1-1-2016

Molecular Mechanisms of Protein Thermal Stability

Lucas Sawle
University of Denver

Molecular Mechanisms of Protein Thermal Stability

A Dissertation

Presented to the Faculty

of Natural Sciences and Mathematics

University of Denver

in Partial Fulfillment

of the Requirements for the Degree

of Doctor of Philosophy

by

Lucas Sawle

June 2016

Advisor: Dr. Kingshuk Ghosh

Author: Lucas Sawle
Title: Molecular Mechanisms of Protein Thermal Stability
Advisor: Dr. Kingshuk Ghosh
Degree Date: June 2016

Abstract

Organisms that thrive under extreme conditions, such as high salt concentration, low pH, or high temperature, provide an opportunity to investigate the molecular and cellular strategies these organisms have adapted to survive in their harsh environments. Thermophilic proteins, those extracted from organisms that live at high temperature, maintain their structure and function at much higher temperatures compared to their mesophilic counterparts, found in organisms that live near room temperature. Thermophilic and mesophilic homolog protein pairs have identical functionality, and show a high degree of structural and sequential similarity, but differ significantly in their response to high temperature. Addressing the principles of enhanced stability and structural resilience to high temperatures environments is important in furthering our understanding of protein folding and stability, and can be quite useful for protein engineering in industrial and biomedical arenas. Furthermore, understanding temperature dependent protein stability can provide valuable insights into aging and certain diseases.

This work will present the observations from multiple large-scale studies that show meaningful general principles that can be a potential mechanism for thermophilic adaptation. First, from the analysis of the largest data set of thermodynamic data, the roles of reduced thermodynamic parameters upon unfolding, and their association with the unfolded state are discussed. Next, from a first-principle polymer physics model, the contribution from electrostatic interactions are shown to reduce the dimensionality of the unfolded state in thermophilic proteins. Finally,

as a result of long time scale molecular dynamics simulations, electrostatic interactions are shown to be the key contributor in the stability of the folded state in thermal stable proteins. The combined results indicate that thermophilic proteins modify their amino acid content to increase the amount of charged side chains to utilize an adaptive strategy of enhancing favorable electrostatic energies. Molecular effects of protein mutations are observed in experimental measurements of protein thermodynamic values and enzymatic activity. However, modified proteins can also be quantitatively linked to cellular health and fitness. The consequences of modified thermodynamic traits seen in thermophilic proteins to the growth rate of several organisms will be discussed.

Table of Contents

List of Tables v
List of Figures vi

1 Introduction **1**
1.1 Factors enhancing thermal stability 2
1.2 Goals of this project 12

2 Thermodynamic Study **15**
2.1 Model 19
2.2 Results 20
2.2.1 Ideal mesophilic and thermophilic proteins have different thermodynamic properties 21
2.2.2 Specific enthalpy and entropy changes at the convergence temperature are lower in thermophiles on average than mesophiles 23
2.2.3 Maximal stability free energy is higher in thermophiles than mesophiles 25
2.2.4 Reduction in folding entropy (per amino acid) is responsible for high melting temperature 26
2.2.5 Homologous protein pairs reveal similar trend 28
2.3 Discussion 30

3 Modeling the Unfolded State **34**
3.1 Polymer physics model 37
3.2 Results 42
3.2.1 Model captures all-atom simulation results 42
3.2.2 Thermophilic proteins have more compact denatured states than their mesophilic orthologs 48
3.2.3 Additional comparison with homolog structures shows a similar trend 53
3.2.4 Thermophilic protein sequences have more segregated charge distribution 54
3.3 Conclusion 56

4 Convergence of All-Atom Biomolecular Simulations **58**
4.1 Materials and Methods . . . 63
4.1.1 Computational Setup . . . 63
4.1.2 Clustering . . . 64
4.1.3 Convergence inspection protocol . . . 64
4.2 Results and Discussion . . . 66
4.2.1 Convergence of the cluster probability distribution can be highly demanding . . . 66
4.2.2 Comparison of observables at different timescales using Cold Shock Protein Implicit solvent simulation . . . 73
4.2.3 Shortcomings of CCE convergence criteria . . . 75
4.2.4 Long simulations show self-consistency is not equivalent to "true" convergence . . . 76
4.3 Conclusions . . . 80

5 Elevated Temperature Simulations **82**
5.1 Computational Setup . . . 82
5.2 Analysis . . . 84
5.3 Results . . . 86
5.3.1 ACP . . . 86
5.3.2 CheA . . . 86
5.3.3 CheW . . . 87
5.3.4 CheY . . . 87
5.3.5 CSP . . . 88
5.3.6 HGCS . . . 88
5.3.7 HPr . . . 89
5.3.8 NusB . . . 89
5.3.9 PaiA . . . 90
5.3.10 RNaseH . . . 90
5.3.11 RNaseP . . . 91
5.3.12 Sigma . . . 91
5.3.13 Thioredoxin (Oxidized) . . . 92
5.3.14 Thioredoxin (Reduced) . . . 92
5.4 Conclusions . . . 92

6 Native State Modeling **93**
6.1 Methods . . . 95
6.1.1 Selection of protein pairs . . . 95
6.1.2 Computational setup . . . 99
6.1.3 Convergence inspection . . . 100
6.1.4 Representative atoms for the charged groups, and the calculation of their interaction probabilities . . . 101
6.1.5 Calculation of protein interior dielectric constant . . . 102

6.1.6 Calculation of average RMSF . . . 102
6.2 Results and Discussion . . . 103
6.2.1 Attractive ionic networks are better connected in thermophiles 103
6.2.2 Thermophilic proteins have more favorable electrostatic interaction energies in their native state . . . 105
6.2.3 Role of solvation . . . 111
6.2.4 Insights gleaned from the sequence comparison between orthologous proteins . . . 114
6.2.5 Thermophilic proteins do not necessarily have suppressed fluctuations in the native state . . . 119
6.2.6 Presence of hydrogen bonds . . . 121
6.3 Conclusion . . . 124

7 Organismal growth rate calculation and comparison with experiment 125
7.1 Model . . . 126
7.2 Results . . . 127

8 Concluding Remarks 131

Bibliography 133

Appendices 161

A Convergence 161

B DelPhi Benchmarking 169

List of Tables

2.1 Mean values of thermodynamic parameters normalized by chain length 24

5.1 Experimental melting temperature data available for the 13 homolog pairs in the simulation study 83

6.1 Homologous mesophile–thermophile protein pairs used in native state study 96
6.2 Electrostatic energies calculated from structural ensembles taken from simulation trajectories 109
6.3 Calculation of electrostatic energies using strictly the PDB structure for each system 110
6.4 Average solvation energy for each charged group found from combinatoric construction of pentamers 113
6.5 Net charge and calculated SCD per protein 115
6.6 Average hydrogen bond content calculated from MD simulations . . . 123

7.1 Fitted parameters from proteome analysis 128

A.1 Decorrelation times for simulated systems 163

B.1 Contributions to the electrostatic free energy from the two ion model system 171
B.2 Contributions to the electrostatic free energy from the four ion model system 173
B.3 Electrostatic free energies for a five ion model predicted from site potentials 175
B.4 Comparison of electrostatic free energy contributions between all-atom partial charge and ionic group representation atom libraries 180

List of Figures

1.1 Free energy curves shifted in response to differing mechanisms 3

2.1 Changes in enthalpy, entropy, and specific heat upon unfolding for master data set of 116 different proteins 21
2.2 Changes in enthalpy, entropy, and specific heat upon unfolding for subset of mesophilic data 22
2.3 Changes in enthalpy, entropy, and specific heat upon unfolding for subset of thermophilic data 22
2.4 Plot of folding free energy for ideal thermophilic and mesophilic proteins, and resultant stability curves from substitutions of thermophilic thermodynamic parameters into mesophilic free energy expression . . 25
2.5 Direct comparison of all pairwise thermophile–mesophile pairs shown as the difference in change of thermodynamic parameter versus difference in melting temperature 28
2.6 Direct comparison of homologous thermophile–mesophile pairings shown as the difference in change of thermodynamic parameter versus difference in melting temperature 30

3.1 List of 30 sequences used to benchmark theoretical model 42
3.2 Comparison of R_g values from theory to those from simulation in the presence and absence of solution salt 44
3.3 Average distances $\langle\langle R_{ij}\rangle\rangle$ versus $|i-j|$ sequence separartion 45
3.4 Average distances $\langle\langle R_{ij}\rangle\rangle$ from theory versus sequence separation $|i-j|$, and compared against simulation results 46
3.5 Results of analyzing ortholog set of 540 mesophile–thermophile pairs . 51
3.6 Results of analyzing the reduced set of 383 mesophile–thermophile orthologs 52
3.7 Results of analyzing a set of 55 orthologous mesophile–thermophile structural domain pairs 54
3.8 Results of analyzing the reduced set of 46 mesophile–thermophile structural domain orthologs 55
3.9 Charge segregation comparison between metrics ζ and SCD 56

4.1 Results of convergence tests for the CheW explicit solvent simulation at7μs 67
4.2 Results of Block Covariance Overlap Method (BCOM) analysis of CheW 7μs trajectory 67
4.3 Probabilities of the top 20 most occupied clusters as a function of simulation time for the 7μs simulation of CheW 68
4.4 Results of convergence test for the CheA simulation after 1μs 70
4.5 Results of convergence test for the CheA explicit simulation at 2μs . . 71
4.6 Results of convergence test for the CSP implicit simulation at 2.5μs . 72
4.7 Convergence tests for a trajectory of 6μs of CSP 72
4.8 Comparison of probability in amino acid contacts from multiple time scale simulations of CSP 74
4.9 Direct comparison of contact probabilities from shorter time-scale simulations to those of the longest available trajectory 74
4.10 Covariance overlap of the 20μs implicit trajectory of CSP 76
4.11 Convergence criteria inspection of 2μs simulation of BPTI 77
4.12 Multiple self-consistent regions in the 1ms BPTI simulation 78
4.13 Three different SCC regions from the simulation of CSP in explicit solvent 80

5.1 Results of elevated temperature simulations of ACP 86
5.2 Results of elevated temperature simulations of CheA 86
5.3 Results of elevated temperature simulations of CheW 87
5.4 Results of elevated temperature simulations of CheY 87
5.5 Results of elevated temperature simulations of CSP 88
5.6 Results of elevated temperature simulations of HGCS 88
5.7 Results of elevated temperature simulations of HPr 89
5.8 Results of elevated temperature simulations of NusB 89
5.9 Results of elevated temperature simulations of PaiA 90
5.10 Results of elevated temperature simulations of RNaseH 90
5.11 Results of elevated temperature simulations of RNaseP 91
5.12 Results of elevated temperature simulations of Anti-σ 91
5.13 Results of elevated temperature simulations of oxidized Thioredoxin . 92
5.14 Results of elevated temperature simulations of reduced Thioredoxin . 92

6.1 Attractive ζ (edges per vertex) as a function of cut-off probability . . 106
6.2 Attractive ζ (edges per vertex) as a function of cut-off probability . . 107
6.3 Cumulative $\langle RMSF \rangle$ as a function of simulation time for thermophile-mesophile pairs 120

7.1 Predicted growth rate versus experimentally determined growth rates for mesophilic and thermophilic organisms 130

A.1 Dependence of convergence criteria on the choice of clustering cutoff . 161

A.2 Probabilities of most occupied clusters as a function of simulation time for the 7μs simulation of CheW at varying cluster d_c 162
A.3 Results of Block Covariance Overlap Method (BCOM) analysis of CheA 1μs (top) and 2μs (bottom) trajectories 164
A.4 Results of BCOM analysis of implicit solvated CSP 2.5 and 6μs simulations 165
A.5 Results of BCOM analysis for multiple SCC regions of the 1ms simulation of BPTI 166
A.6 Representative states for the two most occupied clusters from the BPTI simulation 167
A.7 Results of Block Covariance Overlap Method (BCOM) analysis of explicit CSP simulation 168

B.1 Two ion model used to calculate electrostatic free energies 171
B.2 Four ion model used to calculate electrostatic free energies 173
B.3 The five ion model used for the calculation of site potentials 174
B.4 Dual angle view of the CSP four ion model 176
B.5 Direct comparison of Generalized Born polarization energy versus DelPhi calculated polarization energy 178

Chapter 1

Introduction

Organisms that thrive under extreme conditions, such as high salt concentration, small or large pH levels, or high temperatures, provide an opportunity to investigate the molecular and cellular strategies these organisms have adapted to survive in their harsh environments. Thermophilic proteins, those extracted from organisms that live at high temperature, denature at much higher temperatures compared to their mesophilic counterparts, found in organisms that live at, or near room temperature. However, a complete understanding of the molecular mechanisms responsible for enhanced thermal stability is lacking. Thermophilic and mesophilic homolog protein pairs have identical functionality, and show a high degree of structural and sequential similarity, but differ significantly in their response to high temperature. Thermophilic proteins maintain their structure, and therefore their function and activity at high temperatures, and identifying and understanding the factors that contribute to the stability under extreme conditions has been a long standing problem within the protein community. Addressing the principles of enhanced stability and intrinsic resistance to high temperatures environments is not only important in furthering our understanding of protein folding, stability, and evolution, but has strong appeal in

industrial and biotechnological arenas where engineered enzymes could have a multitude of applications.[1–3] Furthermore, understanding temperature dependent protein stability can provide valuable insights into proteostasis, the complex and coordinated process of regulating protein expression, stability, and degradation to maintain cellular health. Aging and disease (cystic fibrosis, Alzheimer's, Parkinson's, and Huntington's disease) have been attributed to disruptions in the proteostasis network,[4–7] and protein thermal stability is a key component in understanding these perturbations.

1.1 Factors enhancing thermal stability

A number of strategies have been proposed in an attempt to explain enhanced stability in thermophilic proteins, and currently, there is a lack of consensus in any single proposed mechanism.

Thermodynamic measurements

A thermodynamic understanding of thermophilic protein stability demands knowledge of the free energy change of unfolding. Experimental measurements provide a wealth of information in the determination of thermodynamic properties, such as changes in enthalpy (ΔH), entropy (ΔS), and specific heat (ΔC_p) upon protein unfolding, and provide direct, unambiguous validation that thermophilic proteins indeed have a higher denaturing temperature than mesophiles. Analysis of gathered thermodynamic information highlights the role each of these parameters contributes to the temperature dependent unfolding free energy, i.e. the protein stability curve.

