upon unfolding, these studies provided insight into what defines the expansivity, α(*T*) of the native and unfolded states of proteins (for an example of α(*T*) profile, see Fig. 7.1 or Fig. 7.2). The general observations from these studies are that proteins with large numbers of hydrophilic residues on the surface have a larger α(*T*) values and a steeper temperature dependence of α(*T*) at lower temperatures (*263, 266*). It also appears that the absolute value and temperature dependence of α(*T*), and resulting ΔV, are highly dependent on the nature of the co-solvent (*179, 266, 267*). For example, in the presence of denaturant, the absolute value of α(*T*) at low temperatures is smaller than in water, and the temperature dependence is shallower. The measured ΔV also changes sign under these conditions (*179*).

It has been discussed that one of the advantages of PPC compared to other methods (i.e. (*264*)) is that it does not rely on the validity of a two-state model for unfolding for interpreting the data, and as such, can be used to measure volumetric changes upon unfolding in a model-independent manner (*179*). In the current analysis of PPC data, the user defines the native and unfolded state baselines, and a progress baseline is extrapolated between them using a 4^{th} order

polynomial function. The volumetric change upon unfolding ($\Delta V/V$) is then determined by calculating the area between the baseline and the experimental $\alpha(T)$ profile. While a model-independent method might produce similar results to those obtained by other high pressure methods that use a two-state model of unfolding to analyze data (*179, 264*), there are instances, such as broad unfolding transitions or cold denaturation, where the baselines are impossible to define given the experimental data. Furthermore, the model-independent method could make it difficult to reproduce results for proteins where $\Delta V/V$ is small, and therefore more sensitive to the baselines. In this chapter, a two-state model is developed for analyzing PPC experiments, and it will be shown that this model is valid for several model systems: hen egg white lysozyme (HEWL), ribonuclease A (RNaseA), ubiquitin (Ubq-WT and Ubq-SC), cytochrome *c* (CytC), and eglinC (EgC). By analyzing PPC data in the context of a two-state model of unfolding, it becomes possible to analyze pressure effects in cases where baselines are difficult to define, and to directly fit the PPC data to get a more complete thermodynamic description of unfolding.

7.2 Description of the Two-State Model for Analyzing PPC Data

The concepts used to analyze the data from PPC experiments are analogous in many ways to those used to analyze the experimental data from DSC. In other words, the relationship of α to $\Delta V/V$ is akin to the relationship between C_P and ΔH. Figure 7.1 highlights some of the similarities and differences between PPC and DSC experiments, and demonstrates some of the considerations that need to be addressed in the analysis of the experimental data. In this figure, the fits of experimental PPC and DSC data for Ubq-WT (Fig. 7.1A) and CytC (Fig. 7.1B) at pH 3.0 are shown. For both Ubq-WT and CytC, the $\alpha(T)$ profiles suggest negative volumetric changes, although the magnitudes of $\Delta V/V$ are markedly different. Notice that the peak of the C_P profile measured by DSC occurs at a similar temperature to the minimum of the $\alpha(T)$ profile

measured by PPC, suggesting that the T_m values measured by each technique are comparable. The area under the $C_P(T)$ profile is the enthalpy of unfolding (ΔH), while the area under the $\alpha(T)$ profile represents the change in volume upon unfolding ($\Delta V/V$).

Since the results of PPC experiments are analogous to those obtained from DSC, and since the analysis of DSC data has is well stabled, it should be relatively straightforward to analyze the PPC data in the context of a two-state model of unfolding, which has two major advantages. First, it provides a standard method to analyze experimental results, which could make it possible to decrease the current errors in the measurement of $\Delta V/V$ that most likely stem from how the experimental baselines are defined (see (*179, 264*)). Since the temperature dependent behaviors of the native and unfolded state baselines are vital for determining the area under the $\alpha(T)$ profile, and hence the volumetric changes upon pressure denaturation, it is important to develop an analysis that will decrease the errors associated with user-defined baselines. The second advantage of developing a two-state model for analyzing experimental data is that such a model will make it possible to fit data in circumstances where current methods fail, such as cold denaturation. This section will describe important features of the two-state model of unfolding that we developed to be able to analyze data from PPC experiments under a wide variety of conditions (also see Fig. 7.2).

7.2.1 Defining experimental baselines

The first step in fitting the data for a PPC experiment is to define the native and unfolded state baselines (Fig. 7.2A). These baselines are important for defining the $\alpha_{progress}(T)$ profile (Fig 7.2B), and for calculating the volumetric changes upon pressure-induced denaturation (Fig 7.2C). In order to define an appropriate function for the unfolded state baseline ($\alpha_U(T)$), its shape was characterized in two different ways: from amino acid composition based on the $\alpha(T)$ profiles of amino acid side chains in water measured by Brandts and co-workers (*179*), and from a PPC

experiment performed on Ubq at pH 3.0. The native state baseline was characterized using CNBr cleaved Ubq (Ubq-WT-CNBr) which is very thermostable and unfolds reversibly at neutral pH, making it possible to monitor the shape of the native baseline over a broader temperature range.

In order to model the unfolded state baselines based on the amino acid composition of proteins, it is important to remember that the molar expansivity coefficient is not strictly additive (i.e. $\alpha_P \neq \Sigma\alpha_i$). Indeed, if $\alpha_P = (1/V_P)(\partial V_P/\partial T)$, and we assume that the partial volume of the protein is properly described by the sum of the partial volumes of its amino acids (i.e. $V_P = \Sigma v_i$) (*166*), then:

$$\alpha_P = \frac{1}{\sum v_i}\left(\frac{\partial\left(\sum v_i\right)}{\partial T}\right) = \frac{1}{\sum v_i}\left(\frac{\partial v_1}{\partial T} + \frac{\partial v_2}{\partial T} + \ldots\right)$$

$$= \frac{1}{\sum v_i}\left(\frac{v_1}{v_1}\frac{\partial v_1}{\partial T} + \frac{v_2}{v_2}\frac{\partial v_2}{\partial T} + \ldots\right) = \frac{1}{\sum v_i}\left(v_1\alpha_1 + v_2\alpha_2 + \ldots\right)$$

$$\alpha_P = \frac{\sum v_i\alpha_i}{\sum v_i} \qquad \textbf{(7.2)}$$

Since it is known that the specific volumes of proteins also have temperature dependent behavior (*166*), an accurate representation of α_U(T) will also take the temperature dependence of $\bar{v}_p$ into account. Once again, we will utilize the relationship $\alpha_i = (1/V_i)(\partial V_i/\partial T)$ to derive a temperature dependent function for V_i ($V_i(T)$):

$$\alpha_i = \frac{1}{V_i}\frac{\partial V_i}{\partial T}$$

$$\int(\alpha_i)dT = \int\frac{dV_i}{V_i} \qquad \textbf{(7.3)}$$

Brandts and co-workers demonstrated that the $\alpha(T)$ profiles for individual amino acids could be represented by a cubic function ($\alpha(T) = a + bT + cT^2 + dT^3$) (*179*). It follows, then that:

$$\int(\alpha_i)dT = a_iT + \frac{b_i}{2}T^2 + \frac{c_i}{3}T^3 + \frac{d_i}{4}T^4 + \alpha_{i,o} \qquad \textbf{(7.4)}$$

where a_i, b_i, c_i, and d_i are the coefficients for a given amino acid, given in (*179*), and $\alpha_{i,o}$ is an arbitrary constant of integration. The other half of Eq. 7.3 is given by:

$$\int \frac{dV_i}{V_i} = \ln(V_i) + V_{i,o} \qquad \textbf{(7.5)}$$

By combining Eq. 7.3 and 7.4, we can solve for V_i as a function of temperature:

$$V_i(T) = \exp\left(a_i T + \frac{b_i}{2} T^2 + \frac{c_i}{3} T^3 + \frac{d_i}{4} T^4 + F_i \right) \qquad \textbf{(7.6)}$$

where $F_i = \alpha_{i,o} + V_{i,o}$. The partial molar volumes of the amino acids at 25°C were used to define F_i:

$$F_i = \ln\left(\overline{V}_i^{25^o C}\right) - a_i(25) - \frac{b_i}{2}(25)^2 - \frac{c_i}{3}(25)^3 - \frac{d_i}{4}(25)^4 \qquad \textbf{(7.7)}$$

By incorporating Eqs. 7.6 & 7.7 into Eq. 7.2, we were able to develop a model for the temperature dependence of $\alpha_U(T)$.

Figure 7.3 shows the $\alpha_U(T)$ profiles for the six proteins studied in this chapter, calculated based on the amino acid composition, as described above. From this figure, it can be seen that the calculated unfolded state baselines for all proteins are remarkably similar over the entire temperature range. There is only a slight deviation in the positions of the baselines at higher temperatures. Furthermore, these baselines fit well to a cubic polynomial function (solid lines, R^2 = 0.99). In order to determine the accuracy of the calculated unfolded state baselines, we compared the experimentally measured $\alpha(T)$ profile Ubq at pH 3.0 (Fig. 7.3, grey circles) to the baselines calculated from amino acid composition. From Fig. 7.3, it can be seen that the calculated unfolded state baselines at high temperatures are in good agreement with the unfolded state baseline of the Ubq $\alpha(T)$ profile. Future PPC experiments to characterize the $\alpha(T)$ profiles of natively unfolded proteins, such as apo-CytC or apomyoglobin, will be necessary to validate low temperature behavior of the calculated unfolded state baselines.

The shape of the native state baseline was determined using CNBr cleaved wild-type ubiquitin (Ubq-WT-CNBr) because it unfolds reversibly at pH 7.0, where the T_m is 87 °C. This makes it possible to observe the native state baseline at temperatures up to 65 °C. The native state baseline ($\alpha_N(T)$) for this experiment is shown in Fig. 7.3 (white circles). The $\alpha_N(T)$ profile also fits well to a cubic polynomial ($R^2 = 0.75$), suggesting that the overall shapes of the native and unfolded state baselines are similar. Therefore, we propose that both the native and unfolded state baselines should be represented by a 3rd order polynomial as a function of temperature (Fig. 7.2A):

$$\alpha_N(T) = \alpha_{N,ref} + B_N(T - T_m) + C_N(T - T_m)^2 + D_N(T - T_m)^3 \quad \textbf{(7.8)}$$

$$\alpha_U(T) = \alpha_{U,ref} + B_U(T - T_m) + C_U(T - T_m)^2 + D_U(T - T_m)^3 \quad \textbf{(7.9)}$$

where $\alpha_{N,ref}$ and $\alpha_{U,ref}$ are the values of $\alpha_N(T)$ and $\alpha_U(T)$, respectively, at the transition temperature, T_m. For most of the proteins studied here, we found that the simplest scenario, where $B_N = B_U$, $C_N = C_U$, and $D_N = D_U$ was sufficient to describe $\alpha_N(T)$ and $\alpha_U(T)$. However, in some instances, it was necessary to have different values for B_N, B_U, C_N, C_U, D_N, and D_U to more accurately describe the nature of the baselines.

7.2.2 Derivation of a two-state model for analysis of PPC data

Now that we have a good description for how to represent the native and unfolded state baselines in the analysis of a PPC experiment, we can derive the rest of the two-state model. A representation of $\alpha_{exp}(T)$ that unfolds via a two-state mechanism has been previously derived by Rösgen and Hinz (*264*) using temperature independent representations of α_N and α_U. The two-state analysis of their data provided measurements of $\Delta V/V$ that were within 10-20% of the model independent analysis performed by Brandts and co-workers (*179*). The temperature independent representations of α_N and α_U stem from the fact that Rösgen and Hinz were determining $\alpha(T)$ by

taking the temperature derivative of the specific volumes of proteins measured by densitometry. In the densitometry experiments, it was not possible to observe a significant temperature dependence of α below the T_m of the protein, so temperature independent values were used. However, the discussion of baselines provided in the previous section demonstrates that this is not quite an accurate representation of the behavior of $\alpha_N(T)$ and $\alpha_U(T)$, and could be the source of the small discrepancies between their results and those measured by Brands and co-workers (*179*). The following is a derivation of $\alpha_{exp}(T)$ which includes a description for the temperature dependence of the baselines. The thermodynamic parameters of ΔH and T_m are obtained by performing parallel DSC experiments on each protein, and are used to fit for the volumetric changes upon unfolding ($\Delta V/V$).

The relationship between α and ΔV has been previously described (*179*) and is equal to:

$$\alpha_p = \frac{1}{V_P}\frac{\partial V_P}{\partial T} \qquad \textbf{(7.10)}$$

where V_P is the partial volume of the protein. Therefore, the change in volume can be obtained by calculating the area under the experimental $\alpha(T)$ curve ($(\alpha_{exp}(T)$, Fig. 7.2A), which is made up of two components:

$$\alpha_P^{\exp}(T) = \alpha_P^{progress}(T) + \alpha_P^{excess}(T) \qquad \textbf{(7.11)}$$

where $\alpha_P^{progress}(T)$ (Fig. 7.2B) is defined by the fraction of native (F_N) and unfolded (F_U) protein in the sample:

$$\alpha_P^{progress}(T) = F_N \cdot \alpha_N(T) + F_U \cdot \alpha_U(T) \qquad \textbf{(7.12)}$$

where $\alpha_N(T)$ and $\alpha_U(T)$ are defined by equations 7.8 and 7.9, respectively. Subtracting $\alpha_P^{progress}(T)$ from $\alpha_P^{\exp}(T)$ gives the $\alpha_P^{excess}(T)$ profile (Fig. 7.1C). The area under this curve is equal to $\Delta V/V$, and we will now derive the relationship between $\alpha_P^{excess}(T)$ and $\Delta V/V$ for a protein that undergoes two-state unfolding.