Modifications to the individual thermodynamic contributions (ΔH, ΔS, and ΔC_p) of the free energy have consequences to the overall stability and denaturing temperature. With respect to their mesophilic counterparts, increased melting tem-

perature in thermophiles can be achieved by i) increasing the enthalpy change ΔH resulting in an upshift of the free energy stability curve, ii) a reduction in the curvature by decreasing ΔC_p to broaden the stability curve, iii) a lateral shift toward elevated temperatures, or a combination of these strategies.[8,9] Figure 1.1 demonstrates the changes to stability curves resulting from these different mechanisms. Previous stud-

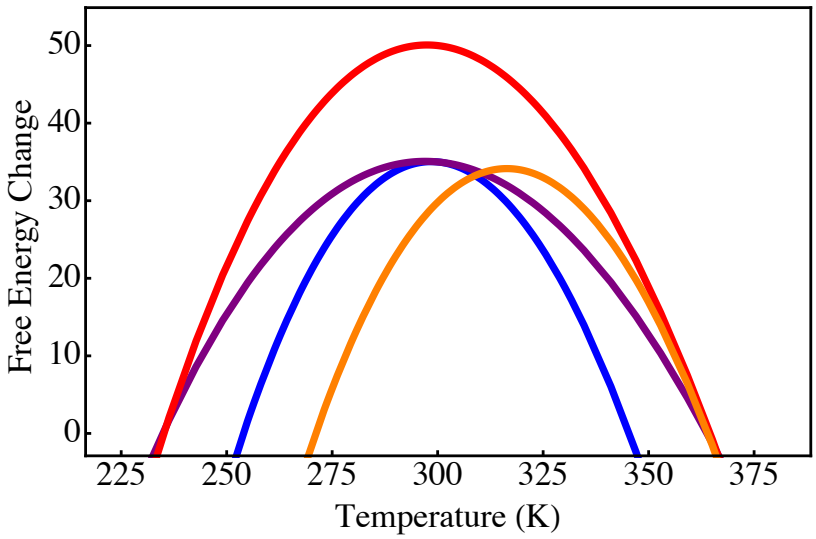

Figure 1.1: Stability curves shifted from a generalized mesophilic thermodynamic parameters (blue curve). An increased melting temperature can be achieved by an upshift (red), broadening (purple), or lateral shift (orange) of the free energy curve.

ies of thermodynamic information showcased that a reduced change in specific heat (ΔC_p) in thermophilic systems broadened the free energy curve, leading to elevated denaturing temperatures, but kept the location and magnitude of the stability maximum unchanged[10–17] (purple curve in Figure 1.1), and direct comparison between several thermophilic and mesophilic homologues support this observation.[13] Other work has shown that an increased enthalpy change (upshift of the free energy curve) in thermophiles may be responsible for heightened thermal stability.[18]

An issue that arises in trying to establish the key players in adaptation is the lack of large scale studies. Particularly, most studies conducted in comparing thermodynamic values between mesophile and thermophiles tend to focus on pairs or

small families worth of information. In a large scale study conducted by this group, analysis showed that while a reduced specific heat change is seen in a large data set, reduced changes in enthalpy and entropy play a key role in thermophilic adaptation.[19] Specifically, the observed reduced entropy of folding imparts the largest positive shift in stability curve when compared to the other thermodynamic parameters. Details of this work will be described in subsequent chapters. While a reduced specific heat change can be linked to a reduced change in solvent accessible surface area by empirical means,[20] and therefore linked to smaller unfolded states in thermophilic proteins, a reduced enthalpy and entropy change from unfolding also points to residual structure and the role of the unfolded state in thermal adaptation.

Although insightful, the studies do not offer many molecular insights. Thermodynamic properties are treated as indirect evidence, while the molecular underpinnings remain elusive. Below, the contributions to enhanced thermal stability from previously proposed structural and biophysical mechanisms are discussed.

Hydrophobicity

The hydrophobic effect (the observed aggregation of nonpolar amino acids in the protein core) is one of the main driving forces in all protein folding,[21,22] but what role does this effect have in protein thermostability? Analysis of 20 genomes, 12 from mesophiles and eight from thermophiles, have shown that proteins from thermophilic organisms posses a significant increase in hydrophobic (nonpolar) amino acids.[23] Furthermore, from a study of 373 structurally well-aligned homologous pairs, 80% of thermophiles displayed higher hydrophobicity than their mesophilic counterpart, and a more refined set of 102 homologs showed 71% of thermophiles had more favorable hydrophobicity.[24] But does the enrichment of hydrophobic residues in the protein necessarily enhance thermophiles to resist elevated temperatures? Theoretically, it

has been shown that even a small increase in the percentage of hydrophobic amino acids presented in the sequence predicts an increase in the overall thermal stability.[25] Experimentally, modification of the hydrophobic content around the protein core has shown to extensively effect stability,[26–30] and selected mutations at the protein surface are stabilizing only if accompanied with a hydrophobic interaction.[31–33] Mutation of a single, unpaired nonpolar residue on the protein surface to a charged residue increased the stability.[34] Recent work from Matsuura et al. with the CutA1 protein demonstrated the effectiveness of hydrophobic substitutions to increase stability, as the double mutation to nonpolar amino acids raised the melting temperature 28°C.[35] In addition, simulation studies have demonstrated that protein stability is degraded by mutations that remove a hydrophobic amino acid from the protein core,[36] and via high temperature modeling, denature easily compared to the wild type.[37]

Hydrogen bonding

Historically, hydrogen bonding and the hydrophobic effect were thought to be the most important contributions to folded state stability of all proteins.[21,22,38–42] The contributions of the hydrophobic effect to the high temperature resistance seen in thermophiles has been discussed, but leaves the question: do thermophilic proteins benefit in any way from hydrogen bonding? The work of Vogt and coworkers have shown by analyzing 16 thermophilic–mesophilic homolog families that in 80% of thermophiles, increases in hydrogen bond content correlated with increased thermal stability.[1,43] In pairwise homologous structure comparisons, the thermophile contained more hydrogen bonds,[44,45] and via the results of molecular dynamics simulations of a homolog pair, the thermophile formed not just more protein-protein hydrogen bonds, but also displayed an affinity to form these bonds with the model solvent.[46] Mutational studies also display the correlation of removed hydrogen bonds and loss of stability.[39,47–49]

The presence of additional hydrogen bonds help maintain the native states of thermophilic systems during high temperature simulations, while the mesophilic homolog denatured.[50,51]

However, in contradiction to the aforementioned studies, and in agreement with a pairwise report,[52] the results of a structural study compromised of 25 protein mesophile and thermophile homolog families demonstrated no difference in the number of hydrogen bonds between the homologous groups, and no tendency for the bond count to be effected by temperature.[53] The hydrogen bond count from simulation studies of homolog pairs are also seen as equivalent,[54,55] leaving the role of hydrogen bond contribution towards increased thermal stability a controversial subject.

Electrostatic contributions

While hydrophobicity and hydrogen bonding have been historical arguments toward protein stability, mounting evidence for electrostatic interactions, specifically those between charged amino acids, point to the dominant contribution for enhanced thermal stability. Starting with purely amino acid sequence content analysis, an increase in the number of charged amino acids was found in thermophilic proteins by comparison of mesophilic and thermophilic genomes,[23] and by in-depth statistical analysis of over 125 sequence pairs from homologous groups.[56] The consistent finding from multiple structure comparison studies show an increased number of ion pairs (salt bridges) and larger interacting ion networks in the thermophilic proteins compared to mesophilic.[24,43,44,53,57–65] The mechanisms observed by researchers with respect to the thermophilic protein folded state is to increase the favorable charge-charge interactions by optimizing spatial placement of charged amino acids on the protein surface,[66–71] and alleviate the unfavorable interactions by charge reversal or negation.[72–76]

The above structural studies illuminate the strategy of optimization in charge–charge interactions, either pairwise or within larger networks, by spatial preference or in relieving repulsive interactions. Simulation studies allow these static assessments to be expanded to include dynamical information. Many simulations have been conducted amongst homologous mesophilic-thermophilic pairs, and at elevated temperatures, to demonstrate that heightened thermal tolerance can be attributed to more favorable charge-charge interactions.[51,54,55,77–81]

A larger abundance of charged amino acids does not equate to an automatic enhancement of stability. Surprisingly, ionic interactions appear to make little contribution towards stability, or even destabilize the protein native state at room temperature.[62,82] The formation of a salt bridge is favorable for the folded protein if the desolvation energy penalty is surpassed by the favorable contributions due to the interaction of the interacting amino acids. The work of Elcock has proposed a way to reconcile the destabilization of ion pairs with their abundance in thermophiles.[83] Specifically, the small contributions to stability of ion pairs is found from the large dehydration penalty incurred from bringing separated charges in the unfolded state together to form salt bridges. The coulombic interaction of an ion pair is obviously favorable, but the energy is not enough to offset the large desolvation penalty incurred upon folding. Due to the lowering of the dielectric constant of water at high temperature, the energy loss from desolvation of the charged residues will be minimized, thus encouraging the formation of ionic pairs.[13,63,77,78,84,85]

Enhanced dielectric

An enrichment of charged side chains in thermophilic proteins is no guarantee to increased stability, due to the larger desolvation penalty upon folding from the elevated charge count. A potential strategy adapted by thermophilic proteins is an

enhanced interior dielectric constant, reducing this desolvation penalty from an increased count of ionic side chains content discussed above. Quantifying structural dynamics provides information about electrostatic free energies. Specifically, from the Frölich–Kirkwood theory of dielectrics, and utilizing the magnitude of dipole fluctuations that arise from protein dynamics, a dielectric constant can be estimated for the protein system from molecular dynamics simulations.[86,87] Dielectrics calculated in this manner are a direct measure of protein polarizability. Dominy and colleagues have shown that by raising their dielectric constant, thermophiles reduce the dehydration penalty incurred upon folding by reducing the dielectric constant discrepancy .[87] Furthermore, the dielectric imbalance that exists between the protein interior and the solvent medium can be alleviated at high temperatures, where the dielectric constant of the water solvent will be reduced, relieving some of the larger dehydration penalty.[83] These measurements, combined with the flexibility studies mentioned below, demonstrate the comparative differences in structural dynamics between homologous pairs of thermophilic and mesophilic proteins.

Native state flexibility

Although the static structures between mesophilic and thermophilic homologous proteins share a high degree of structural similarity, the collective motions and flexibility encoded in these structures must be different for thermophiles to remain active at high temperatures. Motivated by observations that i) thermophilic enzymatic activity is optimal near the natural environment of the organism at high temperatures, and relatively inactive at low temperatures,[88] and ii) function is directly related to structural dynamics,[89,90] a long-standing view is that at room temperature, thermophilic proteins display more mechanical rigidity, and show less activity than their mesophilic homolog, but at each organisms respective physiological tempera-

ture, enzymatic activity is equivalent.[91–97] Enhancement of rigidity by introduction of proline residues, removal of glycines, or the introduction of disulfide bridges have been shown to increase the stability of proteins.[98–101] Early hydrogen exchange experiments have suggested that thermophilic proteins are indeed less flexible at room temperatures compared to their mesophilic counterparts.[94,102] Low protease susceptibility and slower unfolding rate[103] in thermophiles also tend to support this view. However, subsequent experimental investigations to test conformational flexibility of thermophilic and mesophilic homologs have shown otherwise. Hydrogen exchange experiments show sufficient conformational flexibility at room temperature in thermophilic rubredoxin[104,105] and RNaseH.[106] Furthermore, investigations of α-amylase with neutron scattering have shown the thermophilic enzyme is characterized by a higher structural flexibility.[107,108] The work of Querol et al.[109] gathered data for 195 single point mutations linked to greater thermostability and found, by use of crystallographic B-factors, that greater rigidity is not a good indicator of thermal stability.

While experimental studies of reduced flexibility in thermophiles remain inconclusive, computational studies have also remained indecisive. Rigidity theory based analysis from the program FIRST[110] has shown that thermophilic rubredoxin is structurally more rigid than its mesophilic counterpart.[111] Furthermore, room temperature molecular dynamics simulations of homologous protein pairs verify the lack of flexibility in the thermophile with respect to the mesophile.[80,112] A separate analysis based on constraint network analysis has also shown the role of rigidity in thermostability of different mutants from Lipase.[113] However, as clearly articulated by Karshikoff et al., rigidity is a characteristic of a frozen structure and neglects structural fluctuations.[114] Molecular dynamics (MD) simulations provide the ability to capture such fluctuations and dynamics, although at a much shorter time scale than protein domain dynamics probed by hydrogen exchange. MD simulations of protein native states at room

temperature have concluded that thermophiles are not necessarily associated with suppressed fluctuations.[50,54,81,115,116]

Loop depletion

Secondary structure loop regions are often overlooked, but serve an important part in proteins.[117] From proteomic analysis of 20 unique genomes, it has been observed that thermophiles have a shorter sequence than their mesophilic homolog by removal of amino acids in these loop regions, suggesting that by reduction of molecular fluctuation in the folded and unfolded states, thermal resilience is enhanced by adapting a strategy of entropic stabilization.[118,119] Validity for this notion has been seen by many groups whose work showed that by increasing the length of loop regions, stability is reduced,[44,120–123] and by cleaving loop region residues, increased stability is observed.[124] Furthermore, high temperature simulations of BPTI have shown the loop areas act as "targets of attack" for unfolding.[125]

Combination of factors

No individual mechanism or strategy is solely responsible for the increased thermal stability. Instead, thermophilic proteins have adapted by optimizing a combination of the discussed mechanisms. Direct comparison of homologous Cold Shock proteins (CSP) shows a high level of sequence similarity (only 12 of 67 amino acids differ), but a large difference in denaturing temperature (23.3°C). A double mutation in the mesophilic CSP to the corresponding thermophilic amino acids accounted for a 21°C increase in melting temperature by i) alleviation of unfavorable charge-charge interactions, and ii) a charge reversal to optimize an ionic network allowing the formation of a localized rigid cluster more resistant to increased temperature.[13,74,76]

Experimental evolution work of Shamoo et al. display the utilization of several adaptive strategies to enhance protein stability for organism fitness.[126–128] Briefly, the weak-link method allows the study of adaptation of a single gene within a bacterial population. In the weak-link approach, survival of a genetically modified thermophilic organism at elevated temperatures is directly linked to changes in the essential enzyme adenylate kinase (ADK) derived from the mesophile. ADK function is essential for adenylate homeostasis and energy metabolism. The thermophile (native growth temperature range of $48 - 72°C$) expressed the mesophilic ADK, but was unable to grow at temperature higher than 55°C due to high temperature inactivation of the mesophilic ADK. To increase the temperature growth range of the modified organism to that of the native thermophile, evolution of a large population was conducted by a slow temperature increase of 55 to 70°C over 30 days, and population samples were drawn and functional intermediates of the inserted mesophilic ADK were observed. The first mutation was a glutamine to arginine mutation (Q199R) that displayed a marginal increase in the thermal stability of ADK, but dramatically improved enzyme function. The mutation and introduction of a charged amino acid permitted the formation of a small ionic network on the surface of ADK, allowing the rigidification of the C-terminus. The Q199R mutation acted as a progenitor for five double mutations that arose simultaneously within the population at higher incubation temperatures increasing the denaturing temperature above the Q199R mutation. Of these five, four aided in further rigidification of the C-terminus by either electrostatic interaction expanding the aforementioned surface ion network, introduction of hydrogen bonds not present in the unmutated structure, or by increasing hydrophobicity in a cavity made with the terminus structure.

1.2 Goals of this project

In the remainder of this thesis, contributions to understanding the strategies employed in thermophilic adaptation will be presented. First, in an attempt to settle the contradicting viewpoints seen in earlier and smaller studies, the largest data set of experimentally determined thermodynamic data was curated, and from the calculated average parameters for changes in specific heat (ΔC_p), enthalpy (ΔH), and entropy (ΔS), thermophilic proteins show a smaller change in enthalpy and entropy upon unfolding.[19] A reduced change in heat capacity was also observed, in accord with earlier studies of thermodynamic parameters. However, the reduced ΔC_p term has a minimal negative effect on stability, and the smaller change in enthalpy destabilized our *ideal* system substantially. The reduced entropy change observed in thermophilic *ideal* system offsets the destabilizing effects of smaller ΔC_p and ΔH, and amounts to a net large increase in stability. Furthermore, the reduction in the thermodynamic parameters allowed the logical inference to the role of residual structure in the unfolded state.

Second, from the large-scale thermodynamic parameter study, the role of denatured state residual structure was highlighted as a potential strategy for enhanced stability. However, an actual physical mechanism remained undefinable. Using methods from polymer physics theory and first principle approaches, we modeled the unfolding state of a large set of thermophilic and mesophilic homolog protein pairs. By considering the effects of ionic amino acid patterning within the sequence alone, it was observed that nearly 70% of the thermophilic sequences in the 540 pair dataset had a smaller, more compact unfolded state,[129] reinforcing the roles of residual structure, and enrichment and optimization of charged amino acids as strategies for enhanced thermal stability.