For a two-state system, the Gibbs free energy (ΔG) of unfolding is equal to:

$$\Delta G(T) = -RT\ln(K_{eq}) = -RT\ln\left(\frac{F_U}{F_N}\right) \quad \textbf{(7.13)}$$

where R is the universal gas constant and K_{eq} is the unfolding equilibrium constant, F_N is the fraction of folded protein and F_U is the fraction of unfolded protein in the population. The Gibbs free energy can also be related to the changes in enthalpy (ΔH) and entropy (ΔS) upon unfolding via the Gibbs-Helmholtz relationship:

$$\Delta G(T) = \Delta H(T) - T\Delta S(T) \quad \textbf{(7.14)}$$

where $\Delta H(T) = \Delta H(T_m) + \Delta C_P \cdot (T-T_m)$ and $\Delta S(T) = \Delta H(T_m)/T_m + \Delta C_P \cdot \ln(T/T_m)$. Combining Eqs. 7.13 and 7.14 and solving for K_{eq} yields:

$$K_{eq} = \exp\left(\frac{-\Delta G}{RT}\right) = \exp\left(\frac{-\Delta H}{RT}\right)\exp\left(\frac{\Delta S}{R}\right) \quad \textbf{(7.15)}$$

Knowing that $K_{eq} = (F_U/F_N)$ and that $F_U + F_N = 1$, we can solve for F_U:

$$F_U = \frac{\exp\left(\frac{-\Delta G}{RT}\right)}{1+\exp\left(\frac{-\Delta G}{RT}\right)} = \frac{\exp\left(\frac{-\Delta H}{RT}\right)\exp\left(\frac{\Delta S}{R}\right)}{1+\exp\left(\frac{-\Delta H}{RT}\right)\exp\left(\frac{\Delta S}{R}\right)} \quad \textbf{(7.16)}$$

If we assume that ∂V_P is temperature independent, then $\alpha_P^{excess}(T)$ is related to the fraction of unfolded protein in the sample (Fig. 7.2C), in such a way that Eq. 7.10 becomes:

$$\alpha_P^{excess}(T) = \frac{1}{V_P}\frac{\partial}{\partial T}(F_U \cdot \partial V_P) = \frac{\Delta V_P}{V_P}\frac{\partial}{\partial T}(F_U) \quad \textbf{(7.17)}$$

To obtain the complete relationship between $\alpha_P^{excess}(T)$ and ΔV_P, we take the derivative of Eq. 7.17 with respect to temperature:

$$\frac{\partial}{\partial T}\left(F_U\right)=\frac{\partial}{\partial T}\left(\frac{\exp\left(\frac{-\Delta H}{RT}\right)\exp\left(\frac{\Delta S}{R}\right)}{1+\exp\left(\frac{-\Delta H}{RT}\right)\exp\left(\frac{\Delta S}{R}\right)}\right)$$
$$=\frac{\exp\left(\frac{\Delta S}{R}\right)\exp\left(\frac{-\Delta H}{RT}\right)\frac{\Delta H}{RT^2}}{\left(1+\exp\left(\frac{\Delta S}{R}\right)\exp\left(\frac{-\Delta H}{RT}\right)\right)^2} \qquad \textbf{(7.18)}$$

$$\frac{\partial}{\partial T}\left(F_U\right)=\frac{\exp\left(\frac{-\Delta G}{RT}\right)\cdot\frac{\Delta H}{RT^2}}{\left(1+\exp\left(\frac{-\Delta G}{RT}\right)\right)^2}=\frac{K_{eq}}{\left(1+K_{eq}\right)^2}\frac{\Delta H}{RT^2} \qquad \textbf{(7.19)}$$

Substituting Eq. 7.19 into Eq. 7.17 gives the relationship between $\alpha_P^{excess}(T)$ and ΔV_P:

$$\alpha_P^{excess}(T)=\frac{K_{eq}}{\left(1+K_{eq}\right)^2}\frac{\Delta H}{RT^2}\frac{\Delta V_P}{V_P} \qquad \textbf{(7.20)}$$

Finally, by incorporating Eqs. 7.12 and 7.20 into equation 7.11, we obtain the following representation for $\alpha_P^{\exp}(T)$, which fits the PPC data to a two-state model of unfolding (Fig. 7.2D):

$$\alpha_P^{\exp}(T)=F_N\cdot\alpha_N(T)+F_U\cdot\alpha_U(T)+\frac{K_{eq}}{\left(1+K_{eq}\right)^2}\frac{\Delta H}{RT^2}\frac{\Delta V_P}{V_P} \qquad \textbf{(7.21)}$$

where $\alpha_N(T)$ and $\alpha_U(T)$ are in the form given by Eqs. 7.8 and 7.9, respectively, and $\Delta H(T) = \Delta H(T_m) + \Delta C_P\cdot(T-T_m)$. From Eq. 7.21, we can see that placing PPC data in the context of a two-state model of unfolding makes it possible to directly fit the data, not only for $\Delta V/V$, but also for ΔH, ΔC_P, and T_m. This provides the potential to get a full thermodynamic description of protein unfolding in with a single PPC experiment.

7.3 Results & Discussion

7.3.1 The two-state model is a robust method for analyzing PPC data

One of the first aspects of the two-state model that we tested was to see if it was really necessary to define the ΔH and T_m values based on the DSC data, or if we could simultaneously fit PPC data for $\Delta V/V$ and these thermodynamic parameters. Indeed, it is possible to fit the PPC data with both ΔH and T_m as parameters. This will typically result in a fitted PPC T_m that is within 2-3 °C of the DSC T_m. However, the fitted ΔH values based solely on PPC data can differ from the DSC measurements by as much as 15%. This occurs mostly because fewer experimental points are collected in a PPC experiment, where data points are collected every 5 °C outside the transition range and every 2 °C in the transition range, than in a DSC experiment, where data points are collected every 0.1 °C. As a result, the temperature dependence of the equilibrium constant, and consequently the enthalpy of unfolding, cannot be defined with the same accuracy in PPC analysis as in DSC analysis. The combinations of these errors could result in erroneous estimates of the volumetric changes ($\Delta V/V$) upon unfolding (Table 7.1), which might then result in an incorrect interpretation of experimental data. This was especially evident for proteins like CytC (Fig. 7.1B), where $\Delta V/V$ is smaller than $\Delta V/V$ of ubiquitin (Fig. 7.1A). For these reasons, it is recommended to always perform a corresponding DSC experiment so the T_m and ΔH parameters can be properly constrained in the PPC data analysis.

The next aspect of our model that we tested was whether different protein concentrations would yield significantly different measurements of $\Delta V/V$. Figure 7.4 shows the data for PPC experiments performed on ribonuclease A (RNaseA) at four different concentrations. This figure clearly demonstrates that in the concentration range of 1-4 mg/mL, protein concentration has very little effect on the experimental results. It has been previously argued that increasing

concentration will change the baselines (and potentially the measurement of $\Delta V/V$) because increased intermolecular interactions will affect the hydration properties of the native state (*266*). However, these effects were only observed for protein concentrations above 5 mg/mL. It seems that as long as the protein concentration is less than 5 mg/mL, there should be very little dependence of $\alpha_{exp}(T)$ or $\Delta V/V$ on protein concentration. This is an important observation because it means that it is possible to perform accurate PPC experiments with lower protein concentrations than previously been used or recommended.

It is also necessary to test whether our two-state model for analyzing PPC data can be used on a number of different proteins that are known to exhibit two-state unfolding. Figure 7.5 shows the results of PPC experiments performed on five proteins with different sizes, shapes and secondary structural compositions: hen egg-white lysozyme (HEWL Fig. 7.5A), ribonuclease A (RNaseA, Fig. 7.5B), ubiquitin (Ubq, Fig. 7.5C), cytochrome *c* (CytC, Fig. 7.5D), and eglin C (EgC, Fig 7.5E). The temperature dependence of the thermal expansion coefficient was measured at three to five different pH values for each protein. It was possible to fit all five proteins to our two-state model of unfolding, even when the volumetric changes switched signs (CytC), demonstrating that this is a robust method for analyzing PPC data.

7.3.2 High temperature convergence of $\Delta V/V$

The pH-dependent $\alpha_{exp}(T)$ profiles also made it possible to analyze the volumetric changes of different proteins as a function of transition temperature. Figure 7.6 shows a plot of $\Delta V/V$ vs. T_m for the six proteins mentioned above. HEWL, RNaseA, Ubq, and EgC all exhibit similar behaviors, in the sense that $\Delta V/V$ is always negative and varies linearly with T_m. In addition, the volumetric changes of these proteins appear to converge at high temperatures. Interestingly, the temperature of convergence appears to be similar for ΔS (~ 110 °C), ΔH (~ 130 °C), and ΔV (120 - 130 °C). The causes of the convergence at high temperatures of the entropies

and enthalpies of unfolding, normalized by the size of the protein, has previously been observed and discussed (*185, 246, 254, 268-270*). Based on model compound studies, it appears that the entropies of unfolding measured for several globular proteins converge at the temperature where the entropy associated with transferring a nonpolar compound to water is zero (*246, 268*). In other words, at the convergence temperature, the only contribution to the entropic component is the configurational entropy. The convergence temperature of the specific enthalpies of unfolding have also been argued to be due to the hydrophobic effect. In other words, the nonpolar contribution to the enthalpy of unfolding is equal to zero (*268-270*). Since the enthalpies of van der Waals interactions, polar hydration, and nonpolar hydration change differently for different proteins as a function of temperature (*254*), at the convergence temperature, these terms must compensate in such a way that the only contribution to the specific enthalpy of unfolding is hydrogen bonding (*185, 254*). The observed convergence of $\Delta V/V$ for HEWL, RNaseA, Ubq, and EgC suggests that high-temperature convergence of this parameter might also be a general behavior of naturally occurring proteins.

To explore the cause of the high temperature convergence of $\Delta V/V$, we need to return to the parameters that define the specific volume of a protein, as described in Eq. 7.1. If the partial molar volume is defined as described by Chalikian and Breslauer (*261*): $\overline{V}_p = V_{\text{int}} + \Delta V_{hyd} + V_t$, then the change in volume upon unfolding ($\Delta V/V$) can be described as:

$$\Delta V / V = \Delta V_{\text{int}} / V + \Delta\Delta V_{hyd} / V + \Delta V_t / V \qquad \textbf{(7.22)}$$

As mentioned in the introduction V_{int} describes the void volume due to imperfect packing of the native state of the protein. The native state will provide a positive (i.e. $V_{int,N} > 0$) contribution to the specific volume of the protein because the internal cavities are not solvent accessible. Upon unfolding, there are no solvent inaccessible voids, so $V_{int,U} \approx 0$. Therefore, $\Delta V_{int}/V$ will have a slightly negative (i.e. $\Delta V_{int}/V < 0$) contribution to $\Delta V/V$ at lower temperatures. Because the

internal voids will also be present in the native state of proteins at high temperatures, the contributions of $\Delta V_{int}/V$ should also be negative under these conditions. The hydrational term, ΔV_{hyd}, is defined by the interactions of the polar and nonpolar groups of the protein with the solvent. This effect is likely to be very large at room temperature with a magnitude that probably depends on the amount of polar and nonpolar surface area in the protein. However, the expansivity data on single amino acids (*179*) shows that the high temperature values of α for nonpolar and polar amino acids converge at high temperature to a value that is similar to that of bulk water. Therefore, at high temperatures, the contribution of $\Delta\Delta V_{hyd}/V$ to $\Delta V/V$ is expected to be negligible. The thermal volume, V_t, is defined by the thermal fluctuations of the side chains, which effectively pushes the solvent away from the protein molecule, resulting in an increase in the partial specific volume of the protein. Because the unfolded protein molecule is larger than the native state, $V_{t,U}$ is expected to be larger than $V_{t,N}$. As such, $\Delta V_t/V$ is expected to have a large positive contribution to $\Delta V/V$. This contribution is expected to be relatively independent of temperature.

Based on this discussion of the relative contributions of different components of the partial volume of proteins to the volumetric changes upon unfolding, the contribution of the hydrational term is essentially negligible at the high temperatures (~120-130 °C) where convergence of $\Delta V/V$ is observed. Therefore, at these temperatures, the only terms contributing to $\Delta V/V$ are the intrinsic volume and the thermal volume:

$$\Delta V / V = \Delta V_{\text{int}} / V + \Delta V_t / V \qquad \textbf{(7.23)}$$

Since $\Delta V_{int}/V$ is defined by the volumes of internal cavities, it essentially describes the packing density of a protein. Furthermore, the packing densities of most proteins are very similar, so if ΔV_{int} is normalized per residue (as $\Delta V_{int}/V$), then one would also expect $\Delta V_{int}/V$ to be similar for all proteins. The normalized contribution of the thermal volume (as $\Delta V_t/V$) should also be similar for all proteins since the difference in size between the native and unfolded states should be

roughly similar for all proteins. Based on the relative contributions of the intrinsic and thermal volumes to $\Delta V/V$, one would expect the volumetric changes upon unfolding to converge at a small, positive value, and indeed, this is what is observed (Fig. 7.6, $\Delta V/V_{conv} \approx 0.002 - 0.004$).

The argument that high-temperature convergence of $\Delta V/V$ is a general property of all proteins is challenged by our measurements of CytC. The linear fit of $\Delta V/V(T_m)$ for CytC has a steeper temperature dependence than the other proteins, such that eventually, $\Delta V/V$ changes sign. One possible explanation for the anomalous behavior of CytC can be found in its structural characteristics. Unlike any of the other proteins, CytC has a covalently bound heme group, which adds a large, rigid, hydrophobic structure to the globular protein. In addition to perturbing the packing of the protein around the heme group, this rigid structure will have limited expansivity relative to the protein itself, and may therefore affect the volumetric changes we are observing. In fact, the total volume of internal cavities of CytC is more than an order of magnitude larger than the internal cavity volume of the other proteins (Table 7.2), which would suggest that the presence the heme group is perturbing the packing of this protein compared to others. If the internal cavities of the protein are indeed the primary determinants of the magnitude of the volumetric changes upon unfolding, then it is possible that structures with larger cavity volumes will display different $\Delta V/V$ behavior than proteins that have smaller internal cavities. More studies on proteins with large numbers or volumes of internal cavities (lower packing densities) will be needed to confirm this result.

7.3.3 Cold denaturation can be studied by PPC

One of the advantages of using a two-state model is that it becomes possible to interpret the data from a PPC experiment where the current methods fail due to an inability to define appropriate baselines. When a protein undergoes cold denaturation, the native state baseline can no longer be defined as passing through the data points since the fraction of native protein never

reaches 100%, even at low temperatures. As a result, it will be extremely difficult analyze such an experiment using a model independent approach. In order to explore whether our model would be able to handle cases like cold denaturation, we studied the unfolding of supercharged ubiquitin (Ubq-SC), which undergoes cold denaturation in acidic pH conditions (see Chapter 5).

Before starting the PPC experiment, we used our two-state model to simulate what cold denaturation might look like, and if it would provide a detectable signal in the calorimeter. The simulated curve (thin line, Fig. 7.7) was calculated based on the following assumptions: **1.**) The native and unfolded state baselines of Ubq-SC would have a similar behavior to those of Ubq-WT; **2.**) $\Delta V/V$ for the simulated curve was taken by extrapolating the $\Delta V/V$ vs. T_m line for Ubq-WT to the T_m of Ubq-SC; and **3.**) the ΔH and T_m of Ubq-SC were taken from the corresponding DSC experiment. When we saw that we should still get a good signal-to-noise ratio in our experiments, we measured the unfolding of Ubq-SC at pH 3.5 by PPC. From Figure 7.7, it can be seen that the transition region of the protein can be successfully predicted from our simulated curve. The difference between the predicted and experimental baselines is most likely explained by the large number of mutations between Ubq and Ubq-SC. It has previously been demonstrated that the polar/apolar nature of surface residues is extremely important for defining position and temperature dependence of the baselines of $\alpha_{exp}(T)$ (*179, 263, 266*). Nevertheless, the two-state model can successfully fit the cold denaturation data (thick line Fig 7.8), demonstrating that PPC can be used to explore the volumetric changes of proteins over a broad range of thermodynamic stabilities.

To put the concept of cold denaturation, as measured by PPC, into a more familiar context, Figure 7.9 shows simulated $\alpha_{exp}(T)$ (Fig. 7.8A), $F_U(T)$ (Fig. 7.8B), and $C_P(T)$ (Fig. 7.8C) curves for a protein with different stabilities. The parameters used to calculate these curves are given in Table 7.3. The signature of cold denaturation in a DSC experiment is the observation that the $C_P(T)$ profile initially decreases, and then increases again with increasing temperature.

This is further manifested in the $F_U(T)$ curves, where the fraction of unfolded protein never reaches zero. In a PPC experiment, cold denaturation seems to manifest itself as an increase in the absolute position of $\alpha(T)$ at the starting temperature and a large increase in the apparent slope of the native state baseline. The unfolded state baselines for both α_{exp} and C_P curves are independent of whether a protein undergoes cold denaturation, as expected. Furthermore, as with DSC, the position of the peak maximum does not necessarily correlate to the T_m of the protein.

7.4 Implications of Two-State Model for Future PPC Experiments

This chapter has discussed a novel way of analyzing the data from a PPC experiment. The main advantage of using a two-state model of unfolding to calculate the volumetric changes upon unfolding is that placing the results in this context allows one to understand volumetric changes when baselines are not easily defined. The possible existence of the convergence of the volumetric changes at high temperatures was discussed. CytC challenges this notion because its volumetric changes as a function of temperature do not converge with the other proteins. It is possible that this is due to the decreased packing density as a result of the presence of a heme group, which will increase the amount of internal void volume, relative to the other proteins. In order to test this hypothesis, we can examine the volumetric changes of other proteins with low packing densities. In particular, we have already characterized the thermal denaturation of several ubiquitin and eglinC variants which have cavity creating substitutions in the protein core (*8, 17, 163*), and should therefore have lower packing densities than wild-type ubiquitn and eglinC, respectively. By characterizing the pressure-induced denaturation of these variants under similar conditions as described here, it will be possible to determine whether the decreased packing density of CytC, relative to the other proteins, is the source of its anomalous behavior.

In the future, it will also be interesting to extend the use of PPC to examine the role of solvent in protein unfolding. Since $\alpha(T)$ seems to be sensitive to solvent conditions, especially at low T, PPC provides a sensitive method for determining the extent to which substitutions on the protein surface affect interactions with bulk solvent. In our computational design approach, we typically only make a small number of substitutions, so $\alpha(T)$ would not expected to be significantly different between the wild-type and designed variants. However, PPC could prove to be particularly useful in the context of supercharging proteins to increase solubility without sacrificing stability (see Chapter 6). It is possible that choosing substitutions such that we strike a balance between a high net charge, but an $\alpha(T)$ profile that is similar to the wild-type protein, indicating that the solvent/protein interactions have not been significantly altered, then we will be able to develop a better understanding of how intramolecular interactions contribute differently to stability and solubility.

Table 7.1 Comparison of fitted parameters for ubiquitin PPC data

pH	T_m (°C) DSC	T_m (°C) PPC fit	$\Delta H(T_m)$ (kJ/mol) DSC	$\Delta H(T_m)$ (kJ/mol) PPC fit	$\Delta V/V$ DSC ΔH & T_m	$\Delta V/V$ Fitted ΔH & T_m	$\Delta V/V$ % difference
2.4	61.5	60.2	231	245	-9.85×10^{-3}	-9.06×10^{-3}	-8.73%
2.6	63.1	61.8	244	240	-9.49×10^{-3}	-1.01×10^{-2}	5.84%
2.8	66.2	64.5	261	247	-7.88×10^{-3}	-9.31×10^{-3}	15.36%
3.0	71.1	70.3	266	252	-8.07E $\times 10^{-3}$	-8.87×10^{-3}	8.97%
3.2	73.2	72.1	277	239	-8.07E $\times 10^{-3}$	-1.08×10^{-2}	25.29%

Table 7.2 Total cavity volumes (V_{cav}) of five model proteins studied by PPC

Protein Name	**V_{cav} ($Å^3$)** [a]
Lysozyme	1.16
RNaseA	2.92
Ubiquitin	21.75
EglinC	16.10
Cytochrome *C*	159.70

[a] The cavity volumes were calculated with the VOIDOO software package (*271, 272*) using a probe size of 1.4 Å.

Table 7.3 Parameters used to simulate α(T), F_U(T), and C_P(T)

Curve ID[a]	T_m (°C)	$\Delta H(T_m)$ (kJ/mol)	$C_{P,N,ref}$[b] (kJ/mol K)	$\alpha_{N,ref}$[c] (deg^{-1})	$\alpha_{U,ref}$[c] (deg^{-1})	$\Delta V/V$[d]
1 (solid)	19	0	11.9	$6.9x10^{-4}$	$1.0x10^{-3}$	-0.0177
2 (long dash)	49.5	114	11.9	$6.9x10^{-4}$	$1.0x10^{-3}$	-0.0126
3 (medium dash)	61.5	231	11.9	$6.9x10^{-4}$	$1.0x10^{-3}$	-0.00929
4 (short dash)	63.1	244	11.9	$6.9x10^{-4}$	$1.0x10^{-3}$	-0.00921
5 (dotted)	66.2	261	11.9	$6.9x10^{-4}$	$1.0x10^{-3}$	-0.00825
6 (dash-dot)	71.1	266	11.9	$6.9x10^{-4}$	$1.0x10^{-3}$	-0.00804
7 (dash-dot-dot)	73.2	277	11.9	$6.9x10^{-4}$	$1.0x10^{-3}$	-0.00739

(a) The parenthesis correspond to the line patterns in Fig. 7.8.
(b) ΔC_P for all curves was 3.46 kJ/mol K.
(c) The baselines for $\alpha_{exp}(T)$ take the form given in Eq. 7.1 and 7.2, where $B_N = B_U = -1.5x10^{-5}$, $C_N = C_U = 7.1x10^{-8}$, and $D_N = D_U = 0$ for all curves.
(d) The $\Delta V/V$ values for curves 1 and 2 are calculated based on extrapolation of the ubiquitin $\Delta V/V(T_m)$ curve shown in Fig. 7.6

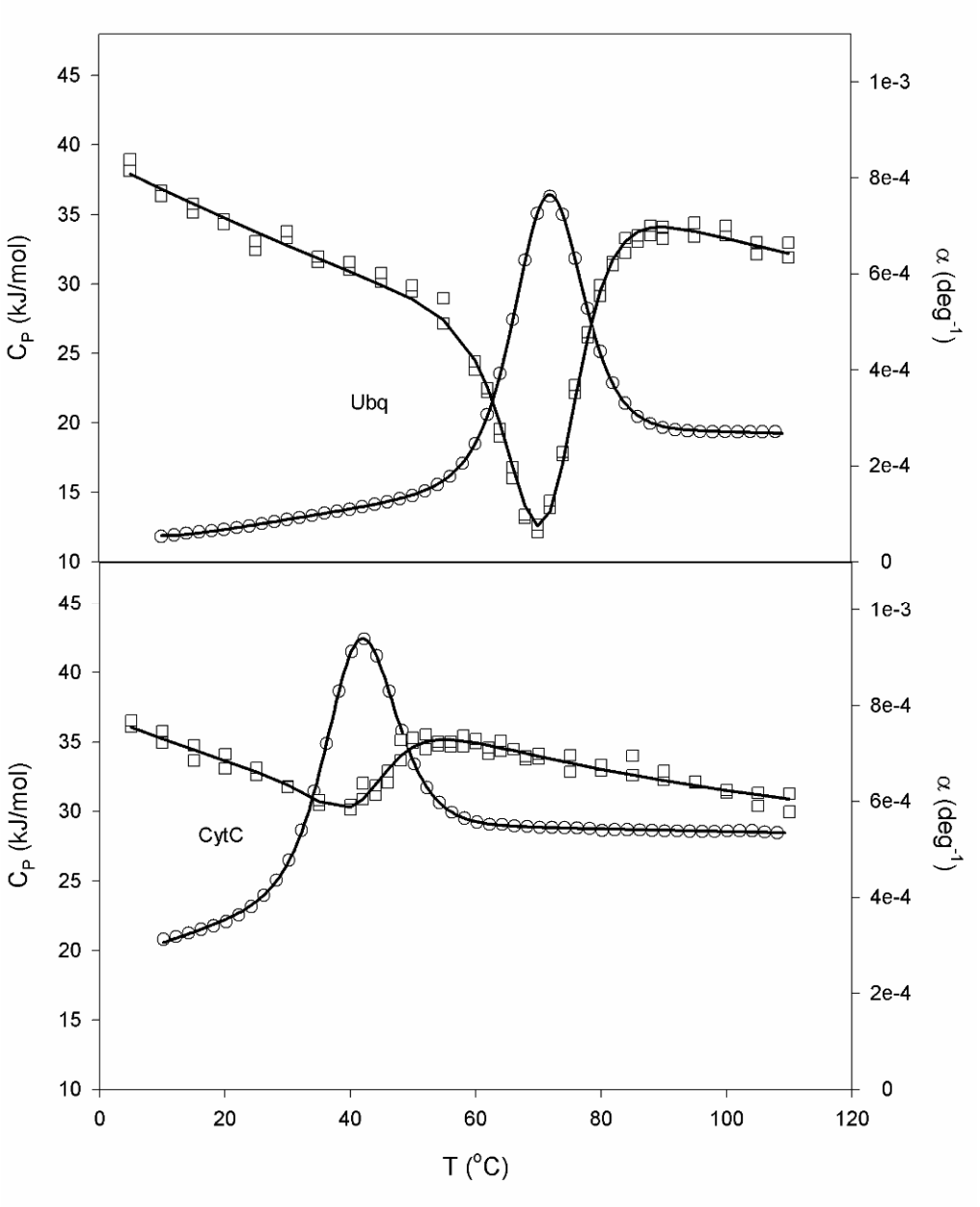

Figure 7.1 Comparison of thermal denaturation curves obtained from DSC and PPC experiments for **A.** Ubiquitin in 50mM Glycine buffer, pH 3.0 and **B.** Cytochrome C in 50mM Glycine buffer, pH 3.0. In both panels, the symbols represent the experimental data (○ – DSC data, shown every 5 ºC for clarity; □ – PPC data). The solid lines represent the fit of the experimental data to a two-state model of unfolding.

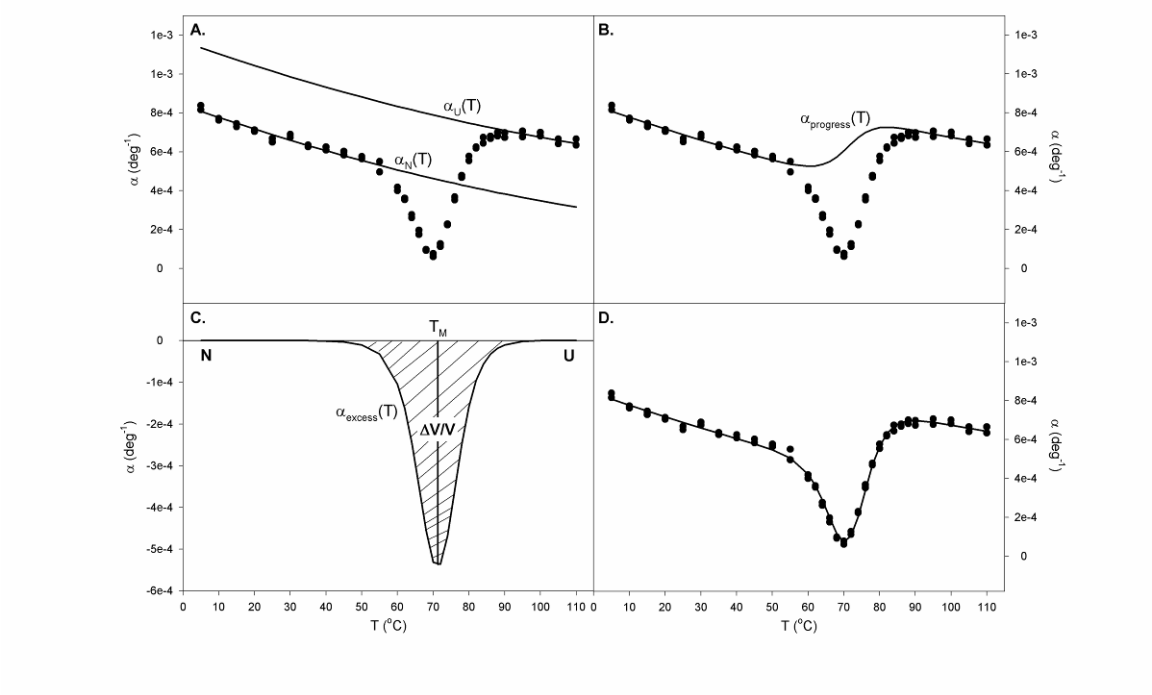

Figure 7.2 Example of PPC data fit to a two-state model of unfolding. In all parts of the figure, the symbols represent the experimental data. **A.** Native ($\alpha_N(T)$) and unfolded ($\alpha_U(T)$) baselines. **B.** Definition of the $\alpha_{progress}(T)$. **C.** Relationship between the $\alpha_{excess}(T)$ and $\Delta V/V$. **D.** The fit of the experimental data to the two-state model of unfolding described here.

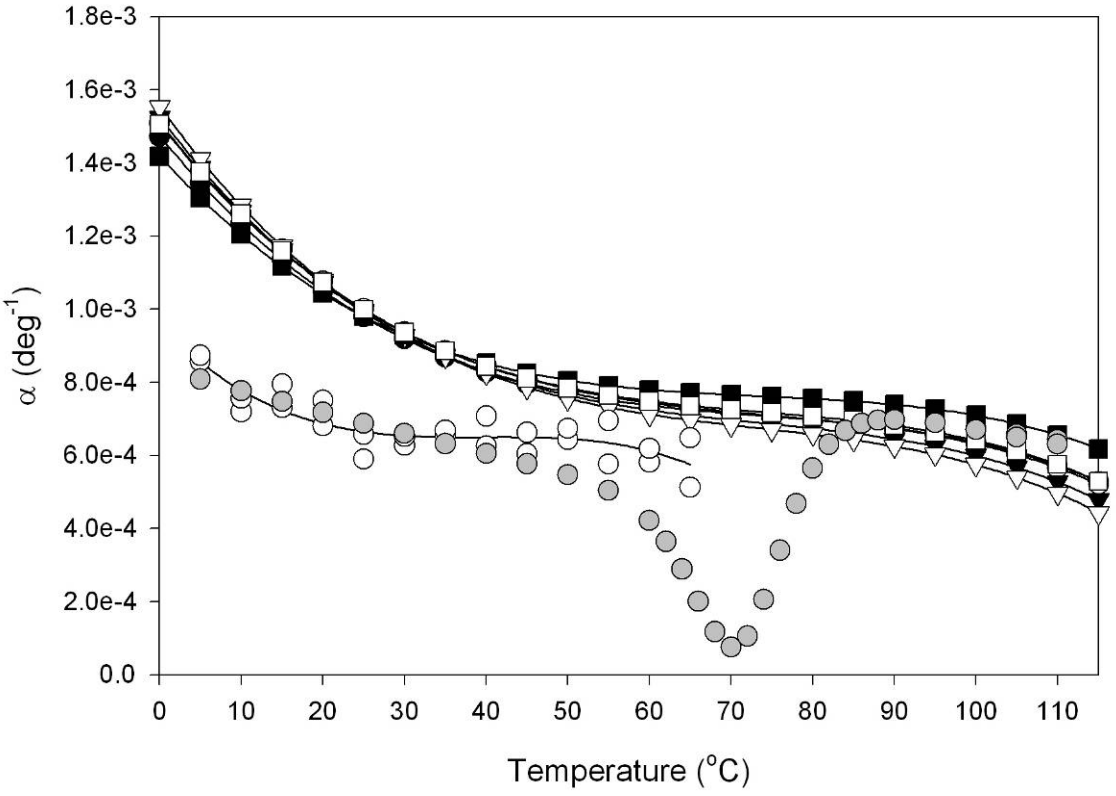

Figure 7.3 Temperature-dependent behaviors of native and unfolded state baselines for PPC experiments. The unfolded state baselines, based on amino acid composition, were determined from the temperature-dependent α profiles for individual amino acids studied in (*179*). The symbols represent the baselines for different proteins: ● – RNaseA, ○ – Ubq, ▼ – Lysozyme, ▽ – CytC, ■ – Ubq-SC, □ EglinC. The solid lines represent the fit of the data to a cubic function. For comparison the native and unfolded state baselines were also experimentally measured for Ubq-WT-CNBr (○) and Ubq at pH 3.0 (●).