Third, the role of native state dynamics was examined by utilizing long time scale, room temperature molecular dynamics simulations for 13 pairs of thermophilic and mesophilic homologous globular proteins. The majority of previous studies have relied on only a select protein pair per study, not allowing any universal mechanisms to be seen. By including many pairs, different competing strategies, and the relative importance of each were highlighted. However, the demands of a study of this scale have raised additional concerns detailed below.

With the exception of only a few simulation studies,[51,54,81] previous work utilized relatively short simulation trajectories with very little attention paid to quality of sampling measures, leaving the reported observables unreliable, or artifacts of inadequate sampling. Therefore, we will first introduce a novel technique to test the convergence of molecular dynamics simulations that allow the sampled trajectories to be claimed as adequately sampled and self-consistent.[130]

Next, due to the lack of experimental unfolding information, only half of the 13 homolog pairs have a documented melting temperature. The pairs were constructed by assuming proteins from thermophilic organisms would intrinsically possess heightened stability at elevated temperatures. Therefore, we conducted a pairwise study of all of the system pairs in high temperature environments ($\approx$ 400K) with the goal of observing resistance to unfolding in the thermophilic protein, and structural instability in the mesophilic protein. This study helped to verify the consistency of the computational force field away from room temperature simulations, and demonstrated that even in the absence of experimental melting temperatures, the thermophilic proteins in this study displayed structural resilience at elevated temperature, while their respective mesophilic proteins denatured.

Now with the ability to adequately study verified mesophilic-thermophilic homolog pairs, the minimum simulation time per system was 1μs, with the in-depth

convergence inspection applied to each simulation, causing some protein pairs to be simulated up to 6μs to simply model the folded state. From the multitude of information, a plethora of significant conclusions were found. Specifically, thermophilic proteins have adapted a strategy to construct well connected attractive electrostatic networks within their folded state. Furthermore, the electrostatic driven adaptation employed more subtle mechanisms, such as a more favorable electrostatic interaction energy, or a reduced dehydration penalty via an increased folded state dielectric constant.

Finally, the effects of protein thermal stability on cellular health and fitness will be discussed. Specifically, the effects to cellular growth rate by alterations to protein thermodynamics. From thermodynamic free energy equations, the free energy distribution of the entire proteome was calculated to estimate the growth rate of several mesophilic and thermophilic organisms. This model showed quantitative agreement with the experimental thermal growth data, but only upon proper selection of protein thermodynamic parameters.[19]

From these studies, we hoped to gain further insights into the mechanisms and strategies of thermophilic adaptation not seen in previous works. Having used larger datasets in these projects, we hoped to remove the ambiguous and often conflicted results from earlier studies. Also, having incorporated a quality of sampling metric into our molecular dynamics study, the resultant observables will be explicit, reliable, and reproducible.

Chapter 2

Thermodynamic Study

Studying different mechanisms by which proteins increase, or decrease stability can teach us fundamentals of protein thermodynamics, and assist in the design of new enzymes with desired stability. The vast majority of biophysical studies have been directed towards understanding the origin of enhanced stability in proteins under conditions of high temperature.[105,131–135] The unusual tolerance to high temperature has raised a few interesting questions: i) Does any systematic principle exist that proteins may utilize to withstand such high temperatures? ii) Do significant alteration of the average enthalpy, entropy, or specific heat, or a combination of these effect thermal stability? iii) Does high melting temperature also imply high maximal stability free energy? Researchers have tried to address these questions by directly comparing different homologues of proteins extracted from mesophilic and thermophilic organisms .[13–15,18,131,132,136–138]

It is not clear which of these mechanisms proteins adopt, or whether they adopt different mechanisms simultaneously to a different extent. It has been widely demonstrated that reduced ΔC_p is primarily responsible for increased stability[13–16] by broadening the stability curve, but keeping the location and magnitude of the maximum of

the stability curve unchanged. Direct comparison between several thermophilic and mesophilic homologues support this observation.[13] However, other studies suggest otherwise by demonstrating little dependence between melting temperature and ΔC_p of unfolding.[18,139]

These different and often contradictory studies make it very hard to identify the principle behind increased stability. Conclusions are very specific to proteins, and there is no systematic study[136] using a large data set of different protein families to explain increased stability. Conclusions are often conflicting depending on the list of proteins studied, and this is mainly due to two reasons, i) lack of systematic analysis of a large dataset as previous studies were mostly restricted to smaller sets of proteins, and ii) lack of proper decoupling of different driving forces. The latter point originates from the fact that enthalpy and entropy changes are significantly temperature dependent, due to non-zero ΔC_p arising from hydrophobic effect, both enthalpic and entropic changes are temperature dependent in the following manner:[140,141]

$$\Delta H(T) = \Delta H(T_h) + \Delta C_p(T - T_h) \tag{2.1}$$

$$\Delta S(T) = \Delta S(T_s) + \Delta C_p \log\left(\frac{T}{T_s}\right) \tag{2.2}$$

where, T_h and T_s are two reference temperatures. This also raises the natural question: at what temperature one should compute these quantities to compare ? From equations 2.1 and 2.2, it is evident that both enthalpy and entropy change have different hydrophobic contribution (due to non-zero ΔC_p) at different temperatures. For example, total change in entropy has a contribution due to configurational entropy of the protein chain as well as mixing entropy of amino acids, increasing the difficulty to isolate and study the role of different driving forces separately.

In order to significantly decouple the hydrophobic effect from conformational entropy and purely enthalpic contributions (polar and van der Waals), one should compute $\Delta H(T)$ and $\Delta S(T)$ at temperatures where the hydrophobic effect is minimal. Extensive studies dating back to Privalov,[142,143] Baldwin,[144] Doig[145] and Robertson-Murphy[140] showed the existence of a temperature where enthalpy and entropy (per residue) for many different proteins converge. These *convergence* temperatures are now believed to be the temperature where hydrophobic effects are zero, and the contribution to enthalpy is purely due to van der Waals or polar interactions. Similarly, at the convergence temperature entropy is primarily conformational in origin as hydrophobic contribution is minimal. Hydrocarbon-transfer experiments by Baldwin demonstrated the existence of a similar convergence temperature, where the transfer entropy is zero and was very close to the protein convergence temperature.[144] The finding strongly suggests, at this temperature the protein chain conformational entropy is significantly decoupled from the hydrophobic entropy. Extensive work of Robertson-Murphy gives the most recent and reliable estimate of these convergence temperatures based on the largest set of proteins: for entropy $T_s = 385$K and for enthalpy, $T_h = 373.5$K. The original work of Robertson-Murphy,[140] and previous work[141] showed $\Delta H(N)$ and $\Delta S(N)$ can be very well approximated as a linear function of chain length N when computed at 373.5K and 385K, respectively. In fact, the correlation coefficient between the changes in enthalpy and entropy versus chain length were the highest at these two temperatures.[140] Based on this evidence, and several other studies,[142,146,147] it is clear, at these temperatures sequence effects are minimal, and the major contribution to enthalpy and entropy is due to the polymeric nature of the protein alone.[140,148] Therefore, the slope and intercept of the linear dependence of these properties with respect to protein chain length gives us an average estimate of changes in enthalpy, conformational entropy and specific heat of a pro-

tein upon folding purely based on the protein chain length N. This defines an *Ideal Thermal Protein*, and can serve as a first estimate for the folding free energy when all other protein information, except chain length, is absent.[141]

Here, we will compute and compare changes in entropy and enthalpy of thermophilic and mesophilic proteins at $T_s = 385$K and $T_h = 373.5$K, respectively, for two reasons: i) at these convergence temperatures, the hydrophobic effect can be separated from enthalpy and conformation entropy, thus different driving forces are maximally decoupled, and ii) it provides a common reference temperature to compare different proteins. This has not yet been explored, and is in striking contrast to common practice where thermodynamic properties are computed at the melting temperature for comparison. However, it is not guaranteed that at the protein melting temperature, the hydrophobic effect is separated from conformational entropy.

Below, from the largest protein thermodynamic unfolding data set constructed to date, almost doubling the current largest set,[140] the analysis to derive *Ideal Thermal Protein* parameter values will be outlined. Next, the data set will be classified into two classes i) based on melting temperatures, and ii) a smaller set of homologous proteins based on the optimal growth temperature of the organism from which proteins were selected. From this classification scheme, new *Ideal Thermal Protein* parameter values were found for thermophilic (with high melting temperature) and mesophilic (with low melting temperature) proteins. Statistical analysis was conducted on the distribution of entropy, enthalpy and specific heat changes (per amino acid) between these two classes of proteins, revealing a new thermodynamic principle that proteins on average may adopt to withstand high temperature. In general, lower entropic loss upon folding may be responsible for enhanced stability in thermophilic proteins. Finally, based on these new parameters, the entire proteome of different organisms is modeled and compared against experimental thermal growth rate data.

2.1 Model

Analysis was conducted on a significantly large data set (116 proteins) of thermodynamic data, almost doubling the number of proteins (63) from the earlier work of Robertson-Murphy.[140] In this analysis, the differences in entropy ΔS, enthalpy ΔH, and specific-heat ΔC_p between the folded and unfolded states were calculated. Thus, ΔS is defined as $S_u - S_f$, where S_u is the entropy of the unfolded state, and S_f is the entropy of the folded state. This definition has been used for enthalpy, specific heat and, free energy change throughout.

As outlined above, it is most instructive to compute ΔH at 373.5K and ΔS at 385K to maximally decouple the effects of sequence and other driving forces. Furthermore, at these two temperatures, $T_h = 373.5\text{K}$ and $T_s = 385\text{K}$, the change in enthalpy and entropy show a strong linear chain length dependence.[140] These quantities can be computed from enthalpy and entropy values reported at melting temperature (T_m) using:

$$\Delta H(373.5\text{K}) = \Delta H(T_m) + \Delta C_p(373.5\text{K} - T_m) \tag{2.3}$$

$$\Delta S(385\text{K}) = \Delta S(T_m) + \Delta C_p \log\left(\frac{385\text{K}}{T_m}\right) \tag{2.4}$$

Specific heat was assumed independent of temperature.[140,142] All proteins that were either multimeric, or non-two state folder were removed from the original list of Robertson-Murphy analysis. Several new two-state folders for which changes in enthalpy, entropy, and specific heat could be found were added to the data set. Furthermore, the search was limited to only monomeric proteins under conditions where reversible transition was observed, and closer to isoelectric point to further decouple

electrostatic contributions. This analysis also extended the range of applicability by including proteins with bigger chain length, and a wider range of melting temperatures than originally considered in the Robertson Murphy analysis.[140]

However, since this set does not distinguish moderate (mesophilic) and high melting temperature (thermophilic) proteins which may have different thermodynamic properties, the set of 116 proteins was further decomposed into two sets by defining a cut-off melting temperature (T_c). Upon this classification, the analysis described above was revisited. The optimal cut-off temperature, T_c, was determined by minimizing the least square error of fitting chain length dependent linear equation for all three thermodynamic quantities ($\Delta H, \Delta S$, and ΔC_p) separately when proteins were subdivided in two classes: i) proteins with melting temperature higher than T_c, and ii) proteins with melting temperature below T_c. This method yielded a choice of $T_c = 341$K where the least square error was minimized for enthalpy, entropy, and specific heat change independently. Therefore, we determine $T_c = 341$K as the melting temperature below which we identify proteins as mesophilic, and above which they are termed thermophilic for this analysis. Also note, the least square fitting error of these quantities with chain length reduced significantly when proteins were subdivided in two families compared to the undivided set.

2.2 Results

The linear correlation of the overall set was found to be slightly lower than the original analysis due to Robertson-Murphy,[140] but the average thermodynamic parameters change only slightly. Figure 2.1 shows results of this analysis. The slopes and intercepts of these different thermodynamic quantities against chain length determines the properties of an *Ideal Thermal Protein.*[141]

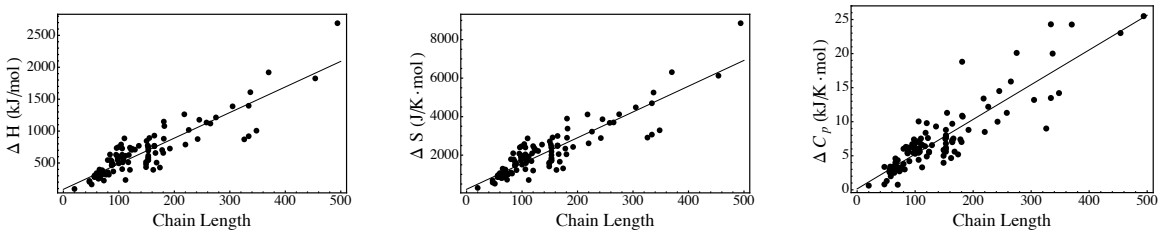

Figure 2.1: Change in enthalpy, entropy and specific heat upon unfolding for the 116 constituent proteins in the master set. Solid lines show the best linear fit, and filled circles show data.

2.2.1 Ideal mesophilic and thermophilic proteins have different thermodynamic properties

It is likely that proteins with higher melting temperature may have evolved with a different set of thermodynamic rules than their mesophilic counterparts. Therefore, it is natural to think that the *Ideal Thermal Protein* parameter values for thermophiles would be significantly different than their mesophilic counterparts. Several indirect experimental results support this. For example, comparison of thermodynamic properties between a thermophilic and mesophilic homologue based on native state hydrogen exchange by Hollien and Marqusee[106] demonstrated that the increased stability is not a result of localized effect, but rather distributed throughout. A proportional increase in stability for all residues results in an overall enhanced stability indicating the possible role of simple properties such as chain length in stability determination, and the importance of studying average parameters. Motivated by this, the master set was divided into two classes as mentioned in the methods section. Based on the analysis outlined above, good correlation between thermodynamic parameters and chain length for mesophilic proteins was found, see Figure 2.2. This set of 59 proteins, out of a total 116, have a melting temperatures below $341K$. Therefore, based on the new analysis of this smaller, modified data set of proteins, new parameters for

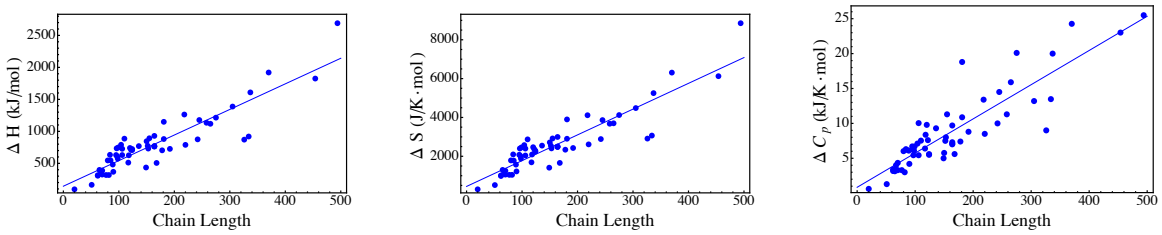

Figure 2.2: Change in enthalpy, entropy and specific heat upon unfolding for the 59 proteins assigned to the mesophilic subset. Solid lines show the best linear fit, and filled circles show data.