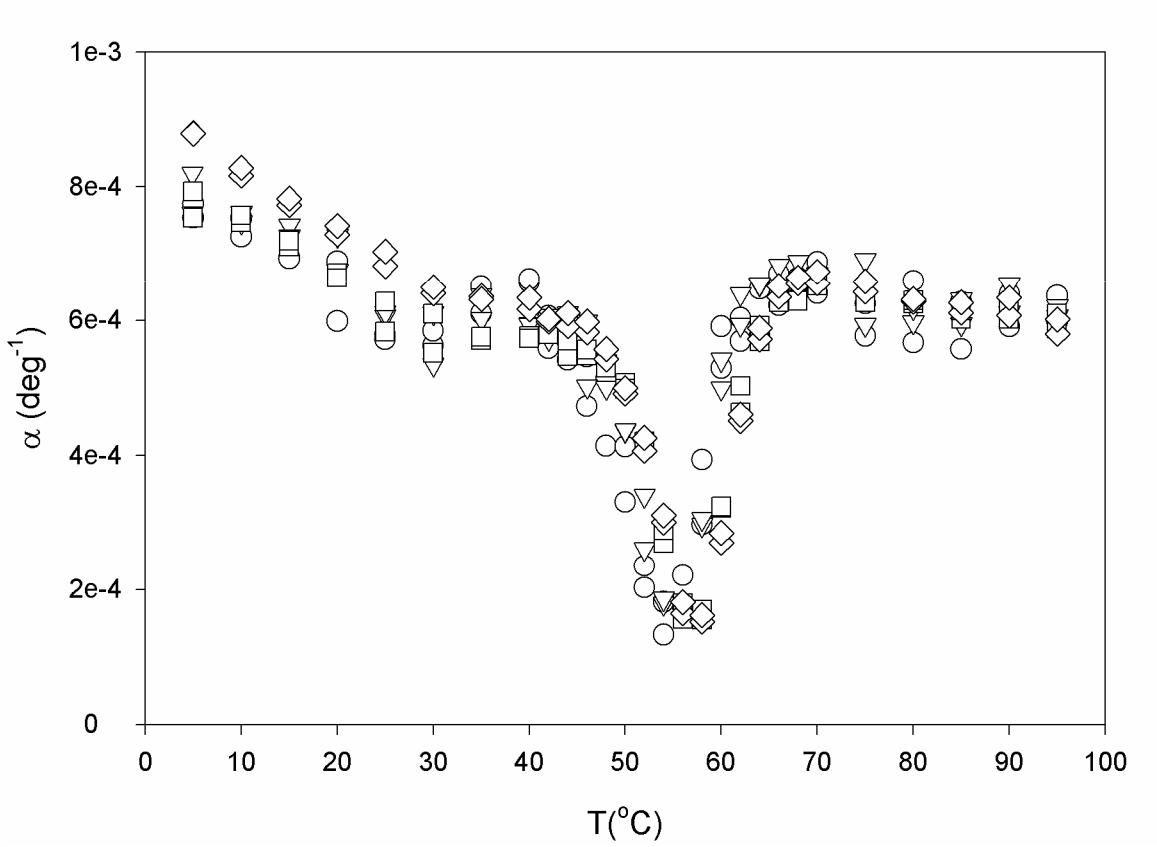

Figure 7.4 Concentration dependence of RNaseA on PPC data. All experiments were performed in 10mM glycine buffer, pH3.2. ○ – 1.249 mg/mL ($\Delta V/V = 4.8 \times 10^{-3}$), □ – 1.642 mg/mL ($\Delta V/V = 4.9 \times 10^{-3}$), ◊ – 3.962 mg/mL ($\Delta V/V = 4.6 \times 10^{-3}$), ▽ – 4.781 mg/mL ($\Delta V/V = 4.4 \times 10^{-3}$). The experimental data for all protein concentrations overlay, indicating that in the concentration range of 1-5mg/mL, PPC results are independent of protein concentration. Furthermore, the fitted volumetric changes are the same within 10%, which is within the accepted error of previous experiments (*179, 264*).

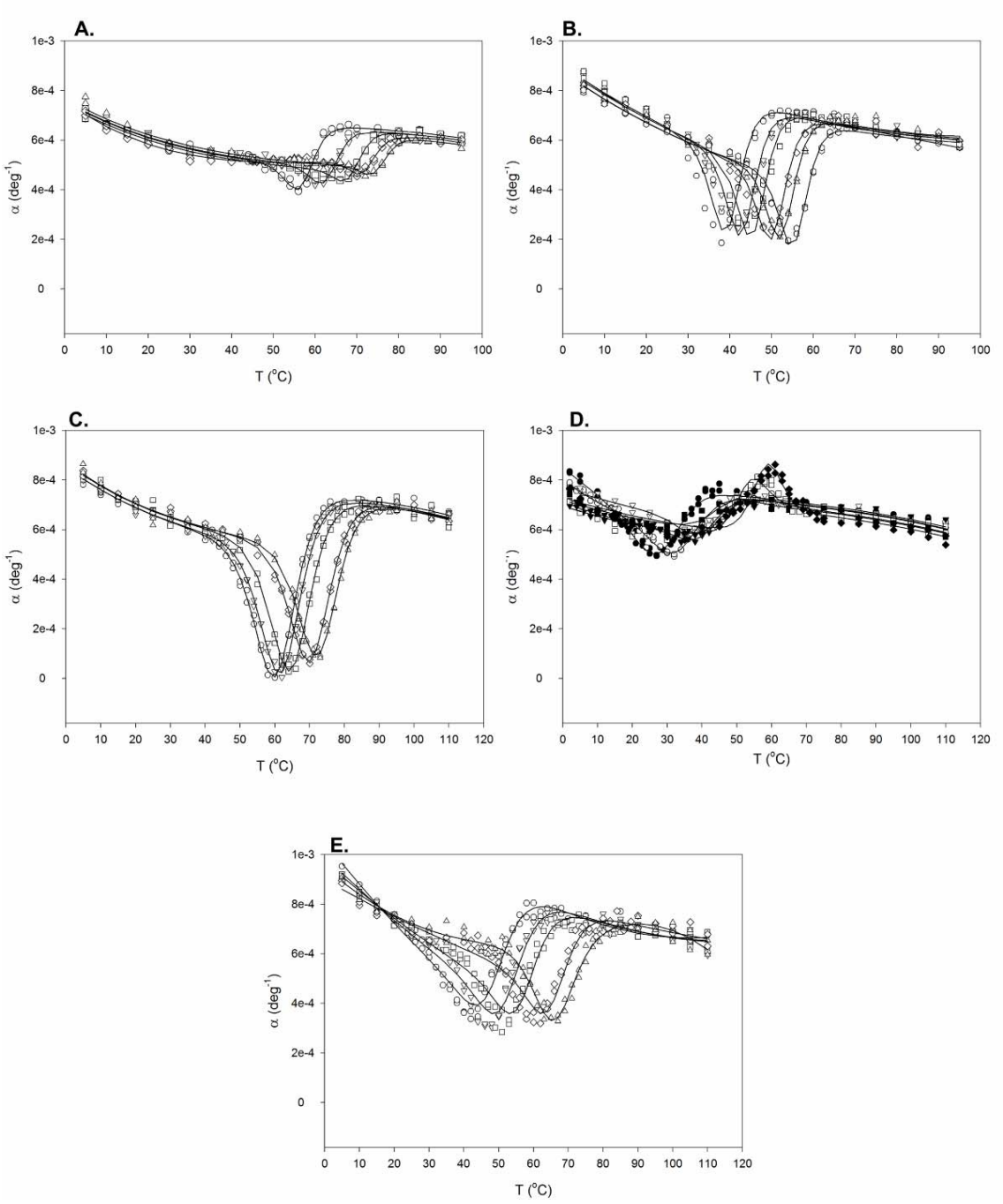

Figure 7.5 pH-dependence of $\alpha_{exp}(T)$ for five model proteins. The solid lines in each part of the figure represent the fits of the data to a two-state model of unfolding. The symbols represent the experimental data for: **A.** Lysozyme – pH 2.2 (○), pH 2.5 (▽), pH2.8 (□), pH 3.1 (◊), pH 3.4 (△). **B.** RNaseA – pH 2.4 (○), pH 2.6 (▽), pH 2.8 (□), pH 3.0 (◊), pH 3.2 (△) , pH 3.45 (⬡). **C.** Ubiquitin – pH 2.4 (○), pH 2.6 (▽), pH 2.8 (□), pH 3.0 (◊), pH 3.2 (△). **D.** Cytochrome *c* – pH 2.4 (●), pH 2.6 (○), pH 2.8 (▼), pH 3.0 (▽), pH 3.2 (■), pH 3.4 (□), pH 3.6 (♦). **E.** EglinC – pH 2.5 (○), pH 2.75 (▽), pH 3.0 (□), pH 3.25 (◊), pH 3.5 (△).

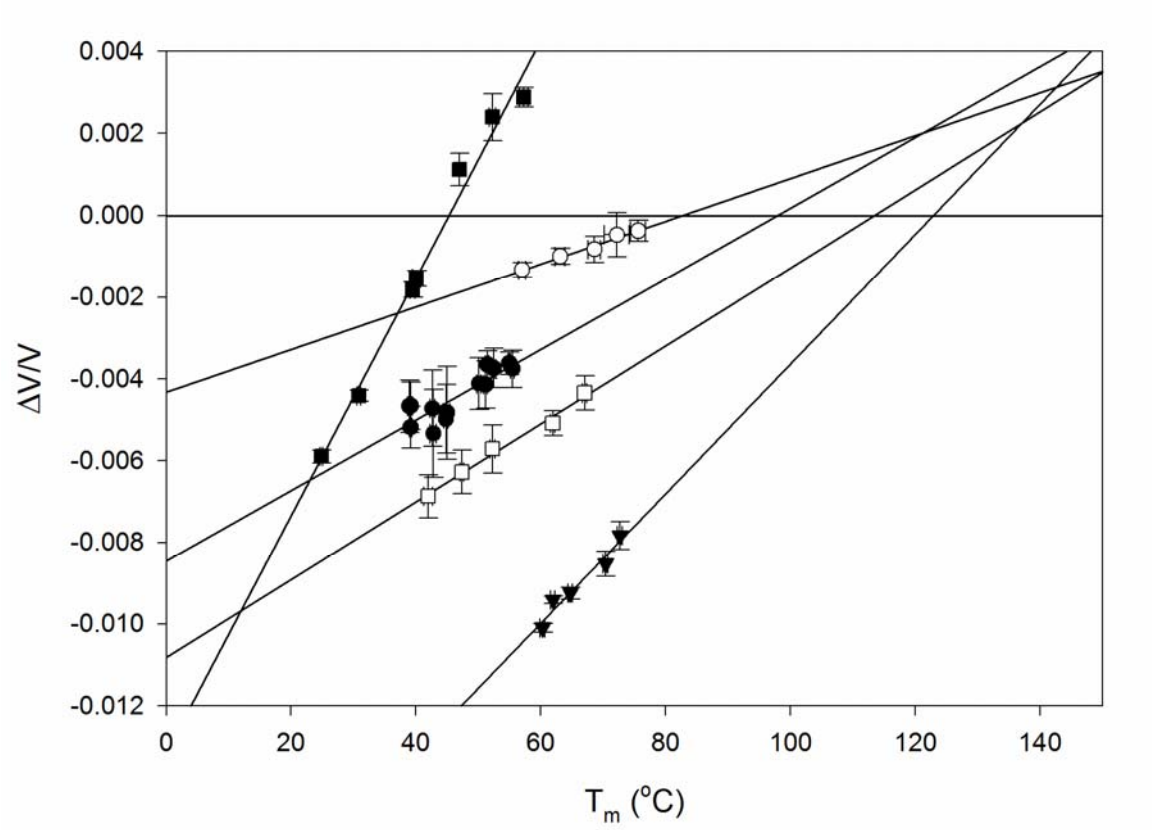

Figure 7.6 Temperature dependence of *ΔV/V*. RNaseA (●), HEWL (○), Ubq (▼), CytC (■), and EgC (□). The symbols represent the average of fitted *ΔV/V* using several different representations of the unfolded state. The error bars are the standard deviation of the averaged *ΔV/V* values at each point. The solid lines represent the linear regressions of the data. For all proteins, except CytC, the *ΔV/V* functions appear to converge in the temperature range of 110-130 °C.

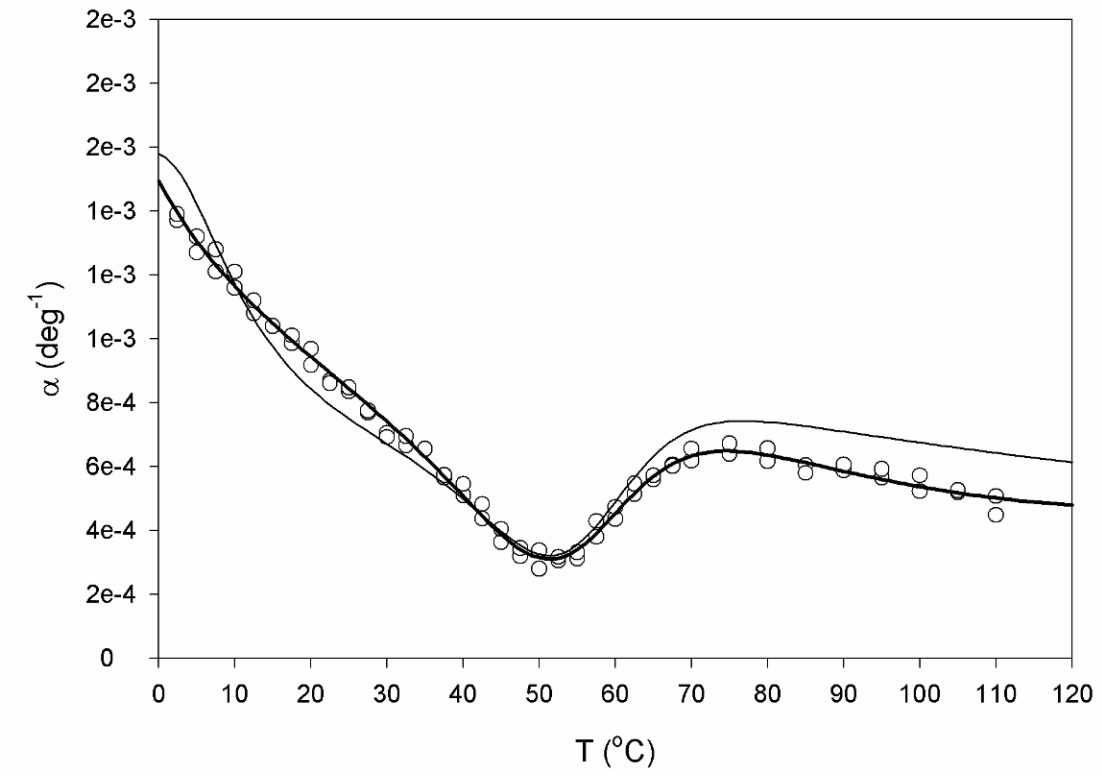

Figure 7.7 Cold denaturation of Ubq-SC measured by PPC. The symbols represent the experimental data. The thin line is the simulated cold denaturation curve, which was simulated using the parameters in Table 7.3, Curve 1.

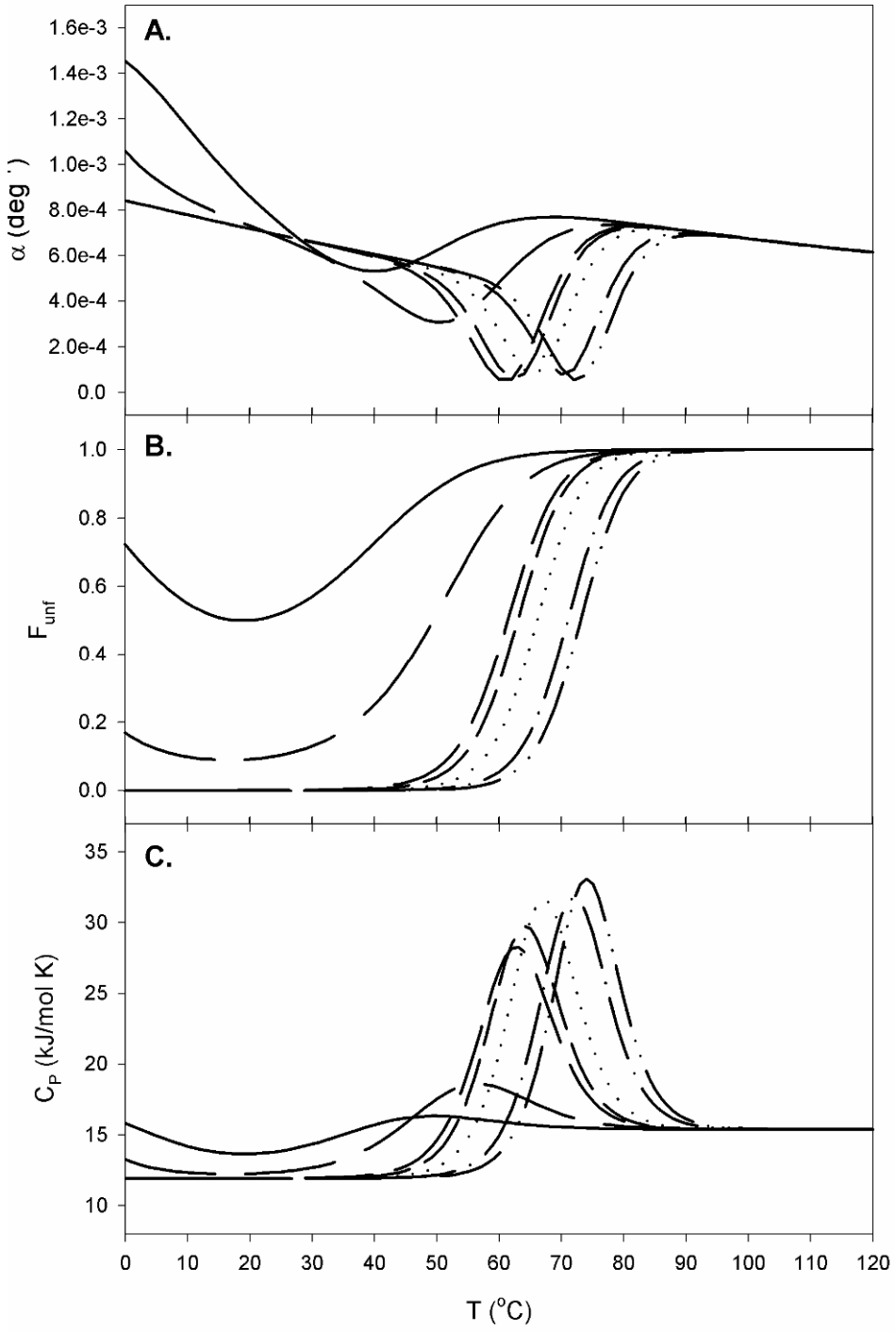

Figure 7.8 Simulated curves to show the relationship between **A.** α_{exp}, **B.** F_{unf}, and **C.** C_P. The curves were calculated using the parameters listed in Table 7.3. In terms of α_{exp}, cold denaturation is distinguished by a shift in the starting position and slope of the $\alpha_{exp}(T)$ profile (solid and long dashed lines), relative to the conditions under which no cold denaturation is observed.

REFERENCES

1. Kauzmann, W. (1959) Some factors in the interpretation of protein denaturation, *Adv. Protein Chem. 14*, 1-63.
2. Dill, K. A. (1990) Dominant forces in protein folding, *Biochemistry 29*, 7133-7155.
3. Matthews, B. W. (1995) Studies on protein stability with T4 lysozyme, *Adv. Protein Chem. 46*, 249-278.
4. Serrano, L., Kellis, J. T., Jr., Cann, P., Matouschek, A., and Fersht, A. R. (1992) The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability, *J Mol Biol 224*, 783-804.
5. Makhatadze, G. I., and Privalov, P. L. (1995) Energetics of protein structure, *Adv Protein Chem 47*, 307-425.
6. Desjarlais, J. R., and Handel, T. M. (1995) De novo design of the hydrophobic cores of proteins, *Protein Sci 4*, 2006-2018.
7. Lazar, G. A., Desjarlais, J. R., and Handel, T. M. (1997) De novo design of the hydrophobic core of ubiquitin, *Protein Sci 6*, 1167-1178.
8. Loladze, V. V., Ermolenko, D. N., and Makhatadze, G. I. (2002) Thermodynamic consequences of burial of polar and non-polar amino acid residues in the protein interior, *J Mol Biol 320*, 343-357.
9. Pace, C. N., Shirley, B. A., McNutt, M., and Gajiwala, K. (1996) Forces contributing to the conformational stability of proteins, *Faseb J 10*, 75-83.
10. Griko, Y. V., Makhatadze, G. I., Privalov, P. L., and Hartley, R. W. (1994) Thermodynamics of barnase unfolding, *Protein Sci 3*, 669-676.
11. Loladze, V. V., Ibarra-Molero, B., Sanchez-Ruiz, J. M., and Makhatadze, G. I. (1999) Engineering a thermostable protein via optimization of charge-charge interactions on the protein surface, *Biochemistry 38*, 16419-16423.
12. Grimsley, G. R., Shaw, K. L., Fee, L. R., Alston, R. W., Huyghues-Despointes, B. M., Thurlkill, R. L., Scholtz, J. M., and Pace, C. N. (1999) Increasing protein stability by altering long-range coulombic interactions, *Protein Sci 8*, 1843-1849.
13. Spector, S., Wang, M., Carp, S. A., Robblee, J., Hendsch, Z. S., Fairman, R., Tidor, B., and Raleigh, D. P. (2000) Rational modification of protein stability by the mutation of charged surface residues, *Biochemistry 39*, 872-879.
14. Perl, D., Mueller, U., Heinemann, U., and Schmid, F. X. (2000) Two exposed amino acid residues confer thermostability on a cold shock protein, *Nat Struct Biol 7*, 380-383.
15. Makhatadze, G. I., Loladze, V. V., Gribenko, A. V., and Lopez, M. M. (2004) Mechanism of thermostabilization in a designed cold shock protein with optimized surface electrostatic interactions, *J Mol Biol 336*, 929-942.
16. Strickler, S. S., Gribenko, A. V., Keiffer, T. R., Tomlinson, J., Reihle, T., Loladze, V. V., and Makhatadze, G. I. (2006) Protein stability and surface electrostatics: a charged relationship, *Biochemistry 45*, 2761-2766.
17. Loladze, V. V., Ermolenko, D. N., and Makhatadze, G. I. (2001) Heat capacity changes upon burial of polar and nonpolar groups in proteins, *Protein Sci 10*, 1343-1352.
18. Privalov, P. L. (1990) Cold denaturation of proteins, *Critical reviews in biochemistry and molecular biology 25*, 281-305.
19. Deutschman, W. A., and Dahlquist, F. W. (2001) Thermodynamic basis for the increased thermostability of CheY from the hyperthermophile Thermotoga maritima, *Biochemistry 40*, 13107-13113.

20. Beadle, B. M., Baase, W. A., Wilson, D. B., Gilkes, N. R., and Shoichet, B. K. (1999) Comparing the thermodynamic stabilities of a related thermophilic and mesophilic enzyme, *Biochemistry 38*, 2570-2576.
21. Grattinger, M., Dankesreiter, A., Schurig, H., and Jaenicke, R. (1998) Recombinant phosphoglycerate kinase from the hyperthermophilic bacterium Thermotoga maritima: catalytic, spectral and thermodynamic properties, *Journal of molecular biology 280*, 525-533.
22. Hollien, J., and Marqusee, S. (1999) A thermodynamic comparison of mesophilic and thermophilic ribonucleases H, *Biochemistry 38*, 3831-3836.
23. Guzman-Casado, M., Parody-Morreale, A., Robic, S., Marqusee, S., and Sanchez-Ruiz, J. M. (2003) Energetic evidence for formation of a pH-dependent hydrophobic cluster in the denatured state of Thermus thermophilus ribonuclease H, *Journal of molecular biology 329*, 731-743.
24. Farinas, E. T., Bulter, T., and Arnold, F. H. (2001) Directed enzyme evolution, *Curr Opin Biotechnol 12*, 545-551.
25. Bloom, J. D., Meyer, M. M., Meinhold, P., Otey, C. R., MacMillan, D., and Arnold, F. H. (2005) Evolving strategies for enzyme engineering, *Curr Opin Struct Biol 15*, 447-452.
26. Kumar, S., Chen, C. S., Waxman, D. J., and Halpert, J. R. (2005) Directed evolution of mammalian cytochrome P450 2B1: mutations outside of the active site enhance the metabolism of several substrates, including the anticancer prodrugs cyclophosphamide and ifosfamide, *J Biol Chem 280*, 19569-19575.
27. Morawski, B., Quan, S., and Arnold, F. H. (2001) Functional expression and stabilization of horseradish peroxidase by directed evolution in Saccharomyces cerevisiae, *Biotechnol Bioeng 76*, 99-107.
28. Giver, L., Gershenson, A., Freskgard, P. O., and Arnold, F. H. (1998) Directed evolution of a thermostable esterase, *Proc Natl Acad Sci U S A 95*, 12809-12813.
29. Hill, C. M., Li, W. S., Thoden, J. B., Holden, H. M., and Raushel, F. M. (2003) Enhanced degradation of chemical warfare agents through molecular engineering of the phosphotriesterase active site, *J Am Chem Soc 125*, 8990-8991.
30. Ness, J. E., Welch, M., Giver, L., Bueno, M., Cherry, J. R., Borchert, T. V., Stemmer, W. P., and Minshull, J. (1999) DNA shuffling of subgenomic sequences of subtilisin, *Nat Biotechnol 17*, 893-896.
31. Bulter, T., Alcalde, M., Sieber, V., Meinhold, P., Schlachtbauer, C., and Arnold, F. H. (2003) Functional expression of a fungal laccase in Saccharomyces cerevisiae by directed evolution, *Appl Environ Microbiol 69*, 987-995.
32. Cherry, J. R., Lamsa, M. H., Schneider, P., Vind, J., Svendsen, A., Jones, A., and Pedersen, A. H. (1999) Directed evolution of a fungal peroxidase, *Nat Biotechnol 17*, 379-384.
33. Murashima, K., Kosugi, A., and Doi, R. H. (2002) Thermostabilization of cellulosomal endoglucanase EngB from Clostridium cellulovorans by in vitro DNA recombination with non-cellulosomal endoglucanase EngD, *Mol Microbiol 45*, 617-626.
34. Hopfner, K. P., Kopetzki, E., Kresse, G. B., Bode, W., Huber, R., and Engh, R. A. (1998) New enzyme lineages by subdomain shuffling, *Proc Natl Acad Sci U S A 95*, 9813-9818.
35. Wintrode, P. L., and Arnold, F. H. (2000) Temperature adaptation of enzymes: lessons from laboratory evolution, *Adv. Protein Chem. 55*, 161-225.
36. Sieber, V., Pluckthun, A., and Schmid, F. X. (1998) Selecting proteins with improved stability by a phage-based method, *Nat Biotechnol 16*, 955-960.
37. Martin, A., Sieber, V., and Schmid, F. X. (2001) In-vitro selection of highly stabilized protein variants with optimized surface, *J Mol Biol 309*, 717-726.

38. Wunderlich, M., and Schmid, F. X. (2006) In vitro evolution of a hyperstable Gbeta1 variant, *J Mol Biol 363*, 545-557.
39. Finucane, M. D., Tuna, M., Lees, J. H., and Woolfson, D. N. (1999) Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries, *Biochemistry 38*, 11604-11612.
40. Finucane, M. D., and Woolfson, D. N. (1999) Core-directed protein design. II. Rescue of a multiply mutated and destabilized variant of ubiquitin, *Biochemistry 38*, 11613-11623.
41. Perl, D., and Schmid, F. X. (2001) Electrostatic stabilization of a thermophilic cold shock protein, *J Mol Biol 313*, 343-357.
42. Brockmann, E. C., Cooper, M., Stromsten, N., Vehniainen, M., and Saviranta, P. (2005) Selecting for antibody scFv fragments with improved stability using phage display with denaturation under reducing conditions, *J Immunol Methods 296*, 159-170.
43. Wunderlich, M., Martin, A., and Schmid, F. X. (2005) Stabilization of the cold shock protein CspB from Bacillus subtilis by evolutionary optimization of Coulombic interactions, *J Mol Biol 347*, 1063-1076.
44. Wunderlich, M., Martin, A., Staab, C. A., and Schmid, F. X. (2005) Evolutionary protein stabilization in comparison with computational design, *J Mol Biol 351*, 1160-1168.
45. Martin, A., Kather, I., and Schmid, F. X. (2002) Origins of the high stability of an in vitro-selected cold-shock protein, *J Mol Biol 318*, 1341-1349.
46. Tao, H., and Cornish, V. W. (2002) Milestones in directed enzyme evolution, *Curr Opin Chem Biol 6*, 858-864.
47. Fukuchi, S., and Nishikawa, K. (2001) Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria, *J Mol Biol 309*, 835-843.
48. Alsop, E., Silver, M., and Livesay, D. R. (2003) Optimized electrostatic surfaces parallel increased thermostability: a structural bioinformatic analysis, *Protein Eng 16*, 871-874.
49. Watanabe, K., Ohkuri, T., Yokobori, S., and Yamagishi, A. (2006) Designing thermostable proteins: ancestral mutants of 3-isopropylmalate dehydrogenase designed by using a phylogenetic tree, *J Mol Biol 355*, 664-674.
50. Neuwald, A. F., Liu, J. S., Lipman, D. J., and Lawrence, C. E. (1997) Extracting protein alignment models from the sequence database, *Nucleic Acids Research 25*, 1665-1677.
51. Shaw, E., and Dordick, J. S. (2002) Predicting amino acid residues responsible for enzyme specificity solely from protein sequences, *Biotechnol Bioeng 79*, 295-300.
52. DiTursi, M. K., Kwon, S. J., Reeder, P. J., and Dordick, J. S. (2006) Bioinformatics-driven, rational engineering of protein thermostability, *Protein Eng Des Sel 19*, 517-524.
53. Strickler, S. S., Gribenko, A. V., Gribenko, A. V., Keiffer, T. R., Tomlinson, J., Reihle, T., Loladze, V. V., and Makhatadze, G. I. (2006) Protein stability and surface electrostatics: a charged relationship, *Biochemistry 45*, 2761-2766.
54. Sun, D. P., Soderlind, E., Baase, W. A., Wozniak, J. A., Sauer, U., and Matthews, B. W. (1991) Cumulative site-directed charge-change replacements in bacteriophage T4 lysozyme suggest that long-range electrostatic interactions contribute little to protein stability, *J Mol Biol 221*, 873-887.
55. Sali, D., Bycroft, M., and Fersht, A. R. (1991) Surface electrostatic interactions contribute little of stability of barnase, *J Mol Biol 220*, 779-788.
56. Desjarlais, J. R., and Handel, T. M. (1995) New strategies in protein design, *Curr Opin Biotechnol 6*, 460-466.
57. DeGrado, W. F., Summa, C. M., Pavone, V., Nastri, F., and Lombardi, A. (1999) De novo design and structural characterization of proteins and metalloproteins, *Annu Rev Biochem 68*, 779-819.