Ideal Mesophilic Protein were calculated as

$$\begin{aligned} \Delta H_{373.5\mathrm{K}}(N) &= 4.0N + 143 \text{ kJ/mol} \\ \Delta S_{385\mathrm{K}}(N) &= 13.27N + 448 \text{ J/mol} \\ \Delta C_p(N) &= 0.049N + 0.85 \text{ kJ/mol} \end{aligned} \tag{2.5}$$

The remaining 57 proteins with melting temperatures at, or above $341K$ were classified as thermophilic proteins, and similar analysis was carried out on this set, Fig. 2.3. And, again good correlation between thermodynamic parameters and protein

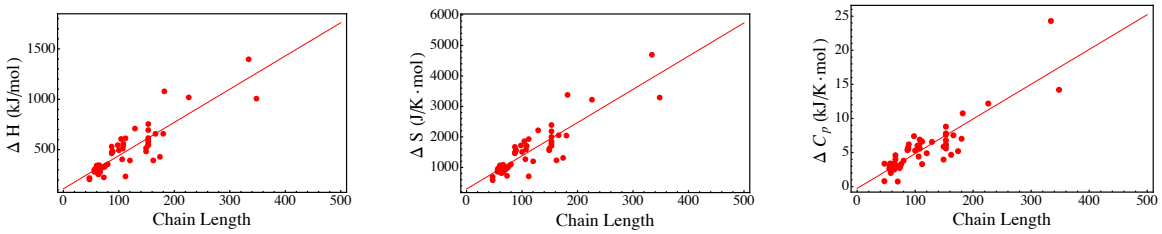

Figure 2.3: Change in enthalpy, entropy and specific heat upon unfolding for the 57 thermophilic proteins. Solid lines show the best linear fit, and filled circles show data.

chain lengths was found for the thermophilic protein set. The new thermodynamic

parameters obtained define an *Ideal Thermophilic Protein* as

$$\begin{aligned} \Delta H_{373.5\mathrm{K}}(N) &= 3.30N + 112 \text{ kJ/mol} \\ \Delta S_{385\mathrm{K}}(N) &= 10.90N + 291 \text{ J/mol} \\ \Delta C_p(N) &= 0.051N - 0.26 \text{ kJ/mol} \end{aligned} \tag{2.6}$$

In the absence of any other information but the protein chain length, thermophilic and mesophilic protein thermodynamic parameters can be estimated based on equation sets 2.5 and 2.6.

2.2.2 Specific enthalpy and entropy changes at the convergence temperature are lower in thermophiles on average than mesophiles

Based on the slopes and intercepts reported above, it is clear that changes in entropy and enthalpy upon folding are lower in thermophilic proteins than in mesophilic. A slightly different, and more rigorous approach will be presented here to establish this fact. Changes in thermodynamic parameters per amino acid for mesophilic and thermophilic sets were calculated, and distributions of $\Delta H(373.5\mathrm{K})/N$, $\Delta S(385\mathrm{K})/N$, and $\Delta C_p/N$ were constructed per set. The means of these quantities were compared between mesophiles and thermophiles. The results are summarized in Table 2.1. From the reported values, it is clear that thermophilic proteins, on average, have a lower value for enthalpic, entropic, and specific heat changes per residue. A two sample t-test was performed on these distributions, from which, it can be claimed that thermophiles have a lower change in entropy per amino acid than mesophiles

Mean Values of Thermodynamic Parameters

	Mesophile	Thermophile	p–value
$\frac{\Delta H(373.5\text{K})}{N}(\frac{kJ}{mol \cdot res})$	5.18	4.52	0.001
$\frac{\Delta S(385\text{K})}{N}(\frac{J}{K \cdot mol \cdot res})$	16.97	14.12	0.00002
$\frac{\Delta C_p}{N}(\frac{kJ}{K \cdot mol \cdot res})$	0.055	0.049	0.008
$\frac{\Delta H(T_m)}{N}(\frac{kJ}{mol \cdot res})$	2.72	3.65	5.3×10^{-8}
$\frac{\Delta S(T_m)}{N}(\frac{J}{K \cdot mol \cdot res})$	8.26	10.22	0.00003
$T_S(K)$	281.6	284.2	0.3
$\frac{\Delta G(T_S)}{N}(\frac{kJ}{mol \cdot res})$	0.21	0.40	3.3×10^{-6}

Table 2.1: Mean values of thermodynamic parameters normalized by chain length. When comparing at convergence temperatures, the changes in enthalpy and entropy are less in thermophiles than in mesophiles. When comparing enthalpy and entropy changes at melting temperatures, thermophilic changes are greater. T_S is the temperature of maximal stability that appears to be similar between thermophiles and mesophiles on average. However, free energy ($\Delta G(T_S)$) at the temperature of maximal stability (per amino acid) is significantly favorable in thermophiles than mesophiles. p–value is a measure of confidence from testing the difference of two means.

with a p-value of 0.00002. Furthermore, changes in enthalpy per amino acid are lower for thermophiles than mesophiles with a p-value of 0.001, whereas the specific heat p-value is 0.008.

However, when enthalpy and entropy changes were computed at their respective melting temperatures, a reverse effect was observed. Based on values reported in the Table 2.1, thermophiles have a higher change in specific entropy and enthalpy than mesophiles when computed at the melting temperature, in agreement with an earlier study.[18]

2.2.3 Maximal stability free energy is higher in thermophiles than mesophiles

Based on two set representation of mesophilic and thermophilic proteins, the average temperature of maximal stability (T_s), and the free energy change at this temperature ($\Delta G(T_s)$) were calcualted. The temperature of maximal stability is similar between thermophiles and mesophiles, but the free energy change at this temperature is almost twice as favorable in thermophiles compared to mesophiles (see Table 2.1), in accordance with previous studies.[18,149] This is also evident from the comparison of stability curves of the *Ideal Mesophilic* (blue) and *Ideal Thermophilic* (red) proteins in Figure 2.4.

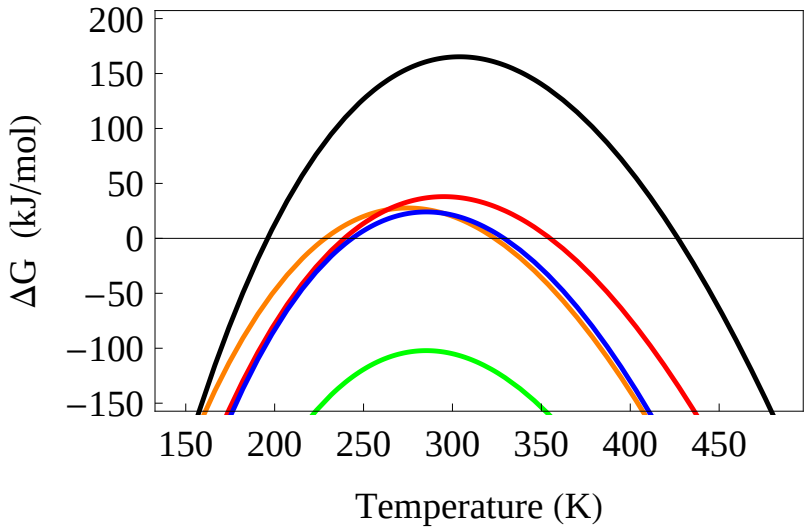

Figure 2.4: Plot of folding free energy of ideal thermophilic (red) protein showing a taller, broader, and right shifted curve compared to ideal mesophilic (blue) protein. Insertion of thermophilic specific heat change into the ideal mesophilic free energy slightly reduces melting temperature (orange), while insertion of thermophilic enthalpy change dramatically reduces stability (green). The substitution of thermophilic entropy into the mesophilic free energy change (black) is the only parameter showing increased stability and melting temperature.

2.2.4 Reduction in folding entropy (per amino acid) is responsible for high melting temperature

Using thermodynamic parameters reported in equation sets 2.5 and 2.6, the temperature dependent free energy $\Delta G(T)$ (in kJ/mol) of an *Ideal Thermophilic* and *Ideal Mesophilic* protein were calculated as

$$\begin{aligned}\Delta G(N,\ T)_{meso} = \quad & (4.0N + 143) + (0.049N + 0.85)(T - 373.5\text{K}) \\ & -T\tfrac{(13.27N+448)}{1000} - T(0.049N + 0.85)\log\left(\tfrac{T}{385\text{K}}\right)\end{aligned} \tag{2.7}$$

$$\begin{aligned}\Delta G(N,\ T)_{thermo} = \quad & (3.30N + 112) + (0.051N - 0.26)(T - 373.5\text{K}) \\ & -T\tfrac{(10.90N+291)}{1000} - T(0.051N - 0.26)\log\left(\tfrac{T}{385\text{K}}\right)\end{aligned} \tag{2.8}$$

Using the average chain length ($N = 136$ of the master 116 protein data set) in the equations above, and plotting both ΔG_{meso} and ΔG_{thermo} as a function of temperature (see Figure 2.4), we make the following points:

- the ideal thermophilic curve (red) is shifted upward, broader, and laterally to the right compared to the ideal mesophilic curve (blue).
- thermophilic melting temperature is 25 Kelvin higher than the mesophilic T_m, 355K versus 330K, these temperature ranges are approximately in the same range as reported earlier.[149]
- cold denaturing temperature of thermophiles (239K) is slightly colder than that of mesophiles (244K) as previously suggested.[150]

- replacing the mesophilic specific heat change (ΔC_p in equation set 2.5) in expression 2.7 with the corresponding thermophilic ΔC_p from equation set 2.6, keeping ΔH and ΔS intact, an almost negligible change in free energy with a slight destabilizing effect (orange curve in Fig. 2.4) is observed.

- substitution of exclusively the mesophilic enthalpy change by thermophilic value (ΔH in equation set 2.6), keeping ΔS and ΔC_p for mesophilic proteins intact, a strong destabilizing effect is seen (green curve in Fig. 2.4).

- replacing only the mesophilic entropy by thermophilic ΔS, a significant upshift in the stability curve is seen, increasing the melting temperature to a much higher value (black curve, Fig. 2.4).

Thus, varying all three parameters individually, it has been clearly demonstrated an *Ideal Thermophilic Protein* gains a high melting temperature by reducing the entropic loss upon folding.

Next, a direct comparison all possible pairwise combinations of thermophilic to mesophilic thermodynamic parameters against the pairings' respective melting temperatures. The above analysis gave, after decomposition, 59 mesophiles and 57 thermophiles leading to a total of 3363 comparable pairs. Also from the above analysis, all thermophilic melting temperatures are greater than mesophilic, so the difference in melting temperatures of the i^{th} thermophile to the j^{th} mesophile, gives $\Delta T_m = [T_m]_i - [T_m]_j > 0$. Computing the difference in specific entropy for the same i^{th} thermophile to j^{th} mesophile pair gives

$$\Delta\left(\frac{\Delta S(385\text{K})}{N}\right) = \left(\frac{\Delta S(385\text{K})}{N}\right)_i - \left(\frac{\Delta S(385\text{K})}{N}\right)_j \tag{2.9}$$

Plotting ΔT_m versus $\Delta\left(\frac{\Delta S(385\text{K})}{N}\right)$ for all 3363 possible i,j pairwise combinations

of mesophile-thermophile data, 70% of the pairs have a higher melting temperature associated with differences in entropic changes (per amino acid) that are less than zero. This implies thermophilic $\Delta S/N$ is less than mesophilic $\Delta S/N$ for 70% of possible pairings (see Figure 2.5). Similar calculations for specific changes in enthalpy and specific heat show in the specific enthalpy change plot 65% of pairings have thermophilic enthalpic gain (per amino acid) lower than their mesophilic counterpart. Finally, 61% of thermophilic $\Delta C_p/N$ is less than its mesophilic counterpart.

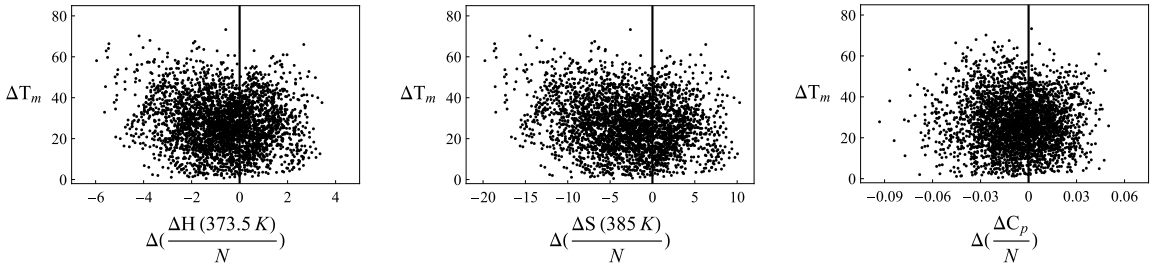

Figure 2.5: Direct comparison of thermophiles to mesophiles shown as the difference in change of thermodynamic parameter (equation 2.9) versus difference in melting temperature per pairing. Points left of Y–axis signify thermophilic proteins having a smaller change in a respective thermodynamic quantity. For enthalpy, 65% of thermophiles are less; for entropy, 70%; for specific heat, 61%.

Analysis based on each protein pair shows a significant correlation between increased melting temperature and reduced entropy change upon folding further justifying the claim based on *Ideal* protein parameter comparison.

2.2.5 Homologous protein pairs reveal similar trend

Thus far, all analysis has been based on a mean field approach on a data set that classified proteins into thermophile and mesophile based on their respective melting temperature. An alternate approach is considered by constraining the data set to consist of only pairs, or groupings, of mesophile and thermophile homologs where the classification is based on the organism from which the proteins have been extracted.

Several proteins derived from thermophilic organisms have been studied[13,131,132,137] and compared with their homologs extracted from mesophilic organisms. The protein pairs considered here show either high structural similarity, or high sequence identity, and have been published as a relevant grouping based on their homology. Our 10 groupings of homologs includes six thermophilic-mesophilic pairs, and four groupings of at least four proteins, giving a total of 16 thermophiles and 17 mesophiles. The majority of the data shared at least 40% sequence identity with its homologs and a backbone RMSD of less than 2Å within its pairing or group. The following groups were published as being homologous, but were either below this criteria, or unavailable to calculate: The pairing of MGMT-AdaC showed only 20% sequence identity, but had a calculated backbone RMSD of 1.9Å.[151] The S16 pair had 33% identity, but a calculation of RMSD was unavailable due to a lack of structural data.[152] Calculation of sequence identity and RMSD was unavailable for Phycocyanin.[153] Within the SH3 Domain-Containing group of 8 proteins, certain pairs (e.g., Sac7d and Fyn) had RMSD as high as 9.9Å.[18,139]

As in the previous section, each thermophilic protein to all other mesophilic protein within the same group was compared to compute differences in changes in specific entropy $\left(\Delta\left(\frac{\Delta S(385\text{K})}{N}\right)\right)$, enthalpy $\left(\Delta\left(\frac{\Delta H(373\text{K})}{N}\right)\right)$, specific heat $\left(\Delta\left(\frac{\Delta C_p)}{N}\right)\right)$, and change in melting temperature (ΔT_m). By directly comparing these quantities only within groupings of homologs, a high percentage (79%) of thermophilic entropy changes (per amino acid) are less than the mesophilic entropy changes (per amino acid) in the same group. Also, 68% of changes in enthalpy (per amino acid), and 75% of changes in specific heat (per amino acid) are less in thermophiles compared to their mesophilic counterparts (see Figure 2.6). Thus, direct comparison of normalized thermodynamic data shows thermophiles have a high propensity to have reduced entropic, enthalpic, and specific heat change when compared to their mesophilic counterparts.

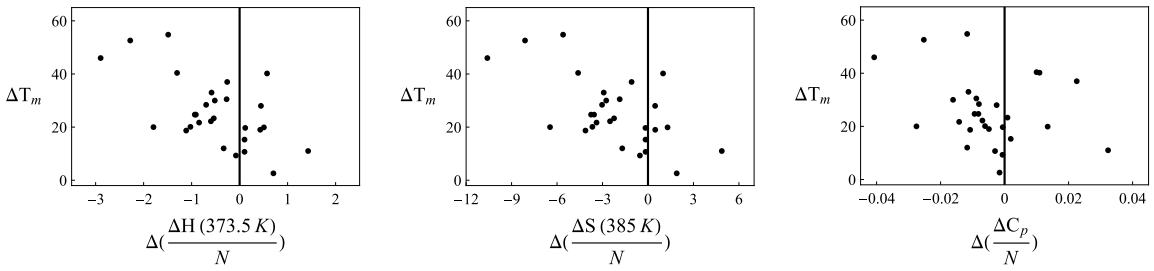

Figure 2.6: Direct comparison of homologous thermophiles and mesophiles shown as the difference in change of thermodynamic parameter (equation 2.9) versus difference in melting temperature per pairing. Points left of Y–axis signify thermophilic proteins having a smaller change in a respective thermodynamic quantity. For enthalpy, 68% of thermophiles are less; for entropy, 79%; for specific heat, 75%.

2.3 Discussion

The largest data set of thermodynamic properties was constructed and analyzed for thermophilic and mesophilic proteins. Unlike any other previous study, enthalpy (per amino acid) and entropy (per amino acid) change were computed at the convergence temperature to compare thermophiles and mesophiles. The rationale was to separate hydrophobic effects from other driving forces, i.e. decouple conformational entropy and solvation entropy. With this definition, in general and with high statistical confidence, the gain in enthalpy (per amino acid) upon folding is less in thermophiles than mesophiles, and is a destabilizing effect. However, with respect to conformational entropy, thermophiles sacrifice less entropy, on average, upon unfolding than mesophiles, and therefore overcompensate the destabilizing effect due to a reduced enthalpy change. The reduced entropy change is responsible for the extra stability in thermophiles. It may appear that this finding is in conflict with studies that predict higher specific entropy and enthalpy in thermophiles than mesophiles.[18] The apparent contradiction is due to the temperatures at which thermodynamic quantities were calculated. When entropy and enthalpy were computed at the melting temperature, the results of Kumar et al[18] were recovered. However, as stated earlier,

the entropy computed at the melting temperature is not purely conformational due to the presence of solvation entropy. The temperatures for maximal stability were found similar among thermophiles and mesophiles, while the folding free energy (per amino acid) at these temperatures is significantly more favorable in thermophiles than mesophiles, in agreement with previous studies.[18,149]

Therefore, computation at the convergence temperature depicting lower folding entropy (conformational) and enthalpy associated with thermophiles is the novelty of this analysis. Furthermore, the combined findings of reduced entropy loss and lowered enthalpy gain alludes to the possibility that thermophiles may retain partial contacts in their denatured state. Presence of strong hydrophobic interactions, disulfide bonds,[154,155] or electrostatic interactions[156–158] may be responsible for such residual structure in the denatured state. This would explain lowered gain in enthalpy upon folding as there are already existing favorable interactions in the denatured state. However, due to the presence of these native/non-native contacts, the conformational entropy in the denatured state will also be lowered. This can explain the reduced loss in folding entropy (conformational) observed in thermophilic proteins. This is consistent with experimental studies indicating reduction of unfolded state entropy responsible for enhanced stability.[100,159,160] The existence of residual structure in the unfolded state,[15,16] or compact denatured state[152] in thermophiles are strong evidence for reduced entropy change upon folding responsible for elevated melting temperatures. Wallgren et al[152] showed thermophilic Ribosomal protein S16 has a more compact denatured state than its mesophilic homolog. Similar observations based on radius measurement, has been made by Licata[159] in connection to their studies on Taq DNA polymerase. Comparative investigation of two thermophilic α–amylases showed a more compact unfolded state when denatured thermally than when denatured chemically. Also, the amylase with the higher thermal stability showed a more

compact state than the amylase with lower melting temperature.[160] Residual structure has been observed in thermophilic RNase H, but not in the mesophilic homolog.[15] Existence of residual structure, and native/non-native contacts and local compactness of denatured ensemble has been addressed in several other contexts as well.[157,161–168] The presence of residual structure in the denatured state will be responsible for a lowered solvent accessible surface area compared to a fully unfolded extended state. This would consequently explain a lower ΔC_p in thermophiles, and is consistent with other works in the literature.[13–16] Thus, the finding of reduction of folding entropy is not in contradiction with studies hinting at reduced ΔC_p, but in accordance. It appears calculations based on convergence temperature reconciles all the existing observations by properly extracting the conformational entropy. However, it should be remembered, by lowering ΔC_p alone, keeping $\Delta H(373.5\text{K})$ and $\Delta S(385\text{K})$ intact, would lead to destabilization rather than stabilization of the protein (see orange curve in Figure 2.4).

Another explanation behind the reduced change in folding entropy could be specific amino acid substitutions that lead to reduced entropy in the unfolded state due to different degrees of flexibility associated with these amino acids .[100,169] This is also consistent with the technique of enhancing stability by reducing conformational entropy of the denatured state by adding proline residues in β turns and at other locations in proteins.[170,171] Nemethy and Scheraga quantified possible changes in unfolded chain entropy from amino-acid substitution.[172] The effect of entropy on increased stability may also arise from different degree of compactness in the native structure as a result of different mutations.[173] Substitution of amino acids could change the entropy of the folded state. Concepts of rigidity, compactness, and rotameric states in the native state play an important role in stabilizing thermophilic proteins, and would be related to the entropy change as well. Through computational studies, rubredoxin

was found to be more globally rigid with respect to temperature than its mesophilic counterpart,[111] and thermophilic RNase H was shown to have less backbone flexibility at same temperatures, and less conformational entropy over a large temperature range than its mesophilic homolog.[174]

However, a more microscopic model will be needed to further investigate the quantitative contribution arising from the unfolded and native state entropy difference. Based on the finding at the convergence temperature, both lowered change in entropy and enthalpy is in accordance with several experimental studies that point to residual structure and reduced specific heat change.[13–16,152,159,160] This finding does not contradict different studies. Rather, it reconciles all of them. However, reduced enthalpy and specific heat upon folding has a destabilizing effect that is over compensated by reduced loss in entropy imparting the higher stability in thermophiles. Therefore, the key factor behind increased thermal tolerance is reduction of folding entropy.

Chapter 3

Modeling the Unfolded State

From the analysis of thermodynamic information above, the suggestion of residual structure being a prevalent mechanism in the denatured state of thermophilic proteins has been illuminated. This structural strategy supports the reduction in thermodynamic parameters observed in thermophilic proteins (Table 2.1), but does not speak to the presence any physical underpinnings that would allow residual structure to occur. Here, we will explore the consequences of electrostatic interactions that arise from the distribution of charged amino acids in the protein sequence.

Sequence specificity of all biomolecules is central to biological function. For example, the particular sequence of amino acids results in a unique and exclusive folded structure of a protein responsible for a specific function.[175] This sequence-structure-function dogma has driven the protein science community to focus primarily on the folded states of proteins. However, the unfolded, or disordered, state of a protein remains relatively unexplored in spite of its importance.[162,164,176] The protein disordered state can also exhibit sequence specific properties, that itself hold important biochemical clues. We know a significant fraction of the eukaryotic proteome is comprised of *intrinsically disordered proteins* (IDP).[177] These proteins lack a definite,

well-defined folded structure, as they are disordered, but are important in biological regulation and signaling.[177,178]

The unfolded state is also important to understand the stability of the folded structure. Thermophilic proteins, those primarily extracted from organisms that thrive at very high temperatures, are important systems for understanding protein stability. Thermophilic proteins transition from the folded to the unfolded state at significantly higher temperatures than mesophilic proteins, those from organisms that require moderate temperatures for survival. Several lines of investigation[15,19,159] lead to the intriguing possibility that thermophilic proteins may retain residual structures in their unfolded states. The residual structure in the unfolded state can entail less entropic loss in transition to the folded state, hence higher fold stability.[19,163] A recent study has also highlighted the role of unfolded protein ensemble in determining folding rates.[179] With these developments, it is now timely to ask how do we quantify the protein unfolded state? In particular, we need to understand the relative differences in the unfolded ensemble between two chains having subtle sequence variations.

While molecular simulation studies (MD) have made significant progress in modeling protein folded states, the studies of unfolded proteins are severely limited by the astronomical number of conformations that a protein molecule can adopt.[180] Advanced computer architectures and novel algorithms are being developed to extend molecular simulation times.[181–183] Despite these advances, computation time still limits the number and size of proteins, and the duration of processes that can be studied by MD. Although all-atom Monte-Carlo (MC) simulations are relatively faster, and have been recently employed as an alternate approach to speed up computation,[184,185] MC simulations are also not scalable to large collections of proteins.

Theoretical polymer physics, using mathematical formalisms, describe disordered ensembles efficiently, although approximately, and provide a tempting alternative.

Traditional polymer physics is primarily based on a bead-spring model, where *identical* beads representing monomers are strung together via springs. When the beads are identical, they describe homopolymers. Theoretical formalisms, including variational approaches, have been developed to describe dimensions of homopolymer chains under good solvent conditions using a uniform value of the excluded volume parameter.[186] These models have been extended to describe flexible and semiflexible polyelectrolytes (uniformly charged polymers).[187–191] Several methods have been developed to study simple random sequences of beads representing few monomer types.[192–196] Scaling arguments and theoretical models have been proposed to describe conformations of polyampholytes (polymers with both positive and negative charges).[192,197–200] Traditional polyampholyte models rely solely on charge composition, and fail to distinguish between sequences that have different charge patterning, but same charge composition. However, it is well known that different sequences with same charge composition can exhibit different conformational properties.[185,201] Models for random polymers average over different sequences assuming the disorder is annealed. However, a given protein has a static sequence, hence the disorder is quenched, and the models above are not suitable.

Polymer physics based approaches[193–195,202] have also modeled heteropolymers with *specific* sequences of a few monomer types as toy systems. These models, and other polymer physics based studies, have been successful in inferring general principles in protein folding.[193,194,203–208] However, none of these models have been used to compute unfolded ensemble configurational properties as a function of the sequence decoration. Here, we describe a general formalism to compute configurational properties of a polymer with different types of individual monomers threaded in a particular sequence in the backbone. Sequence specificity is captured by accounting for monomer-pair specific interactions describing i) short-range excluded volume interac-

tion, and ii) long range electrostatics. The excluded volume interaction, represented by a pseudo-potential,[209] effectively accounts for monomer-solvent interaction. This is particularly relevant to model the protein unfolded state where solution condition mimic that of a good solvent. However, many other potentials, with a hard core repulsive term, and a weak attractive component, can also be approximated by a corresponding pseudo-potential.[209,210] These monomer-pair specific excluded volume parameters can be directly used in our model to describe sequence specificity.

Below, we first present our formalism, and subsequently provide a few applications of the model. First, we benchmark our predictions against all-atom Monte Carlo results of Das and Pappu[185] for 30 distinct sequences (see Figure 3.1) each having an equal number of Glutamic Acid (E) and Lysine (K) amino acids, but uniquely distributed. Next, we demonstrate the large scale application of the model by carrying out a relative comparison of the denatured state dimensions between 540 orthologous pairs of thermophilic and mesophilic proteins using sequence charge information only.

3.1 Polymer physics model

The Hamiltonian for a polymer chain in the presence of inter-monomer excluded volume, and electrostatic interaction is given by[209]

$$\begin{aligned}\beta H_t = & - \frac{3}{2l}\int_0^L ds \left(\frac{d\mathbf{R}(s)}{ds}\right)^2 + \int_0^L ds \int_0^s ds' \omega(s,s')\delta[\mathbf{R}(s)-\mathbf{R}(s')] \\ & + \frac{l_b}{l^2}\int_0^L ds \int_0^s ds' q(s)q(s')\frac{\exp(-\kappa|\mathbf{R}(s)-\mathbf{R}(s')|)}{|\mathbf{R}(s)-\mathbf{R}(s')|} \end{aligned} \tag{3.1}$$

where s is the contour length variable in the backbone, $\mathbf{R}(s)$ is the position vector at s, L is the total contour length with l denoting the Kuhn length, T is the temperature, and k_b is Boltzmann's constant with $\beta = 1/(k_bT)$. We follow a coarse-

grained approach where all the atomic details of the amino acid are represented by the physicochemical properties of a single monomeric unit. Excluded volume interaction between two monomers at s and s' is modeled by $\omega(s, s')$. The residue at s carries a charge $q(s)$ (in the units of electron charge). If fully ionized, $q(s) = +1$ for an amino acid with basic side chain, $q_s = -1$ for an acidic amino acid, and zero for neutral residues. Furthermore, $l_b = e^2/(4\pi\epsilon_0\epsilon k_b T)$, with e being the charge of an electron, and ϵ representing the dielectric constant of the medium. For water, $l_b = 7.2$Å. We use Debye-Huckle theory to model the interaction between charges.[211] Debye length is defined as $\lambda_d = \kappa^{-1} = (8\pi l_b c_s)^{-1/2}$, where c_s is the salt concentration.

Variational approach has been used to determine the average size of homopolymer chains in the presence of inter-monomer interactions.[186–188] Within the variational scheme, we map the Hamiltonian (H_t), with *all* interactions, to a renormalized Hamiltonian (H_r) such that

$$\beta H_r = \left(-\frac{3}{2l_r} \int_0^L ds \left(\frac{d\mathbf{R}(s)}{ds} \right)^2 \right) \tag{3.2}$$

where l_r is the renormalized Kuhn length, and is a function of the excluded volume and electrostatic interaction parameters. The ensemble average of some physical observable A, with respect to the total Hamiltonian H_t up to the first order in $(H_r - H_t)$, can be expressed as

$$\langle A \rangle = \langle A \rangle_r + \langle A \rangle_r \langle (H_r - H_t) \rangle_r - \langle A(H_r - H_t) \rangle_r + O[(H_r - H_t)^2] \tag{3.3}$$

where $\langle ... \rangle_r$ denotes averages with respect to the renormalized Hamiltonian H_r, and $\langle ... \rangle$ denotes average over the original Hamiltonian H_t. Averages are computed over all possible configurations, and thus involve functional integrals. The effective Kuhn length l_r can be determined by demanding $\langle A \rangle \approx \langle A \rangle_r$. Therefore, the equation for

the renormalized Kuhn length is given by

$$\langle A\rangle_r\langle(H_r - H_t)\rangle_r = \langle A(H_r - H_t)\rangle_r. \tag{3.4}$$

The choice of A depends on the quantity of interest. When we are interested in determining the average end-to-end distance, $\mathbf{R}_{ee}$, we set $A = [\mathbf{R}_{ee}]^2 = (\mathbf{R}(L) - \mathbf{R}(0))^2$.[186] Although this method was originally pioneered by Edwards-Singh[186] for an excluded volume chain, similar calculations were later carried out to describe charged polymers .[187–190] Subsequent work[191] has shown the same variational approach can be extended to determine the ensemble average distance between any two amino acids s_2 and s_1 by setting $\mathbf{A} = (\mathbf{R}(s_2) - \mathbf{R}(s_1))^2$. This condition determines the renormalized Kuhn length $l_r(s_2, s_1)$ for a specific pair of amino acid residues (s_2, s_1) by demanding

$$\langle(\mathbf{R}(s_2) - \mathbf{R}(s_1))^2(H_r - H_t)\rangle_r = \langle(\mathbf{R}(s_2) - \mathbf{R}(s_1))^2\rangle_r\langle(H_r - H_t)\rangle_r. \tag{3.5}$$

Switching from continuous variables s_2, s_1 to discrete indices i, j for monomers on the protein backbone, the equation above gives an equation for $l_{r,ij}$.