58. Street, A. G., and Mayo, S. L. (1999) Computational protein design, *Structure 7*, R105-109.
59. Pokala, N., and Handel, T. M. (2001) Review: protein design--where we were, where we are, where we're going, *J Struct Biol 134*, 269-281.
60. Dantas, G., Kuhlman, B., Callender, D., Wong, M., and Baker, D. (2003) A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins, *Journal of molecular biology 332*, 449-460.
61. Korkegian, A., Black, M. E., Baker, D., and Stoddard, B. L. (2005) Computational thermostabilization of an enzyme, *Science 308*, 857-860.
62. Hurley, J. H., Baase, W. A., and Matthews, B. W. (1992) Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme, *J Mol Biol 224*, 1143-1159.
63. Kuhlman, B., and Baker, D. (2000) Native protein sequences are close to optimal for their structures, *Proc Natl Acad Sci U S A 97*, 10383-10388.
64. Baldwin, E. P., and Matthews, B. W. (1994) Core-packing constraints, hydrophobicity and protein design, *Curr Opin Biotechnol 5*, 396-402.
65. Desjarlais, J. R., and Handel, T. M. (1999) Side-chain and backbone flexibility in protein core design, *J Mol Biol 290*, 305-318.
66. Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M. A., Jaenicke, R., and Schmid, F. X. (1998) Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins, *Nat Struct Biol 5*, 229-235.
67. Sanchez-Ruiz, J. M., and Makhatadze, G. I. (2001) To charge or not to charge?, *Trends Biotechnol 19*, 132-135.
68. Makhatadze, G. I., Loladze, V. V., Ermolenko, D. N., Chen, X., and Thomas, S. T. (2003) Contribution of surface salt bridges to protein stability: guidelines for protein engineering, *J Mol Biol 327*, 1135-1148.
69. Huang, X., and Zhou, H. X. (2006) Similarity and difference in the unfolding of thermophilic and mesophilic cold shock proteins studied by molecular dynamics simulations, *Biophys J 91*, 2451-2463.
70. Motono, C., Gromiha, M. M., and Kumar, S. (2008) Thermodynamic and kinetic determinants of Thermotoga maritima cold shock protein stability: a structural and dynamic analysis, *Proteins 71*, 655-669.
71. Hamamatsu, N., Nomiya, Y., Aita, T., Nakajima, M., Husimi, Y., and Shibanaka, Y. (2006) Directed evolution by accumulating tailored mutations: thermostabilization of lactate oxidase with less trade-off with catalytic activity, *Protein Eng Des Sel 19*, 483-489.
72. Berezovsky, I. N., and Shakhnovich, E. I. (2005) Physics and evolution of thermophilic adaptation, *Proc Natl Acad Sci U S A 102*, 12742-12747.
73. Anderson, D. E., Becktel, W. J., and Dahlquist, F. W. (1990) pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme, *Biochemistry 29*, 2403-2408.
74. Horovitz, A., Serrano, L., Avron, B., Bycroft, M., and Fersht, A. R. (1990) Strength and co-operativity of contributions of surface salt bridges to protein stability, *J Mol Biol 216*, 1031-1044.
75. Serrano, L., Horovitz, A., Avron, B., Bycroft, M., and Fersht, A. R. (1990) Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles, *Biochemistry 29*, 9343-9352.
76. Predki, P. F., Agrawal, V., Brunger, A. T., and Regan, L. (1996) Amino-acid substitutions in a surface turn modulate protein stability, *Nat Struct Biol 3*, 54-58.

77. Nagi, A. D., and Regan, L. (1997) An inverse correlation between loop length and stability in a four-helix-bundle protein, *Fold Des 2*, 67-75.
78. Fernandez, A. M., Villegas, V., Martinez, J. C., Van Nuland, N. A., Conejero-Lara, F., Aviles, F. X., Serrano, L., Filimonov, V. V., and Mateo, P. L. (2000) Thermodynamic analysis of helix-engineered forms of the activation domain of human procarboxypeptidase A2, *Eur J Biochem 267*, 5891-5899.
79. Tanford, C., and Kirkwood, J. G. (1957) Theory of protein titration curves. I. General equations for impenetrable spheres, *J Am Chem Soc 79*, 5333-5339.
80. Matthew, J. B., Gurd, F. R., Garcia-Moreno, B., Flanagan, M. A., March, K. L., and Shire, S. J. (1985) pH-dependent processes in proteins, *CRC Crit Rev Biochem 18*, 91-197.
81. Matthew, J. B., and Gurd, F. R. (1986) Stabilization and destabilization of protein structure by charge interactions, *Methods Enzymol 130*, 437-453.
82. Matthew, J. B., and Gurd, F. R. (1986) Calculation of electrostatic interactions in proteins, *Methods Enzymol 130*, 413-436.
83. Richmond, T. J. (1984) Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect, *J Mol Biol 178*, 63-89.
84. Ibarra-Molero, B., Loladze, V. V., Makhatadze, G. I., and Sanchez-Ruiz, J. M. (1999) Thermal versus guanidine-induced unfolding of ubiquitin. An analysis in terms of the contributions from charge-charge interactions to protein stability, *Biochemistry 38*, 8138-8149.
85. Whitten, S. T., and Garcia-Moreno, E. B. (2000) pH dependence of stability of staphylococcal nuclease: evidence of substantial electrostatic interactions in the denatured state, *Biochemistry 39*, 14292-14304.
86. Swint-Kruse, L., and Robertson, A. D. (1995) Hydrogen bonds and the pH dependence of ovomucoid third domain stability, *Biochemistry 34*, 4724-4732.
87. Oliveberg, M., Arcus, V. L., and Fersht, A. R. (1995) pKA values of carboxyl groups in the native and denatured states of barnase: the pKA values of the denatured state are on average 0.4 units lower than those of model compounds, *Biochemistry 34*, 9424-9433.
88. Kuhlman, B., Luisi, D. L., Young, P., and Raleigh, D. P. (1999) pKa values and the pH dependent stability of the N-terminal domain of L9 as probes of electrostatic interactions in the denatured state. Differentiation between local and nonlocal interactions, *Biochemistry 38*, 4896-4903.
89. Cho, J. H., and Raleigh, D. P. (2005) Mutational analysis demonstrates that specific electrostatic interactions can play a key role in the denatured state ensemble of proteins, *J Mol Biol 353*, 174-185.
90. Anil, B., Craig-Schapiro, R., and Raleigh, D. P. (2006) Design of a hyperstable protein by rational consideration of unfolded state interactions, *J Am Chem Soc 128*, 3144-3145.
91. Pace, C. N., Alston, R. W., and Shaw, K. L. (2000) Charge-charge interactions influence the denatured state ensemble and contribute to protein stability, *Protein Sci 9*, 1395-1398.
92. Elcock, A. H., and McCammon, J. A. (1998) Electrostatic contributions to the stability of halophilic proteins, *J Mol Biol 280*, 731-748.
93. Zhou, H. X. (2002) A Gaussian-chain model for treating residual charge-charge interactions in the unfolded state of proteins, *Proc Natl Acad Sci U S A 99*, 3569-3574.
94. Pace, C. N., Laurents, D. V., and Thomson, J. A. (1990) pH dependence of the urea and guanidine hydrochloride denaturation of ribonuclease A and ribonuclease T1, *Biochemistry 29*, 2564-2572.

95. Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000) Comparative protein structure modeling of genes and genomes, *Annu Rev Biophys Biomol Struct 29*, 291-325.
96. Permyakov, S. E., Makhatadze, G. I., Owenius, R., Uversky, V. N., Brooks, C. L., Permyakov, E. A., and Berliner, L. J. (2005) How to improve nature: study of the electrostatic properties of the surface of alpha-lactalbumin, *Protein Eng Des Sel 18*, 425-433.
97. Schweiker, K. L., Zarrine-Afsar, A., Davidson, A. R., and Makhatadze, G. I. (2007) Computational design of the Fyn SH3 domain with increased stability through optimization of surface charge-charge interactions, *Protein Sci 16*, 2694-2702.
98. Lee, C. F., Makhatadze, G. I., and Wong, K. B. (2005) Effects of charge-to-alanine substitutions on the stability of ribosomal protein L30e from Thermococcus celer, *Biochemistry 44*, 16817-16825.
99. Godoy-Ruiz, R., Perez-Jimenez, R., Garcia-Mira, M. M., Plaza del Pino, I. M., and Sanchez-Ruiz, J. M. (2005) Empirical parametrization of pK values for carboxylic acids in proteins using a genetic algorithm, *Biophys Chem 115*, 263-266.
100. Ibarra-Molero, B., and Sanchez-Ruiz, J. M. (2002) Genetic algorithm to design stabilizing surface-charge distributions in proteins, *J Phys Chem B 106*, 6609-6613.
101. Pace, C. N. (1986) Determination and analysis of urea and guanidine hydrochloride denaturation curves, *Methods Enzymol 131*, 266-280.
102. Frankenberg, N., Welker, C., and Jaenicke, R. (1999) Does the elimination of ion pairs affect the thermal stability of cold shock protein from the hyperthermophilic bacterium Thermotoga maritima?, *FEBS letters 454*, 299-302.
103. Strop, P., and Mayo, S. L. (2000) Contribution of surface salt bridges to protein stability, *Biochemistry 39*, 1251-1255.
104. Tissot, A. C., Vuilleumier, S., and Fersht, A. R. (1996) Importance of two buried salt bridges in the stability and folding pathway of barnase, *Biochemistry 35*, 6786-6794.
105. Marqusee, S., and Sauer, R. T. (1994) Contributions of a hydrogen bond/salt bridge network to the stability of secondary and tertiary structure in lambda repressor, *Protein Sci 3*, 2217-2225.
106. Nicholson, H., Anderson, D. E., Dao-pin, S., and Matthews, B. W. (1991) Analysis of the interaction between charged side chains and the alpha-helix dipole using designed thermostable mutants of phage T4 lysozyme, *Biochemistry 30*, 9816-9828.
107. Merkel, J. S., Sturtevant, J. M., and Regan, L. (1999) Sidechain interactions in parallel beta sheets: the energetics of cross-strand pairings, *Structure 7*, 1333-1343.
108. Lassila, K. S., Datta, D., and Mayo, S. L. (2002) Evaluation of the energetic contribution of an ionic network to beta-sheet stability, *Protein Sci 11*, 688-690.
109. Blasie, C. A., and Berg, J. M. (1997) Electrostatic interactions across a beta-sheet, *Biochemistry 36*, 6218-6222.
110. Gribenko, A. V., and Makhatadze, G. I. (2007) Role of the Charge-Charge Interactions in Defining Stability and Halophilicity of the CspB Proteins, *J Mol Biol 366*, 842-856.
111. Chakrabartty, A., Doig, A. J., and Baldwin, R. L. (1993) Helix capping propensities in peptides parallel those in proteins, *Proc Natl Acad Sci U S A 90*, 11332-11336.
112. Doig, A. J., and Baldwin, R. L. (1995) N- and C-capping preferences for all 20 amino acids in alpha-helical peptides, *Protein Sci 4*, 1325-1336.
113. Viguera, A. R., and Serrano, L. (1995) Experimental analysis of the Schellman motif, *J Mol Biol 251*, 150-160.
114. Gong, Y., Zhou, H. X., Guo, M., and Kallenbach, N. R. (1995) Structural analysis of the N- and C-termini in a peptide with consensus sequence, *Protein Sci 4*, 1446-1456.

115. Aurora, R., and Rose, G. D. (1998) Helix capping, *Protein Sci 7*, 21-38.
116. Thomas, S. T., and Makhatadze, G. I. (2000) Contribution of the 30/36 hydrophobic contact at the C-terminus of the alpha-helix to the stability of the ubiquitin molecule, *Biochemistry 39*, 10275-10283.
117. Thomas, S. T., Loladze, V. V., and Makhatadze, G. I. (2001) Hydration of the peptide backbone largely defines the thermodynamic propensity scale of residues at the C' position of the C-capping box of alpha-helices, *Proc Natl Acad Sci U S A 98*, 10670-10675.
118. Ermolenko, D. N., and Makhatadze, G. I. (2002) Bacterial cold-shock proteins, *Cell Mol Life Sci 59*, 1902-1913.
119. Marshall, S. A., Morgan, C. S., and Mayo, S. L. (2002) Electrostatics significantly affect the stability of designed homeodomain variants, *J Mol Biol 316*, 189-199.
120. Dominy, B. N., Perl, D., Schmid, F. X., and Brooks, C. L., 3rd. (2002) The effects of ionic strength on protein stability: the cold shock protein family, *J Mol Biol 319*, 541-554.
121. Kao, Y. H., Fitch, C. A., Bhattacharya, S., Sarkisian, C. J., Lecomte, J. T., and Garcia-Moreno, E. B. (2000) Salt effects on ionization equilibria of histidines in myoglobin, *Biophys J 79*, 1637-1654.
122. Lee, K. K., Fitch, C. A., and Garcia-Moreno, E. B. (2002) Distance dependence and salt sensitivity of pairwise, coulombic interactions in a protein, *Protein Sci 11*, 1004-1016.
123. Gribenko, A. V., Patel, M. M., and Makhatadze, G. I. (2007) to be published.
124. Lopez, M. M., Yutani, K., and Makhatadze, G. I. (1999) Interactions of the major cold shock protein of Bacillus subtilis CspB with single-stranded DNA templates of different base composition, *J Biol Chem 274*, 33601-33608.
125. Lopez, M. M., and Makhatadze, G. I. (2000) Major cold shock proteins, CspA from Escherichia coli and CspB from Bacillus subtilis, interact differently with single-stranded DNA templates, *Biochim Biophys Acta 1479*, 196-202.
126. Lopez, M. M., Yutani, K., and Makhatadze, G. I. (2001) Interactions of the cold shock protein CspB from Bacillus subtilis with single-stranded DNA. Importance of the T base content and position within the template, *J Biol Chem 276*, 15511-15518.
127. Zeeb, M., Max, K. E., Weininger, U., Low, C., Sticht, H., and Balbach, J. (2006) Recognition of T-rich single-stranded DNA by the cold shock protein Bs-CspB in solution, *Nucleic Acids Res 34*, 4561-4571.
128. Max, K. E., Zeeb, M., Bienert, R., Balbach, J., and Heinemann, U. (2006) T-rich DNA single strands bind to a preformed site on the bacterial cold shock protein Bs-CspB, *J Mol Biol 360*, 702-714.
129. Ramponi, G., Treves, C., and Guerritore, A. (1967) Hydrolytic activity of muscle acyl phosphatase on 3-phosphoglyceryl phosphate, *Experientia 23*, 1019-1020.
130. Ramponi, G., Treves, C., and Guerritore, A. A. (1966) Aromatic acyl phosphates as substrates of acyl phosphatase, *Archives of biochemistry and biophysics 115*, 129-135.
131. Shoichet, B. K., Baase, W. A., Kuroki, R., and Matthews, B. W. (1995) A relationship between protein stability and protein function, *Proc Natl Acad Sci U S A 92*, 452-456.
132. Meiering, E. M., Serrano, L., and Fersht, A. R. (1992) Effect of active site residues in barnase on activity and stability, *J Mol Biol 225*, 585-589.
133. Zhang, X. J., Baase, W. A., Shoichet, B. K., Wilson, K. P., and Matthews, B. W. (1995) Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive, *Protein Eng 8*, 1017-1022.