We define $x_{i,j} = l_{r,ij}/l$ as the ratio of the renormalized Kuhn length for amino acid pair (i, j) to the bare Kuhn length l. Equation 3.5 yields an equation for $x_{i,j}$ (see Reference 129 for detailed derivation):

$$x_{i,j}^{3/2}(i - j)(1 - 1/x_{i,j}) = \left(\frac{3}{2\pi}\right)^{3/2}\frac{\Omega_{i,j}}{x_{i,j}} + \frac{1}{9}\frac{4\pi l_b}{l}\frac{2}{(2\pi)^2}Q_{i,j}^{el} \tag{3.6}$$

where, $N \geq i > j \geq 1$, N is the total number of residues in a chain, and $\Omega_{i,j}$ represents

the excluded volume contribution defined as

$$\begin{aligned}\Omega_{i,j} &= \sum_{m=j}^{i}\sum_{n=1}^{j-1}\omega_{m,n}\frac{(m-j)^2}{(m-n)^{5/2}} + \sum_{m=j+1}^{i}\sum_{n=j}^{m-1}\omega_{m,n}(m-n)^{-1/2} \\ &+ \sum_{m=i+1}^{N}\sum_{n=1}^{j-1}\omega_{m,n}\frac{(i-j)^2}{(m-n)^{5/2}} + \sum_{m=i+1}^{N}\sum_{n=j}^{i}\omega_{m,n}\frac{(i-n)^2}{(m-n)^{5/2}} \end{aligned} \tag{3.7}$$

We assume, $m > n$. The electrostatic contribution, $Q^{el}_{i,j}$, is similarly defined as

$$\begin{aligned}Q^{el}_{i,j} &= \sum_{m=j}^{i}\sum_{n=1}^{j-1} q_m q_n A(m,n,x_{i,j},\kappa l)(m-j)^2 \\ &+ \sum_{m=j+1}^{i}\sum_{n=j}^{m-1} q_m q_n A(m,n,x_{i,j},\kappa l)(m-n)^2 \\ &+ \sum_{m=i+1}^{N}\sum_{n=1}^{j-1} q_m q_n A(m,n,x_{i,j},\kappa l)(i-j)^2 \\ &+ \sum_{m=i+1}^{N}\sum_{n=j}^{i} q_m q_n A(m,n,x_{i,j},\kappa l)(i-n)^2 \end{aligned} \tag{3.8}$$

with $A(m,n,x_{i,j},\kappa l)$ as

$$\begin{aligned}A(m,n,x_{i,j},\kappa l) = x_{i,j}^{3/2}\Bigg[&\frac{\pi^{1/2}}{4}\left(\frac{6}{x_{i,j}}\right)^{3/2}\frac{1}{(m-n)^{3/2}} \\ &-\frac{\pi^{1/2}}{2}(\kappa l)^2\left(\frac{6}{x_{i,j}}\right)^{1/2}\frac{1}{(m-n)^{1/2}} \\ &+\frac{\pi}{2}(\kappa l)^3\exp\left(\frac{(\kappa l)^2 x_{i,j}(m-n)}{6}\right) erfc\left(\sqrt{\frac{(\kappa l)^2 x_{i,j}(m-n)}{6}}\right)\Bigg]\end{aligned} \tag{3.9}$$

We now present some limits of this equation. In the limit of zero salt, the electrostatic contribution in equation 3.6 simplifies to

$$Q_{i,j}^{el} = \frac{(6\pi)^{3/2}}{4\pi}\left[\sum_{m=j}^{i}\sum_{n=1}^{j-1} q_m q_n \frac{(m-j)^2}{(m-n)^{3/2}} + \sum_{m=j+1}^{i}\sum_{n=j}^{m-1} q_m q_n (m-n)^{1/2} + \sum_{m=i+1}^{N}\sum_{n=1}^{j-1} q_m q_n \frac{(i-j)^2}{(m-n)^{3/2}} + \sum_{m=i+1}^{N}\sum_{n=j}^{i} q_m q_n \frac{(i-n)^2}{(m-n)^{3/2}}\right] \tag{3.10}$$

In another limit, $i = N$, $j = 1$, $q_i = 1$ and $\omega_{ij} = \omega$, ensuring uniform values of charge and excluded volume parameters, we recover the results for end-to-end distance of a flexible uniformly charged polyelectroylte derived by Muthukumar.[187] However, our formalism differs from those theories[192] that derive configurational properties as a function of compositional quantities. These theories do not account for sequence specificity. Sequence specific theories, such as the one presented here, do not require compositional variables and higher order correction terms.[192] Here, we intend to apply our formalism for proteins that have non-uniform charge decoration, and amino acids that can interact with each other in residue-pair specific manner. Values of x_{ij} obtained from the equations above can be used to compute the distance between any two amino acids i, j as

$$\langle(\mathbf{R}_i - \mathbf{R}_j)^2\rangle = |i-j|l^2 x_{ij} \tag{3.11}$$

Consequently, the radius of gyration (R_g) can be calculated using

$$R_g^2 = \frac{1}{N^2}\sum_{i>j}\langle(\mathbf{R}_i - \mathbf{R}_j)^2\rangle \tag{3.12}$$

Equations 3.6, 3.7, 3.8, and 3.9 are the central results of this work. These equations, along with equation 3.11, completely describe the disordered ensemble for a given sequence specified by the set of $\{\omega_{m,n}\}$ and $\{q_m\}$.

3.2 Results

3.2.1 Model captures all-atom simulation results

Equations 3.8 and 3.9 show chain dimensions can vary with sequence charge decoration by explicitly accounting for charges q_i, q_j (on monomers i, j repectively) and their sequence separation $i-j$. It is easy to see these terms can vary significantly between two sequences, even when the sequences share the same charge composition. Das and Pappu[185] recently constructed 30 sequence variants of Glutamic Acid (E) and Lysine (K) preserving the same composition (25 Es and 25 Ks in a chain of 50 amino acids; see Figure 3.1). They determined the sizes of these different sequences using all-atom Monte Carlo (MC) simulation.[185] We test our model against these simulation results.

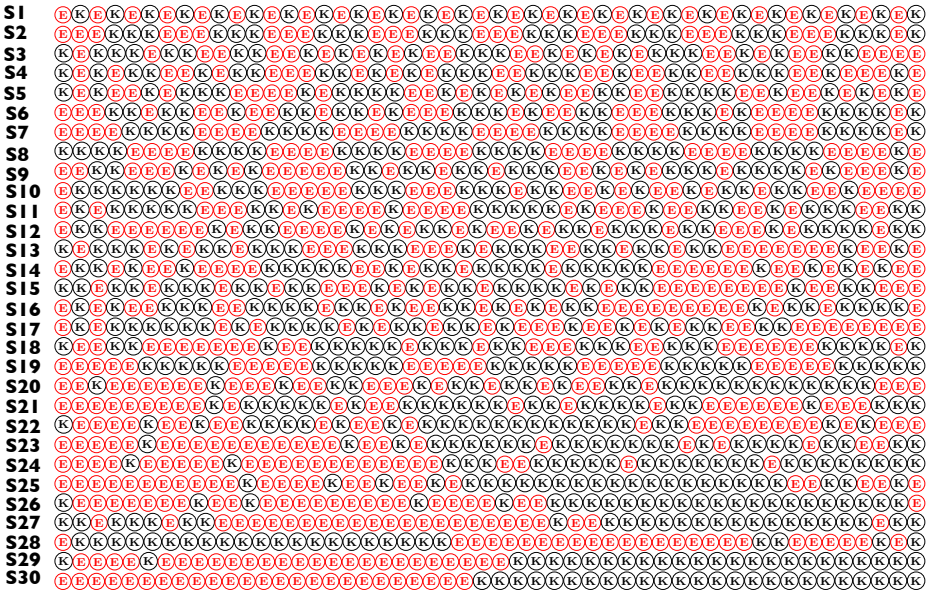

Figure 3.1: List of 30 sequences studied to test theory against all-atom simulation results of Das and Pappu.[185] Red circle with E denotes Glutamic acid, and black circle with K denotes Lysine. Sequence indices are noted on the left.

To predict sequence dependent dimensions, values for the Kuhn length, l, and the excluded volume parameter, $\omega_{m,n}$, are required. From the Flory random coil simulations of Das and Pappu,[185] the Kuhn length was estimated as $l = 5.8\mathring{A}$. To determine a value for $\omega_{m,n}$, the IS limit simulation radius of gyration calculations from Das and Pappu[185] were utilized. The IS limit captures the "Intrinsic Solvation" of the polymer chain in the absence of any electrostatic interactions. These simulation results provide a reference state for a given sequence, and are used to extract the excluded volume parameter. From IS simulations, we obtain $x_{ij,IS}$ defined as the ratio of the internal distances in the IS limit ($\langle(\mathbf{R}_i - \mathbf{R}_j)^2_{IS}\rangle$) to that in the Flory random coil state. Then, we assume excluded volume parameters are independent of the amino acids for a given sequence and choice of residues i, j. This assumption simplifies equation 3.7, leaving only one parameter for the excluded volume. Inserting $x_{ij,IS}$ into equation 3.6, and assigning $Q_{el} = 0$, we determine the numerical value of the mean-field excluded volume parameter. This excluded volume parameter is used to determine x_{ij} in the presence of electrostatics, i.e., non-zero Q_{el}. Here, we compute sequence specific Q_{el} (in equation 3.8 and equation 3.9) by assuming all charged amino acids are fully ionized with $q_i = -1$ for Glutamic Acid and $q_i = +1$ for Lysine. The Debye length is computed using a salt concentration of 15 mM salt, and the Bjerrum length is assumed to be $7.2\mathring{A}$, consistent with the simulation conditions. Although slight variations in the dielectric constant, and degrees of ionization are possible from sequence to sequence, we ignore them for simplicity. Equations 3.11 and 3.12, and the values of x_{ij} determined above are now used to predict the radius of gyration for each sequence. Figure 3.2 shows the comparison between the predicted values of the radius of gyration against the simulated results without any fit parameter. Each point in Figure 3.2 represents a synthetic protein molecule of chain length 50, consisting of 25 positive and 25 negative charged amino acids in different arrangements (see

Figure 3.1). We captured the strong sequence dependence observed in the all-atom simulations with a correlation of $R^2 = 0.9$. Figure 3.2 (right panel) further shows comparison between predicted radius values using simplified equation 3.10 in the zero salt limit against simulated values. This simplified equation also captures ($R^2 = .94$) strong sequence dependence of radius values.

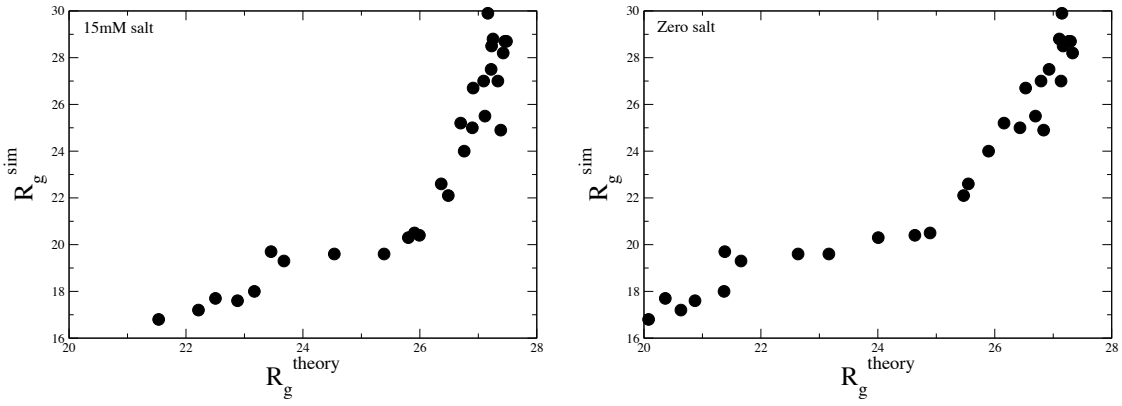

Figure 3.2: (Left) Predicted radius (R_g in Angstroms) from theory (using Debye length for 15mM salt) is plotted in x-axis against simulation in y-axis for each sequence (shown in Figure 3.1). (Right) Predicted radius (R_g in Angstroms) from theory (using zero salt limit equation 3.10) plotted along the x-axis, with simulation along the y-axis for each sequence. Predicted values have a strong correlation ($R^2 = 0.9$ for the left and $R^2 = 0.94$ for the right) with the all-atom simulation. Simulated R_g values were obtained from.[185]

We compute the distance profiles $\langle\langle R_{ij}\rangle\rangle$ as a function of the sequence separation, $|i-j|$ for five illustrative sequences from[185](see Figure 3.3). These sequences contain an equal number of positive and negative charges, but charges are distributed uniquely per sequence (see Figure 3.1). The values of $\langle\langle R_{ij}\rangle\rangle$ were calculated by taking the average of $\langle R_{ij}\rangle$ for different values of i, j for a fixed sequence separation $|i-j|$. Thus the inner $\langle ...\rangle$ denotes averaging over different chain conformations for a given choice of i, j while the outer $\langle ...\rangle$ denotes averaging over all pairings of i, j , but for a given sequence separation $|i-j|$. We notice non-monotonic behavior of distance profiles, consistent with the all-atom simulation (Figure 3.4). The non-monotonic

behavior is significant, both in the simulation and theoretical prediction, for sequences where positive and negative charges are well segregated forming large clusters of similar charges. Das and Pappu have provided a scaling argument[185] to explain undulations in these profiles. For short sequence separations the chain does not experience electrostatic interaction, and behaves like an excluded volume chain before transitioning to an intermediate length scale where the excluded volume interaction is suppressed due to electrostatic attractions. Finally, the intermediate region crosses over to scales where excluded volume behavior between distant chain segments is restored, but at a smaller effective scale. Figure 3.4 shows a comparison between our theoretical prediction and the scaling laws (in solid lines) of Das and Pappu.[185]

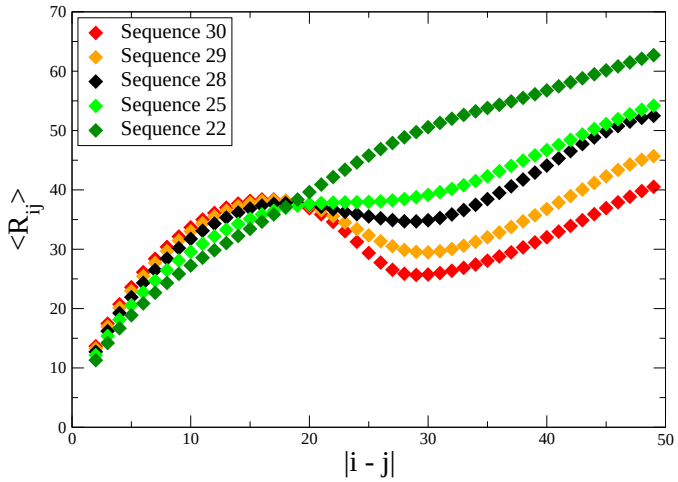

Figure 3.3: Average distances $\langle\langle R_{ij}\rangle\rangle$ are plotted against sequence separation $|i-j|$ for five different sequences 22, 25, 28, 29, and 30. The indexing scheme of these sequences is same as the one shown in Figure 3.1 and are taken from.[185] These sequences were chosen as illustrative examples to show heterogeneous distance scaling and their dependence on different sequences.

The model, without any fit parameter, well predicts the relative sizes between different sequences (Figure 3.2), and internal scaling (Figure 3.3) for a given sequence.

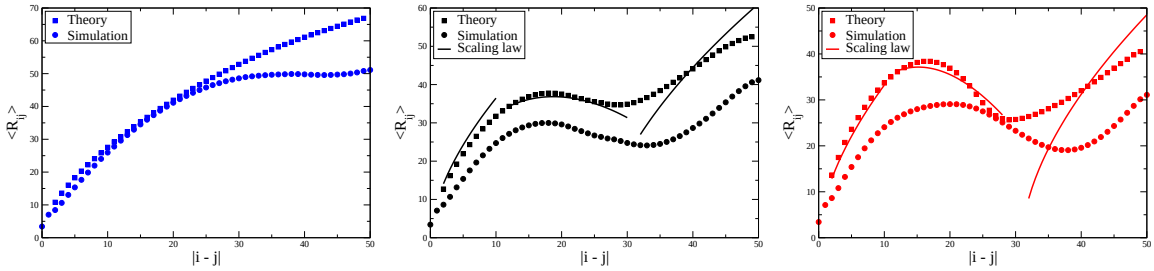

Figure 3.4: Average distances $\langle\langle R_{ij}\rangle\rangle$ from theory (solid squares) are plotted against sequence separation $|i-j|$ and compared against simulation (filled circles) for three different sequences 10 (left), 28 (middle), and 30 (right). Comparison between theoretical prediction and the scaling function of Das and Pappu [185] is shown (solid lines) for sequences 28 and 30 that exhibit non monotonic profile.

However, our model overestimates R_g values for the sequences where simulated values of R_g is low (lower left side in Figure 3.2). These discrepancies are noticed for sequences where positive and negative charges are highly segregated, e.g., sequence 30 (the bottommost sequence in Figure 3.1). For identical sequences, our model also significantly overestimates internal distance profiles when compared to the results of all-atom simulation, although the overall profile is qualitatively similar to the simulation (see Figure 3.4).

Those sequences with highly segregated charges may witness the onset of a small degree of ordering, which can lead to several effects not included in our present model, causing the observed differences. A possible source of error can stem from the dielectric constant employed in the calculation of l_b. Formation of compact structures in the highly segregated sequences would reduce the dielectric constant compared to the dielectric constant of water. Furthermore, these highly segregated sequences may experience the formation of hydrogen bonds observed in the all-atom simulation, but not included in our coarse-grained model. A better agreement with the model's predictions to simulation's results can be obtained by lowering the excluded volume parameter. The present value of the excluded volume parameter, used to generate Figure 3.2, is estimated using the IS limit, which does not include formation of

backbone-sidechain hydrogen bonds. Differences between theoretical prediction and simulation results for the internal scaling profile of sequence 10 at large separation (Figure 3.4) can be attributed to the formation of some ordering, leading to short range interactions not included in the excluded volume parameter.

Variation in the degree of ionization can cause further differences between our predictions and the observed values in the all-atom simulations. In our analysis, salt effects have been modeled using the Debye–Huckle approximation with a Debye length corresponding to 15 mM salt concentration to match simulation conditions. While our theory reasonably captures simulation results at 15mM (Figure 3.2 left panel), we notice i) a slightly better agreement with simulation results using the zero salt limit of our model (Figure 3.2 right panel), and ii) a significant screening at 125 mM, in contrast to simulation for three sequences for which data is available (comparison data not presented). These results suggest the Debye–Huckle approximation, due to its mean-field nature, may not be a suitable approximation at high salt concentration for sequences exhibiting strong charge segregation. Following work of Muthukumar,[210] an advanced model can be envisioned to self-consistently predict the conformation dependent degree of ionization to address this issue. In summary, it is possible to further improve agreement between theory and the all-atom simulation results by taking into account these finer effects with l_b, excluded volume parameters, and degree of ionization as fit parameters. Despite these approximations, the model is capable of capturing observed differences in the configurational properties due to subtle variations in the chain sequence primarily due to long-range coulomb interactions. Thus, the presented model can provide insights to intrinsically disordered proteins where electrostatics strongly influences chain dimension.[212,213] Furthermore, the model may also delineate relative differences between protein sequences that have subtle variations in their charge decorations, as demonstrated below.

3.2.2 Thermophilic proteins have more compact denatured states than their mesophilic orthologs

Using all-atom molecular dynamics to simulate the unfolded state is prohibitively expensive due to sampling problems. Experimental studies are also limited, and mostly rely on indirect measures such as changes in specific heat, which is related to the solvent accessible surface area.[214] A more direct approach would be to compare unfolded state dimensions and configurational properties between a thermophilic and mesophilic proteins. Thus, we need a model, akin to the one presented here, that can capture differences in the denatured state configurational properties arising from subtle differences in the sequence between a thermophilic and a mesophilic protein. Our analytical model is capable of deciphering such differences in the conformational properties due to preferential placement of charges between two similar sequences. Another advantage of our model, being analytical, is its ability to compute configurational properties for many sequences in a high throughput manner. Motivated by these observations, we apply this formalism to evaluate the relative sizes of the denatured states between thermophiles and mesophiles using sequence charge information alone.

As an application of the model, we use slightly simplified version of Equation 3.6 by considering only the end-to-end distance in the zero salt limit. The choice of the zero salt limit is further justified by its performance against all-atom simulation results in the previous section (Figure 3.2 right panel). We set $i = N, j = 1$, and

$\kappa = 0$. The simplified equation for x, ignoring the subscripts, becomes

$$\begin{aligned} x^{3/2}(1-1/x) &= \left(\frac{3}{2\pi}\right)^{3/2} \frac{1}{x}\frac{\omega}{N}\left[\sum_{m=2}^{N}\sum_{n=1}^{m-1}(m-n)^{-1/2}\right] \\ &+ \frac{l_b}{l}\left(\frac{2}{3\pi}\right)^{1/2} \frac{1}{N}\left[\sum_{m=2}^{N}\sum_{n=1}^{m-1} q_m q_n (m-n)^{1/2}\right] \end{aligned} \tag{3.13}$$

We also assumed that the excluded volume parameter is independent of the types of amino acids, i.e., $\omega_{m,n} = \omega$. We further assumed each amino acid to be a Kuhn segment to avoid complication of ambiguous charge assignment. To find an empirically reasonable estimate of ω, we analyzed a set of experimentally determined unfolded state radius of gyration values measured by Hofmann et al.[215] This dataset reported R_g values characterized from different proteins of varied chain length under native conditions. From this radius of gyration dataset we calculate x for each protein by taking the ratio of the end to end distance under denatured condition to that of a Flory random coil. The end to end distance (R_{ee}) is calculated from experimentally measured R_g values by using $R_g^2 = R_{ee}^2/6$. The Flory random coil end-to-end distance ($R_{ee,frc}$) is computed as $R_{ee,frc}^2 = Ll$, where $L = Nb$, b is bond length, and l is the Kuhn length. We assume $b = 0.38$nm and $l = 0.8$ nm[215,216] to determine $R_{ee,FRC}$. From these calculated values of x, using the parameters mentioned above, and the simplified equation 3.13, with the electrostatic term set to zero, we acquired a set of excluded volume parameters (ω). We use the average value, $\langle\omega\rangle = 0.11$, to represent a uniform excluded volume interaction amongst all amino acid pairs during calculation.

A well curated dataset by Fang et al[217] consisting of 540 non-redundant mesophilic–thermophilic protein ortholog sequence pairs was used. In construction of this dataset, the proteins in each pairing were confirmed to contain the same domains, and trans-membrane proteins were omitted, leaving strictly soluble, and highly probable true

ortholog pairs.[217] Because the simplified expression for x requires only sequence positions (m, n) and side chain charges at those positions (q_m, q_n), the sequence of amino acids were translated from single letter codes to their respective side chain ionic charges. Letters **D** and **E** representing negatively charged side chains became -1, positively charged side chains **R** and **K** were converted to +1, and all remaining amino acids were replaced with a 0, leaving pairs of sequences consisting of mixed zeros and ± 1 for analysis.

Solutions for x from solving equation 3.13 represent the relative compaction of the unfolded state, i.e., the ratio of the unfolded end-to-end distance when including electrostatics and excluded volume effects to the Flory random coil distance. The values of x were calculated for each thermophile and mesophile sequence in the 540 ortholog pair dataset. From these values, respective averages of x were calculated within thermophilic and mesophilic sets. We found averages of $\langle x_{meso} \rangle = 1.84$ and $\langle x_{thermo} \rangle = 1.42$. The comparison of the averages gives a P–value of 5×10^{-15}.

A more direct analysis was conducted for each thermophile–mesophile ortholog pair. For each pairing, the ratio of relative compactness x_T of the thermophilic sequence to the compactness x_M of the mesophilic sequence was calculated, $\alpha = x_T/x_M$. Values of α less than unity reflect those pairs with a more compact unfolded state in the thermophile sequence. Values of α greater than one show the pairings which have a more compact denatured state in the mesophile, contrary to our expectation. Here, we introduce a 5% error in calculation of the α ratio, i.e., if the α fraction is between 0.95 and 1.05, we claim neither the thermophilic nor mesophilic sequence is more or less compact. This error region is shown in orange in Figure 3.5. With this error, accounting for inaccuracies that may have resulted for various approximations, we find 65% of the pairings have a more compact unfolded state in the thermophile, 25% of the orthologs are more collapsed in the mesophile sequence, and the remaining 11%

are inconclusive (Figure 3.5). These statistics along with the comparison of averages ($\langle x_{meso} \rangle > \langle x_{thermo} \rangle$) strongly suggests thermophilic sequences, on average, have more compact unfolded states than mesophilic proteins.

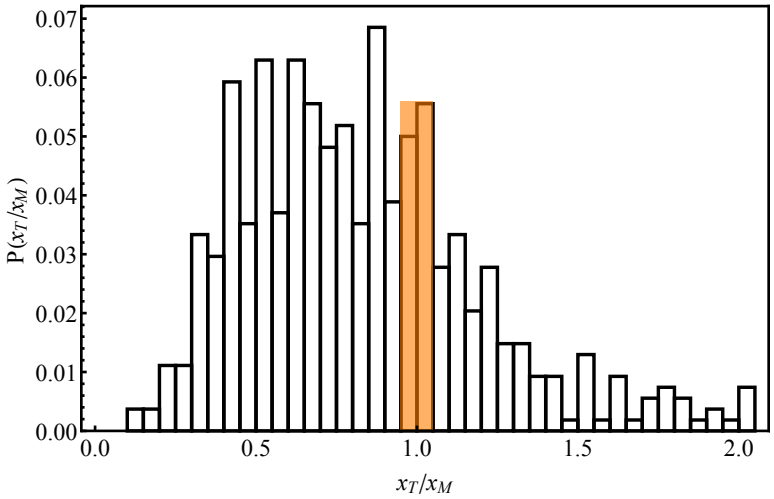

Figure 3.5: Results of analyzing ortholog set of 540 mesophile–thermophile pairs from Fang et al.[217] We plot the distribution of the ratios of thermophile relative compactness to mesophile compactness per pairing. Orange shows the tolerance region between 0.95 and 1.05 accounting for a 5% uncertainty to account for inaccuracies due to approximations.

To investigate the origin of electrostatics in the compactness of the unfolded state a bit more deeply, we reduced the set of 540 mesophile–thermophile pairs to include only those that showed an attractive interaction in the electrostatic component of equation 3.13 in the thermophile sequence of the pairings. This restriction reduced the set of 540 pairs to 383 ortholog pairs. Repulsive overall electrostatic interactions, manifest in the last term of equation 3.13 being positive, will not promote reduction below the reference excluded volume size. Averages of $\langle x_{meso} \rangle = 1.66$ and $\langle x_{thermo} \rangle = 1.04$ are calculated, and from comparison of the averages gives a P–value of 1.1×10^{-37}.

Here, in head-to-head comparison between two proteins in a given pair with the $\pm 5\%$ error margin about $\alpha = 1$, 76% of thermophile sequences had a more compact unfolded state than their mesophilic partner, 15% of the pairs showed a

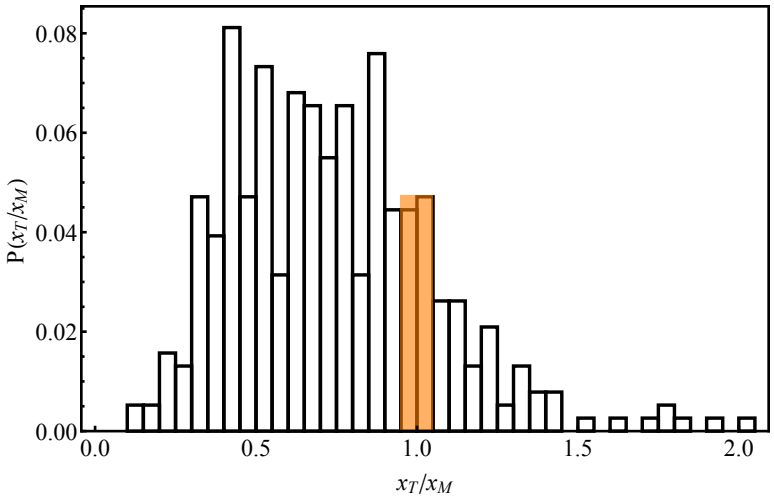

Figure 3.6: Results of analyzing the reduced set of 383 mesophile–thermophile orthologs. We plot distribution of the ratios of thermophile relative compactness to mesophile compactness per pairing. Orange shows the tolerance region between 0.95 and 1.05 accounting for a 5% uncertainty to account for inaccuracies due to approximations.

smaller mesophilic denatured state, while 9% were inconclusive in the orange region of Figure 3.6. Including the unfolded state attractive electrostatic criteria displays a stronger discriminator toward reduced compactness in thermophilic proteins.

We have also used experimentally determined unfolded state radius of gyration under highly denaturing condition measured by Kohn et al.[205] to determine the excluded volume strength (ω), and found the overall statistics to be very similar. Based on 540 pairs, $\langle x_{meso} \rangle = 2.42$ and $\langle x_{thermo} \rangle = 2.1$ are calculated, and from comparison of the averages gives a P–value of 2.2×10^{-12}. Furthermore, we have done our analysis by lifting the assumption that Kuhn length is equal to the bond length. In the new analysis, we assume each Kuhn segment to consist of two amino acids, allowing the possibility that Kuhn length in proteins is roughly twice the bond length.[215,216] Net charge of the Kuhn segments were assumed to be the addition of charges of the two consecutive amino acids that make the Kuhn segment. This analysis also yielded similar statistics highlighting the robustness of our results. For 540 pairs, $\langle x_{meso} \rangle = 1.9$ and $\langle x_{thermo} \rangle = 1.3$ are calculated, with a P–value of 7.0×10^{-16}.

3.2.3 Additional comparison with homolog structures shows a similar trend

While the above data set was well curated, invoking algorithms to confirm the sequences contained same domains, possessed high sequential similarity, and were probably "true" homolog pairs, there was no guarantee of structural similarity. Here, we will utilize a second data set from Robinson-Rechavi and Godzik[218] that consists of strictly well-aligned thermophilic-mesophilic homolog structures.

This structural set is substantially smaller in size, with only 55 ortholog pairs of non-hypothetical protein domains. We followed an identical protocol in calculating unfolded state relative compactness, and we find the average values of $\langle x_{meso} \rangle$ and $\langle x_{thermo} \rangle$ to be 1.96 and 1.81, respectively. Furthermore, the ratio $\alpha = x_T/x_M$ shows a tendency of reduced compactness in thermophilic structures in 58%, while the mesophilic unfolded state is smaller in 22%, see Figure 3.7. The remaining structures fall in the 5% error region highlighted in orange.