134. Varley, P. G., and Pain, R. H. (1991) Relation between stability, dynamics and enzyme activity in 3-phosphoglycerate kinases from yeast and Thermus thermophilus, *J Mol Biol 220*, 531-538.
135. Schutz, C. N., and Warshel, A. (2001) What are the dielectric "constants" of proteins and how to validate electrostatic models?, *Proteins 44*, 400-417.
136. Warshel, A. (2003) Computer simulations of enzyme catalysis: methods, progress, and insights, *Annu Rev Biophys Biomol Struct 32*, 425-443.
137. Garcia-Moreno, E. B., and Fitch, C. A. (2004) Structural interpretation of pH and salt-dependent processes in proteins with computational methods, *Methods Enzymol 380*, 20-51.
138. Antosiewicz, J., McCammon, J. A., and Gilson, M. K. (1994) Prediction of pH-dependent properties of proteins, *J Mol Biol 238*, 415-436.
139. Antosiewicz, J., McCammon, J. A., and Gilson, M. K. (1996) The determinants of pKas in proteins, *Biochemistry 35*, 7819-7833.
140. Alexov, E. G., and Gunner, M. R. (1997) Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties, *Biophys J 72*, 2075-2093.
141. Georgescu, R. E., Alexov, E. G., and Gunner, M. R. (2002) Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins, *Biophys J 83*, 1731-1748.
142. Mehler, E. L., and Guarnieri, F. (1999) A self-consistent, microenvironment modulated screened coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins, *Biophys J 77*, 3-22.
143. Lee, P. S., Chu, Z. T., and Warshel, A. (1993) Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the POLARIS and ENZYMIX programs, *J Comput Chem 14*, 161-185.
144. Sham, Y. Y., Muegge, I., and Warshel, A. (1998) The effect of protein relaxation on charge-charge interactions and dielectric constants of proteins, *Biophys J 74*, 1744-1753.
145. Trefethen, J. M., Pace, C. N., Scholtz, J. M., and Brems, D. N. (2005) Charge-charge interactions in the denatured state influence the folding kinetics of ribonuclease Sa, *Protein Sci 14*, 1934-1938.
146. Fersht, A., and Winter, G. (1992) Protein engineering, *Trends Biochem Sci 17*, 292-295.
147. Loladze, V. V., and Makhatadze, G. I. (2005) Both helical propensity and side-chain hydrophobicity at a partially exposed site in alpha-helix contribute to the thermodynamic stability of ubiquitin, *Proteins 58*, 1-6.
148. Ermolenko, D. N., Thomas, S. T., Aurora, R., Gronenborn, A. M., and Makhatadze, G. I. (2002) Hydrophobic interactions at the Ccap position of the C-capping motif of alpha-helices, *J Mol Biol 322*, 123-135.
149. Zollars, E. S., Marshall, S. A., and Mayo, S. L. (2006) Simple electrostatic model improves designed protein sequences, *Protein Sci 15*, 2014-2018.
150. Bolon, D. N., Marcus, J. S., Ross, S. A., and Mayo, S. L. (2003) Prudent modeling of core polar residues in computational protein design, *J Mol Biol 329*, 611-622.
151. Dantas, G., Corrent, C., Reichow, S. L., Havranek, J. J., Eletr, Z. M., Isern, N. G., Kuhlman, B., Varani, G., Merritt, E. A., and Baker, D. (2007) High-resolution Structural and Thermodynamic Analysis of Extreme Stabilization of Human Procarboxypeptidase by Computational Protein Design, *J Mol Biol 366*, 1209-1221.
152. Schweiker, K. L., Zarrine-Afsar, A., Davidson, A. R., and Makhatadze, G. I. (2007) Computational design of the Fyn SH3 domain with increased stability through optimization of surface charge charge interactions, *Protein Sci 16*, 2694-2702.

153. Vijay-Kumar, S., Bugg, C. E., and Cook, W. J. (1987) Structure of ubiquitin refined at 1.8 A resolution, *Journal of molecular biology 194*, 531-544.
154. Garcia-Saez, I., Reverter, D., Vendrell, J., Aviles, F. X., and Coll, M. (1997) The three-dimensional structure of human procarboxypeptidase A2. Deciphering the basis of the inhibition, activation and intrinsic activity of the zymogen, *The EMBO journal 16*, 6906-6913.
155. Thunnissen, M. M., Taddei, N., Liguri, G., Ramponi, G., and Nordlund, P. (1997) Crystal structure of common type acylphosphatase from bovine testis, *Structure 5*, 69-79.
156. Musacchio, A., Saraste, M., and Wilmanns, M. (1994) High-resolution crystal structures of tyrosine kinase SH3 domains complexed with proline-rich peptides, *Nature structural biology 1*, 546-551.
157. Leahy, D. J., Hendrickson, W. A., Aukhil, I., and Erickson, H. P. (1992) Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein, *Science (New York, N.Y 258*, 987-991.
158. Oubridge, C., Ito, N., Evans, P. R., Teo, C. H., and Nagai, K. (1994) Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin, *Nature 372*, 432-438.
159. Rudolph, M. G., Wittinghofer, A., and Vetter, I. R. (1999) Nucleotide binding to the G12V-mutant of Cdc42 investigated by X-ray diffraction and fluorescence spectroscopy: two different nucleotide states in one crystal, *Protein Sci 8*, 778-787.
160. Maxwell, K. L., and Davidson, A. R. (1998) Mutagenesis of a buried polar interaction in an SH3 domain: sequence conservation provides the best prediction of stability effects, *Biochemistry 37*, 16172-16182.
161. Edelhoch, H. (1967) Spectroscopic determination of tryptophan and tyrosine in proteins, *Biochemistry 6*, 1948-1954.
162. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G., and Gray, T. (1995) How to measure and predict the molar absorption coefficient of a protein, *Protein Sci 4*, 2411-2423.
163. Gribenko, A. V., Keiffer, T. R., and Makhatadze, G. I. (2006) Amino acid substitutions affecting protein dynamics in eglin C do not affect heat capacity change upon unfolding, *Proteins 64*, 295-300.
164. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., and Bairoch, A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis, *Nucleic Acids Res 31*, 3784-3788.
165. Makhatadze, G. I. (1998) Measuring protein thermostability by differential scanning calorimetry, in *Current Protocols in Protein Chemistry, 2* (Wiley, T. J., Ed.), John Wiley & Sons, New York.
166. Makhatadze, G. I., Medvedkin, V. N., and Privalov, P. L. (1990) Partial molar volumes of polypeptides and their constituent groups in aqueous solution over a broad temperature range, *Biopolymers 30*, 1001-1010.
167. Lopez, M. M., and Makhatadze, G. I. (2002) Differential scanning calorimetry, *Methods Mol Biol 173*, 113-119.
168. Privalov, P. L. (1979) Stability of proteins: small globular proteins, *Adv Protein Chem 33*, 167-241.
169. Sherrod, P. H. (1998) Nonlinear Regression Analysis Program, NLREG Version 4.1, Phillip H. Sherrod, Nashville, TN.
170. Pace, C. N., and Shaw, K. L. (2000) Linear extrapolation method of analyzing solvent denaturation curves, *Proteins Suppl 4*, 1-7.

171. Myers, J. K., Pace, C. N., and Scholtz, J. M. (1995) Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding, *Protein Sci 4*, 2138-2148.
172. Jha, A. K., Colubri, A., Freed, K. F., and Sosnick, T. R. (2005) Statistical coil model of the unfolded state: resolving the reconciliation problem, *Proc Natl Acad Sci U S A 102*, 13099-13104.
173. Tran, H. T., and Pappu, R. V. (2006) Toward an accurate theoretical framework for describing ensembles for proteins under strongly denaturing conditions, *Biophys J 91*, 1868-1886.
174. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters, *Proteins 65*, 712-725.
175. Case, D. A., Darden, T. A., Cheatham III, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Merz, K. M., Pearlman, D. A., Crowley, M., RWalker, R. C., Zhang, W., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Wong, K. F., Paesani, F., Wu, X., Brozell, S., Tsui, V., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Beroza, P., Mathews, D. H., Schafmeister, C., Ross, W. S., and Kollman, P. A. (2006) AMBER 9, University of California, San Fransisco.
176. Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics, *J Am Chem Soc 112*, 6127-6129.
177. Clarke, J., Hamill, S. J., and Johnson, C. M. (1997) Folding and stability of a fibronectin type III domain of human tenascin, *J Mol Biol 270*, 771-778.
178. Maxwell, K. L., Wildes, D., Zarrine-Afsar, A., De Los Rios, M. A., Brown, A. G., Friel, C. T., Hedberg, L., Horng, J. C., Bona, D., Miller, E. J., Vallee-Belisle, A., Main, E. R., Bemporad, F., Qiu, L., Teilum, K., Vu, N. D., Edwards, A. M., Ruczinski, I., Poulsen, F. M., Kragelund, B. B., Michnick, S. W., Chiti, F., Bai, Y., Hagen, S. J., Serrano, L., Oliveberg, M., Raleigh, D. P., Wittung-Stafshede, P., Radford, S. E., Jackson, S. E., Sosnick, T. R., Marqusee, S., Davidson, A. R., and Plaxco, K. W. (2005) Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins, *Protein Sci 14*, 602-616.
179. Lin, L. N., Brandts, J. F., Brandts, J. M., and Plotnikov, V. (2002) Determination of the volumetric properties of proteins and other solutes using pressure perturbation calorimetry, *Analytical biochemistry 302*, 144-160.
180. Northey, J. G., Maxwell, K. L., and Davidson, A. R. (2002) Protein folding kinetics beyond the phi value: using multiple amino acid substitutions to investigate the structure of the SH3 domain folding transition state, *J Mol Biol 320*, 389-402.
181. Larson, S. M., Ruczinski, I., Davidson, A. R., Baker, D., and Plaxco, K. W. (2002) Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation, *J Mol Biol 316*, 225-233.
182. Northey, J. G., Di Nardo, A. A., and Davidson, A. R. (2002) Hydrophobic core packing in the SH3 domain folding transition state, *Nat Struct Biol 9*, 126-130.
183. Zarrine-Afsar, A., Mittermaier, A., Kay, L. E., and Davidson, A. R. (2006) Protein stabilization by specific binding of guanidinium to a functional arginine-binding surface on an SH3 domain, *Protein Sci 15*, 162-170.
184. Di Nardo, A. A., Larson, S. M., and Davidson, A. R. (2003) The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core, *J Mol Biol 333*, 641-655.

185. Privalov, P. L., and Khechinashvili, N. N. (1974) A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study, *J Mol Biol 86*, 665-684.
186. Filimonov, V. V., Azuaga, A. I., Viguera, A. R., Serrano, L., and Mateo, P. L. (1999) A thermodynamic analysis of a family of small globular proteins: SH3 domains, *Biophys Chem 77*, 195-208.
187. Gribenko, A. V., Patel, M. M., and Makhatadze, G. I. (2009) Rational stabilization of enzymes by computational redesign of surface charge-charge interactions, *Proc Natl Acad Sci U S A accepted manuscript*.
188. Cho, J. H., Sato, S., and Raleigh, D. P. (2004) Thermodynamics and kinetics of non-native interactions in protein folding: a single point mutant significantly stabilizes the N-terminal domain of L9 by modulating non-native interactions in the denatured state, *J Mol Biol 338*, 827-837.
189. Cho, J. H., and Raleigh, D. P. (2006) Electrostatic interactions in the denatured state and in the transition state for protein folding: effects of denatured state interactions on the analysis of transition state structure, *J Mol Biol 359*, 1437-1446.
190. Tan, Y. J., Oliveberg, M., Davis, B., and Fersht, A. R. (1995) Perturbed pKA-values in the denatured states of proteins, *J Mol Biol 254*, 980-992.
191. Yang, A. S., and Honig, B. (1993) On the pH dependence of protein stability, *J Mol Biol 231*, 459-474.
192. Pfeil, W., and Privalov, P. L. (1976) Thermodynamic investigations of proteins. III. Thermodynamic descrption of lysozyme, *Biophys Chem 4*, 41-50.
193. Zhou, H. X. (2002) Residual electrostatic effects in the unfolded state of the N-terminal domain of L9 can be attributed to nonspecific nonlocal charge-charge interactions, *Biochemistry 41*, 6533-6538.
194. Zhou, H. X. (2004) Polymer models of protein stability, folding, and interactions, *Biochemistry 43*, 2141-2154.
195. Goldenberg, D. P. (2003) Computational simulation of the statistical properties of unfolded proteins, *J Mol Biol 326*, 1615-1633.
196. Wang, Y., Trewhella, J., and Goldenberg, D. P. (2008) Small-angle X-ray scattering of reduced ribonuclease A: effects of solution conditions and comparisons with a computational model of unfolded proteins, *J Mol Biol 377*, 1576-1592.
197. Kundrotas, P. J., and Karshikoff, A. (2002) Modeling of denatured state for calculation of the electrostatic contribution to protein stability, *Protein Sci 11*, 1681-1686.
198. Kundrotas, P. J., and Karshikoff, A. (2002) Model for calculation of electrostatic interactions in unfolded proteins, *Physical review 65*, 011901.
199. Kundrotas, P. J., and Karshikoff, A. (2004) Charge sequence coding in statistical modeling of unfolded proteins, *Biochim Biophys Acta 1702*, 1-8.
200. Luisi, D. L., Wu, W. J., and Raleigh, D. P. (1999) Conformational analysis of a set of peptides corresponding to the entire primary sequence of the N-terminal domain of the ribosomal protein L9: evidence for stable native-like secondary structure in the unfolded state, *J Mol Biol 287*, 395-407.
201. Li, Y., Picart, F., and Raleigh, D. P. (2005) Direct characterization of the folded, unfolded and urea-denatured states of the C-terminal domain of the ribosomal protein L9, *J Mol Biol 349*, 839-846.
202. Millett, I. S., Doniach, S., and Plaxco, K. W. (2002) Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins, *Adv Protein Chem 62*, 241-262.

203. Vitalis, A., Wang, X., and Pappu, R. V. (2008) Atomistic simulations of the effects of polyglutamine chain length and solvent quality on conformational equilibria and spontaneous homodimerization, *J Mol Biol 384*, 279-297.
204. Gelin, B. R., and Karplus, M. (1979) Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment, *Biochemistry 18*, 1256-1268.
205. Bieri, O., and Kiefhaber, T. (1999) Elementary steps in protein folding, *Biological chemistry 380*, 923-929.
206. Privalov, P. L., Tiktopulo, E. I., Venyaminov, S., Griko Yu, V., Makhatadze, G. I., and Khechinashvili, N. N. (1989) Heat capacity and conformation of proteins in the denatured state, *J Mol Biol 205*, 737-750.
207. Tanford, C., Buzzell, J. G., Rands, D. G., and Swanson, S. A. (1955) The Reversible Expansion of Bovine Serum Albumin in Acid Solutions, *Journal of the American Chemical Society 77*, 6421-6428.
208. Tanford, C. (1955) Intrinsic Viscosity and Kinematic Viscosity, *Journal of Physical Chemistry 59*, 798-799.
209. Buzzell, J. G., and Tanford, C. (1956) The Effect of Charge and Ionic Strength on the Viscosity of Ribonuclease, *Journal of Physical Chemistry 60*, 1204-1207.
210. Tanford, C., and Buzzell, J. G. (1956) The Viscosity of Aqueous Solutions of Bovine Serum Albumin between Ph 4.3 and 10.5, *Journal of Physical Chemistry 60*, 225-231.
211. Vitalis, A., Wang, X., and Pappu, R. V. (2007) Quantitative characterization of intrinsic disorder in polyglutamine: insights from analysis based on polymer theories, *Biophys J 93*, 1923-1937.
212. Baker, D., and Agard, D. A. (1994) Kinetics versus thermodynamics in protein folding, *Biochemistry 33*, 7505-7509.
213. Kelch, B. A., and Agard, D. A. (2007) Mesophile versus thermophile: insights into the structural mechanisms of kinetic stability, *J Mol Biol 370*, 784-795.
214. Kelch, B. A., Eagen, K. P., Erciyas, F. P., Humphris, E. L., Thomason, A. R., Mitsuiki, S., and Agard, D. A. (2007) Structural and mechanistic exploration of acid resistance: kinetic stability facilitates evolution of extremophilic behavior, *J Mol Biol 368*, 870-883.
215. Manning, M., and Colon, W. (2004) Structural basis of protein kinetic stability: resistance to sodium dodecyl sulfate suggests a central role for rigidity and a bias toward beta-sheet structure, *Biochemistry 43*, 11248-11254.
216. Johnson, E. C., Lazar, G. A., Desjarlais, J. R., and Handel, T. M. (1999) Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin, *Structure 7*, 967-976.
217. Fersht, A. R., Matouschek, A., and Serrano, L. (1992) The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding, *J Mol Biol 224*, 771-782.
218. Fersht, A. R., Itzhaki, L. S., elMasry, N. F., Matthews, J. M., and Otzen, D. E. (1994) Single versus parallel pathways of protein folding and fractional formation of structure in the transition state, *Proc Natl Acad Sci U S A 91*, 10426-10429.
219. Plaxco, K. W., Guijarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D., and Dobson, C. M. (1998) The folding kinetics and thermodynamics of the Fyn-SH3 domain, *Biochemistry 37*, 2529-2537.
220. Villegas, V., Martinez, J. C., Aviles, F. X., and Serrano, L. (1998) Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain, *J Mol Biol 283*, 1027-1036.
221. Martinez, J. C., and Serrano, L. (1999) The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved, *Nat Struct Biol 6*, 1010-1016.

222. Grantcharova, V. P., Riddle, D. S., and Baker, D. (2000) Long-range order in the src SH3 folding transition state, *Proc Natl Acad Sci U S A 97*, 7084-7089.
223. Hamill, S. J., Steward, A., and Clarke, J. (2000) The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology, *J Mol Biol 297*, 165-178.
224. Cota, E., Steward, A., Fowler, S. B., and Clarke, J. (2001) The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold, *J Mol Biol 305*, 1185-1194.
225. Garcia-Mira, M. M., Boehringer, D., and Schmid, F. X. (2004) The folding transition state of the cold shock protein is strongly polarized, *J Mol Biol 339*, 555-569.
226. Lopez, M. M., Zarrine-Afsar, A., and Davidson, A. R. (2007) unpublished data.
227. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I., and Baker, D. (1999) Experiment and theory highlight role of native state topology in SH3 folding, *Nat Struct Biol 6*, 1016-1024.
228. Di Nardo, A. A., Korzhnev, D. M., Stogios, P. J., Zarrine-Afsar, A., Kay, L. E., and Davidson, A. R. (2004) Dramatic acceleration of protein folding by stabilization of a nonnative backbone conformation, *Proc Natl Acad Sci U S A 101*, 7954-7959.
229. Zarrine-Afsar, A., Dahesh, S., and Davidson, A. R. (2007) Protein folding kinetics provides a context-independent assessment of beta-strand propensity in the Fyn SH3 domain, *J Mol Biol 373*, 764-774.
230. Neudecker, P., Zarrine-Afsar, A., Davidson, A. R., and Kay, L. E. (2007) Phi-value analysis of a three-state protein folding pathway by NMR relaxation dispersion spectroscopy, *Proc Natl Acad Sci U S A 104*, 15717-15722.
231. Perl, D., Holtermann, G., and Schmid, F. X. (2001) Role of the chain termini for the folding transition state of the cold shock protein, *Biochemistry 40*, 15501-15511.
232. Gvritishvili, A. G., Gribenko, A. V., and Makhatadze, G. I. (2008) Cooperativity of complex salt bridges, *Protein Science 17*, 1285-1290.
233. Greaves, R. B., and Warwicker, J. (2007) Mechanisms for stabilisation and the maintenance of solubility in proteins from thermophiles, *BMC structural biology 7*, 18.
234. Kumar, S., and Nussinov, R. (2002) Close-range electrostatic interactions in proteins, *Chembiochem 3*, 604-617.
235. Loladze, V. V., and Makhatadze, G. I. (2002) Removal of surface charge-charge interactions from ubiquitin leaves the protein folded and very stable, *Protein Sci 11*, 174-177.
236. Shaw, K. L., Grimsley, G. R., Yakovlev, G. I., Makarov, A. A., and Pace, C. N. (2001) The effect of net charge on the solubility, activity, and stability of ribonuclease Sa, *Protein Sci 10*, 1206-1215.
237. Trevino, S. R., Scholtz, J. M., and Pace, C. N. (2007) Amino Acid Contribution to Protein Solubility: Asp, Glu, and Ser Contribute more Favorably than the other Hydrophilic Amino Acids in RNase Sa, *J Mol Biol 366*, 449-460.
238. Mayo, K. H., Ilyina, E., and Park, H. (1996) A recipe for designing water-soluble, beta-sheet-forming peptides, *Protein Science 5*, 1301-1315.
239. Mosavi, L. K., and Peng, Z. Y. (2003) Structure-based substitutions for increased solubility of a designed protein, *Protein Engineering 16*, 739-745.
240. Slovic, A. M., Summa, C. M., Lear, J. D., and DeGrado, W. F. (2003) Computational design of a water-soluble analog of phospholamban, *Protein Science 12*, 337-348.
241. Tang, Y. C., and Deber, C. M. (2004) Aqueous solubility and membrane interactions of hydrophobic peptides with peptoid tags, *Biopolymers 76*, 110-118.

242. Kato, A., Maki, K., Ebina, T., Kuwajima, K., Soda, K., and Kuroda, Y. (2007) Mutational analysis of protein solubility enhancement using short peptide tags, *Biopolymers 85*, 12-18.
243. Lawrence, M. S., Phillips, K. J., and Liu, D. R. (2007) Supercharging proteins can impart unusual resilience, *J Am Chem Soc 129*, 10110-10112.
244. Wintrode, P. L., Makhatadze, G. I., and Privalov, P. L. (1994) Thermodynamics of ubiquitin unfolding, *Proteins 18*, 246-253.
245. Ooi, T., and Oobatake, M. (1988) Effects of hydrated water on protein unfolding, *Journal of biochemistry 103*, 114-120.
246. Baldwin, R. L. (1986) Temperature dependence of the hydrophobic interaction in protein folding, *Proc Natl Acad Sci U S A 83*, 8069-8072.
247. Makhatadze, G. I., Gill, S. J., and Privalov, P. L. (1990) Partial molar heat capacities of the side chains of some amino acid residues in aqueous solution. The influence of the neighboring charges, *Biophys Chem 38*, 33-37.
248. Makhatadze, G. I., and Privalov, P. L. (1990) Heat capacity of proteins. I. Partial molar heat capacity of individual amino acid residues in aqueous solution: hydration effect, *J Mol Biol 213*, 375-384.
249. Khechinashvili, N. N. (1990) Thermodynamic properties of globular proteins and the principle of stabilization of their native structure, *Biochim Biophys Acta 1040*, 346-354.
250. Privalov, P. L., and Makhatadze, G. I. (1990) Heat capacity of proteins. II. Partial molar heat capacity of the unfolded polypeptide chain of proteins: protein unfolding effects, *J Mol Biol 213*, 385-391.
251. Livingstone, J. R., Spolar, R. S., and Record, M. T., Jr. (1991) Contribution to the thermodynamics of protein folding from the reduction in water-accessible nonpolar surface area, *Biochemistry 30*, 4237-4244.
252. Privalov, P. L., and Makhatadze, G. I. (1992) Contribution of hydration and non-covalent interactions to the heat capacity effect on protein unfolding, *J Mol Biol 224*, 715-723.
253. Spolar, R. S., Livingstone, J. R., and Record, M. T., Jr. (1992) Use of liquid hydrocarbon and amide transfer data to estimate contributions to thermodynamic functions of protein folding from the removal of nonpolar and polar surface from water, *Biochemistry 31*, 3947-3955.
254. Makhatadze, G. I., and Privalov, P. L. (1993) Contribution of hydration to protein folding thermodynamics. I. The enthalpy of hydration, *J Mol Biol 232*, 639-659.
255. Privalov, P. L., and Makhatadze, G. I. (1993) Contribution of hydration to protein folding thermodynamics. II. The entropy and Gibbs energy of hydration, *J Mol Biol 232*, 660-679.
256. Oobatake, M., and Ooi, T. (1993) Hydration and heat stability effects on protein unfolding, *Progress in biophysics and molecular biology 59*, 237-284.
257. Kauzmann, W. (1987) Thermodynamics of unfolding, *Nature 325*, 763-764.
258. Gross, M., and Jaenicke, R. (1994) Proteins under pressure. The influence of high hydrostatic pressure on structure, function and assembly of proteins and protein complexes, *Eur J Biochem 221*, 617-630.
259. Brandts, J. F. (1969) in *Structure and Stability of Biological Macromelecules* (Timasheff, S. N., and Fasman, G. D., Eds.), pp 213-290, New York.
260. Mozhaev, V. V., Heremans, K., Frank, J., Masson, P., and Balny, C. (1996) High pressure effects on protein structure and function, *Proteins 24*, 81-91.
261. Chalikian, T., and Breslauer, K. (1996) On volume changes accompanying conformational transitions of biopolymers, *Biopolymers 39*, 619-626.

262. Royer, C. A. (2002) Revisiting volume changes in pressure-induced protein unfolding, *Biochim Biophys Acta 1595*, 201-209.
263. Mitra, L., Rouget, J. B., Garcia-Moreno, B., Royer, C. A., and Winter, R. (2008) Towards a Quantitative Understanding of Protein Hydration and Volumetric Properties, *Chemphyschem*.
264. Rosgen, J., and Hinz, H. J. (2000) Response functions of proteins, *Biophys Chem 83*, 61-71.
265. Silva, J. L., Foguel, D., and Royer, C. A. (2001) Pressure provides new insights into protein folding, dynamics and structure, *Trends Biochem Sci 26*, 612-618.
266. Mitra, L., Smolin, N., Ravindra, R., Royer, C., and Winter, R. (2006) Pressure perturbation calorimetric studies of the solvation properties and the thermal unfolding of proteins in solution--experiments and theoretical interpretation, *Phys Chem Chem Phys 8*, 1249-1265.
267. Ravindra, R., and Winter, R. (2004) Pressure perturbation calorimetry: a new technique provides surprising results on the effects of co-solvents on protein solvation and unfolding behaviour, *Chemphyschem 5*, 566-571.
268. Baldwin, R. L., and Muller, N. (1992) Relation between the convergence temperatures Th* and Ts* in protein unfolding, *Proc Natl Acad Sci U S A 89*, 7110-7113.
269. Murphy, K. P., Privalov, P. L., and Gill, S. J. (1990) Common features of protein unfolding and dissolution of hydrophobic compounds, *Science 247*, 559-561.
270. Fu, L., and Freire, E. (1992) On the origin of the enthalpy and entropy convergence temperatures in protein folding, *Proc Natl Acad Sci U S A 89*, 9335-9338.
271. Kleywegt, G. J., and Jones, T. A. (1994) Detection, delineation, measurement and display of cavities in macromolecular structures, *Acta crystallographica 50*, 178-185.
272. Kleywegt, G. J., Zou, J. Y., Kjeldgaard, M., and Jones, T. A. (2001) in *International Tables for Crystallography*, pp 353-356,366-367.

VITA
Katrina L. Schweiker

Education:
Penn State College of Medicine, Hershey, PA (2004-2009)
PhD – Integrative Biosceinces/Chemical Biology option
University of Minnesota, Minneapolis, MN (1999-2004)
BS – Physics

Awards:
NASA Graduate Student Research Fellowship (2006-2007)
NASA Academy Summer Internship Program – Glenn Research Center (2005)
Huck Institutes of the Life Sciences Fellowship (2004-2006)
Graham Endowed Fellowship (2004-2005)

Professional Publications:
Schweiker KL, Zarrine-Asfar A, Davidson AR, Makhtadaze GI. (2007) Computational design of the Fyn SH3 domain with increased stability through optimization of surface charge-charge interactions. *Protein Science* **16:**2694-2702.

Schweiker KL, Makhatadze GI. (2009) Protein stabilization by the rational design of surface charge-charge interactions. *Methods in Molecular Biology* **490:**261-284.

Schweiker KL, Makhatadze GI. (2009) A computational approach for the rational design of stable proteins and enzymes: optimization of surface charge-charge interactions. *Methods in Enzymology* **454:**175-211.

Schweiker KL, Fitz V, Makhatadze GI. (2009) – High temperature convergence of the volume changes upon protein unfolding: Implications for protein hydration. *Proc Natl Acad Sci. USA. In preparation.*

Invited Oral Presentations:
The 21st Annual Gibbs Conference on Biothermodynamics: Schweiker KL, Zarrine-Asfar A, Davidson AR, Makhtadaze GI. "Computational design of the Fyn SH3 domain with increased stability through optimization of surface charge-charge interactions." (2007)

East Tennessee State University Department of Biology Seminar Series: Schweiker KL, Makhatadze GI. "Engineering proteins with enhanced stability: optimization of surface charge-charge interactions." (2007)

Poster Presentations:
The 20th Annual Gibbs Conference on Biothermodynamics: Schweiker KL, Makhatadze GI. "Thermodynamic mechanism of stabilization of ubiquitin variants with surface hydrophobic substitutions." (2006)

The Gordon Research Conference on Biopolymers: Schweiker KL, Zarrine-Asfar A, Davidson AR, Makhtadaze GI. "Computational design of the Fyn SH3 domain with increased stability through optimization of surface charge-charge interactions." (2008)