Again, we reduced the set of 55 structural homolog pairs to include only those that showed an attractive interaction in the electrostatic component of equation 3.13 in the thermophile sequence of the pairings, leaving 46 pairs. Averages of $\langle x_{meso} \rangle = 1.90$ and $\langle x_{thermo} \rangle = 1.64$ were calculated, and from comparison of the averages, a P–value of 2.3×10^{-4} was found.

In head-to-head comparison between constituents of the 46 thermophilic-mesophilic pairs, 67% of thermophile sequences had a more compact unfolded state than their mesophilic partner, 20% of the pairs showed a smaller mesophilic denatured state, while 13% were considered inconclusive in the orange region of Figure 3.8.

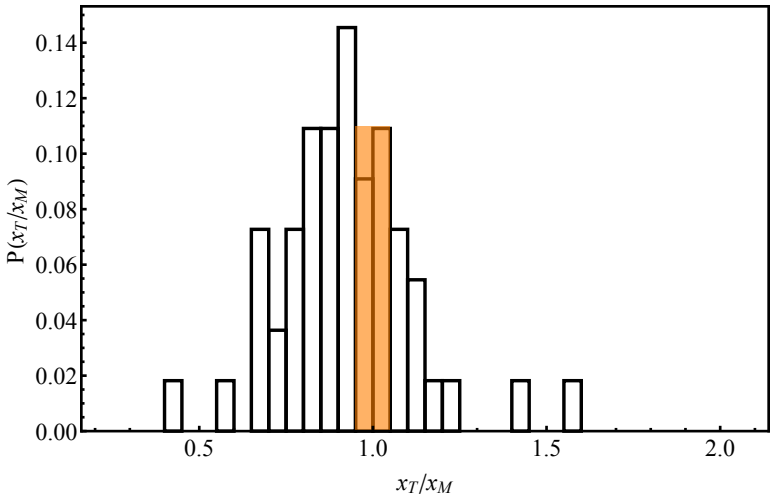

Figure 3.7: Results of analyzing set of orthologous mesophile–thermophile structural pairs from Robinson-Rechavi and Godzik.[218] We plot the distribution of the ratios of thermophile relative compactness to mesophile compactness per pairing. Orange shows the tolerance region between 0.95 and 1.05 accounting for a 5% uncertainty to account for inaccuracies due to approximations.

3.2.4 Thermophilic protein sequences have more segregated charge distribution

From the above analysis, we realize the relative compactness of protein denatured state is directly related to the negative value of the electrostatic contribution (second term on the right hand side of equation 3.13). This term accounts for charge patterning encoded in the sequence. If the excluded volume terms are the same, then from the zero salt limit (equation 3.13) we notice a simple form of this sequence charge decoration (SCD) metric that dictates changes in the unfolded state dimensions. SCD is defined as

$$SCD = \frac{1}{N}\left[\sum_{m=2}^{N}\sum_{n=1}^{m-1} q_m q_n (m-n)^{1/2}.\right] \tag{3.14}$$

Das and Pappu[185] have defined a somewhat different charge decoration metric (ζ) to classify the thirty sequences using the degree of charge mixing. Note, we use a

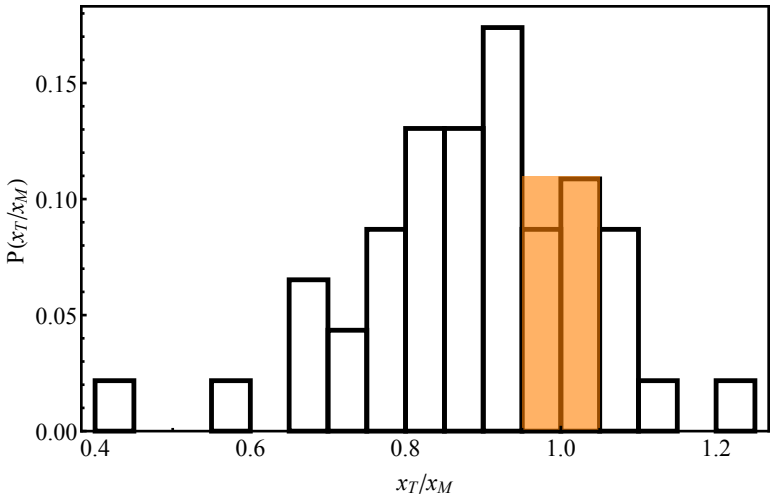

Figure 3.8: Results of analyzing the reduced set of 46 mesophile–thermophile structural domain orthologs. We plot distribution of the ratios of thermophile relative compactness to mesophile compactness per pairing. Orange shows the tolerance region between 0.95 and 1.05 accounting for a 5% uncertainty to account for inaccuracies due to approximations.

separate symbol ζ than the original symbol (κ) of Das and Pappu to avoid confusion with the inverse Debye length. Their metric ζ ranks sequences 1 to 30 in increasing order, indicating sequence 1 has the lowest degree of charge separation ($\zeta = 0$), while sequence 30 has the highest degree of charge separation ($\zeta = 1$) .[185] As a test of our proposed charge patterning metric, we compared SCD against the empirical charge decoration metric (ζ) of Das and Pappu,[185] and noticed a strong correlation. Figure 3.9 shows SCD is indeed highly correlated to ζ indicating highly negative SCD implies higher degree of charge segregation. Furthermore, based on our proteome-wide analysis, we realize highly negative values of SCD lead to reduced denatured state dimensions. These two findings together suggest more compact denatured states in thermophiles may arise from relatively high degree of charge segregation (negative value of SCD) in thermophiles than in mesophiles.

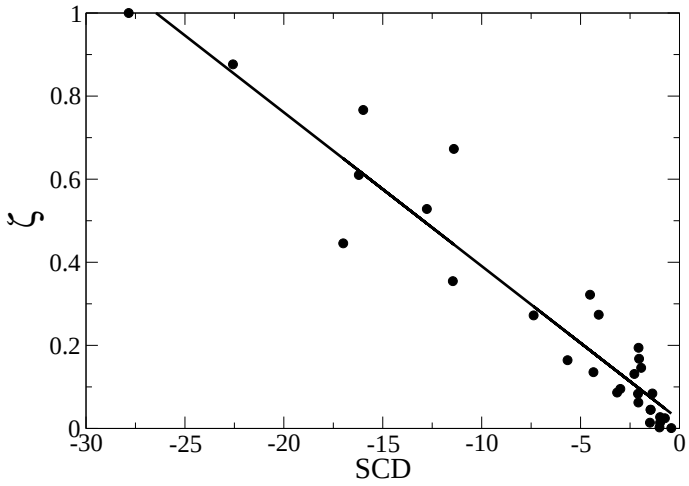

Figure 3.9: Charge segregation metric ζ introduced by Das and Pappu (termed κ in their work) correlates well ($R^2 = 0.95$) with a sequence charge decoration metric SCD defined in equation 3.14. Each point corresponds to a sequence in Figure 3.1.

3.3 Conclusion

To summarize, based on charge decoration alone, we find thermophilic proteins have more compact denatured states compared to their mesophilic counterparts. This observation is in line with two independent observations made by earlier studies: i) thermophilic proteins have more residual structure in their unfolded state, and ii) electrostatics plays a key role in thermophilic adaptation. Furthermore, we notice a primary mechanism to achieve this is by promoting effective attractive interaction in the thermophilic sequence quantified by the negative values of the electrostatic contribution in equation 3.13. A closer inspection further reveals that about 86% of the protein pairs within this set also have lower net charge for thermophiles than mesophiles. These observations, taken together, imply charge balance,[87] is a necessary condition to impart residual structure in thermophilic unfolded state. Although necessary, charge balance may not be sufficient, as different sequences can be cre-

ated with varying contribution to the electrostatics component in equation 3.13 while maintaining the same net charge.[185] Similar conclusions, highlighting the importance of specific patterning of charges, were also drawn describing the role of dielectric in thermophilic adaptation.[87] Further, introducing a charge patterning metric, SCD, we find thermophiles, on average, may have more segregated charges. We reiterate that the sequence patterning, in addition to the composition, of charges is crucial for thermophilic adaptation.

Chapter 4

Convergence of All-Atom Biomolecular Simulations

Properties of the protein native state will be investigated by utilizing long time scale, room temperature molecular dynamics simulations. Most previous studies used relatively short simulation trajectories, with little attention paid to quality of sampling, leaving the reported observables unreliable, or artifacts of inadequate sampling. Therefore, we will introduce a novel technique to test the convergence of molecular dynamics simulations that allow the sampled trajectories to be claimed as adequately sampled and self-consistent.[130]

Ensuring convergence in biomolecular simulations is necessary to guarantee the quality of data. How long must one simulate to ensure convergence? As simulation time T_{sim} approaches infinity, the value of an observable $\mathcal{A}$ derived from such a long trajectory should settle to a constant value $\mathcal{A}_0$ (defined as "true" for subsequent discussion). However, the notion of infinity does not exist in numerical simulation. Consequently, the argument above is often reversed, and the resolution seems to rely on the constancy of the observable ($\mathcal{A}$) for a sufficiently long time. The notion of

"sufficiently long" is vague, and there is no way of knowing whether $\mathcal{A}$ has reached the "true" value without knowing $\mathcal{A}_0$, a priori. Thus, different tests for convergence, without the knowledge of the "true", have been devised to quantify "how long is sufficiently long"?

While some convergence tests are specific to individual observables, others are global, and try to ensure convergence of the system irrespective of any particular observable. For a good discussion on single observable error estimations, and global convergence assessment methods, see the review by Grossfield and Zuckerman.[219] However, as repeatedly stated,[219–223] none of these tests can guarantee "true" convergence, instead they can only provide a "self-consistency" check (SCC) of the simulation. The demand of self-consistency can significantly vary in complexity, and consequently in the requirement of different lengths of simulation trajectories for a given protein. Perhaps the most simple of these criteria is from Smith, Daura and van Gunsteren (referred to as SDVG)),[224] which demand simulation trajectories should be sufficiently long such that it has explored all possible conformations. In subsequent work, Lyman and Zuckerman[220] (termed LZ for future reference) has shown an inadequacy of this criteria. According to the LZ criteria, in addition to the constancy in the total number of found clusters, one must ensure these clusters have been adequately populated. This is enforced by demanding the cluster probability distributions from the first half and the second half of the trajectory match within some error. Clearly, this is more stringent than just demanding that all of the clusters have been visited. Hess has provided a separate criteria altogether by using a principal component analysis (PCA) that does not depend on the definition of clustering.[225] Accordingly, self-consistency is determined by monitoring a metric quantifying the covariance overlap of the eigenvalue spectrum between the full and partial subsets of the trajectory. Because the comparison is made with respect to the end of the

simulation, the metric always tends to unity regardless of the length of the trajectory. In more recent work, Romo and Grossfield[223] have proposed a Block Covariance Overlap Method (BCOM) to monitor self-consistency by determining the slowest relaxation time scale. According to this analysis on GPCR proteins, microsecond (μs) long simulation trajectories could not be considered converged. While the SDVG, Hess criteria, and other PCA based criteria have been tested frequently by numerous researchers,[226–233] the LZ criteria has only been tested in folding-unfolding simulation of small peptides. It was shocking to realize that for such a small system, hundreds of nanoseconds of simulation were required to satisfy LZ criteria.[220] This raises a natural question: what time scales are required to satisfy the LZ criteria in simulating real, globular proteins?

One of the primary applications of molecular simulations of proteins is in predicting the folded structure, and the associated convergence criteria can be different than that for studying only native state properties. For example, Lin and Shell proposed to monitor the highest populated cluster which is relevant to ensure folding to the correct structure.[234] While the majority of the studies mentioned above, with one exception,[223] have mostly focused on convergence in folding-unfolding trajectories, studies of convergence for folded state simulations are rare. Analyzing native state properties is also important in protein science; for example in the studies of substrate and ion binding/unbinding,[235–237] effects of mutation on dynamics,[51,229,238] structural stability and flexibility,[229,230,239,240] and function.[181,241–244]

Molecular dynamics simulations of protein native states have also been extensively applied to understand the role of the native state in thermophilic adaptation. Studies have claimed thermophilic protein native states have higher rigidity compared to their mesophilic counterparts,[50,54,55,245,246] while others have investigated the role of ion pairs in thermophilic proteins.[61,63,79] A higher dielectric constant[86,87] in the

thermophilic folded state have also been associated with high temperature tolerance. Ensuring convergence, and adequate sampling to deduce such claims is important.

It is tempting to assume folded state simulations may be relatively easy compared to folding-unfolding studies that are faced with a highly heterogeneous ensemble. However, work by Romo and Grossfield[223] have suggested, for GPCR proteins, microsecond long simulations may not be sufficient to pass the BCOM self-consistency check. Given these demands and expectations, it is now timely to ask: how long does it take to establish convergence in native state simulations of real globular proteins? Here, this issue will be critically examined by systematically studying multiple protein trajectories (at 300 K) subjected to multiple self-consistency check criteria. The following systems were chosen: the 1ms simulation of BPTI of Shaw et al;[247] and from our own lab, a 20μs trajectory of the cold shock protein (CSP) in implicit solvent, a 12μs simulation of CSP, a 7μs of CheW, and a 2μs trajectory of CheA all in explicit solvent. The length of the simulation needed to satisfy the same self-consistency criteria varied significantly between proteins. Simultaneously, for a given protein, there is significant variation among the different SCC criteria. For example, the implicit simulation of CSP may pass the SDVG criteria after 2μs, but does not pass the self-consistency criteria of LZ or BCOM after 6μs of simulation.

Why do systems vary significantly in their convergence time requirement? While the answer to this question fundamentally lies in the sequence-structure-function properties of the protein, it has been observed that a given protein may visit a particular state more frequently than other states, exhibiting a trap-like feature in these long simulations of the native state. In order to detect such traps in the simulation, an additional self-consistency metric is introduced that is more reliant on the cluster probability distribution that is more stringent than the SDVG criteria. This metric is based on the constancy of the cluster entropy (termed as CCE) of the cluster prob-

ability distribution. It is true that an adequate sampling must find as many clusters available, but the proper sampling of these states is equally important. A simple calculation to inspect the trajectory sampling is to calculate the entropy of the cluster probability distribution. If the simulation has settled to a stable distribution, the plot of the entropy as a function of time will be relatively constant. However, substantial changes to the entropy plot, as seen in many of the protein systems presented below, reflect large changes in the probabilities of the discovered clusters. One primary advantage of this metric is to ensure if a protein is "stuck" in a metastable state not detected by the SDVG criteria alone. Also, some of the algorithms designed to determine decorrelation times, including those based on kinetic clustering,[248] do not detect these traps. Therefore, one should constantly monitor CCE, and if a trap is detected by a continuous decrease in cluster entropy, the simulation should either be allowed to continue until the entropy stabilizes, or the simulation should be aborted.

In summary, as a first assessment of adequate sampling, the number of discovered clusters must first reach a constant value. Second, the cluster entropy should be stable, as changes to the entropy reflect large changes in the underlying probability distribution. Once this dual criteria is ensured, more rigorous testings of convergence can be imposed. And as a forewarning, this process may require tens of microseconds for native state simulations of globular proteins.

Below, the performance of multiple self-consistency criteria (SCC) from our simulations of CheW and CheA in explicit solvent, and CSP in implicit solvent are presented. Finally, the 1ms BPTI simulation of Shaw et al,[247] and our own 20μs trajectory of CSP in explicit solvent are analyzed, and the relative performances of SCCs to reveal insights about regions of self-consistency and its non-equivalence to "true" convergence.

4.1 Materials and Methods

4.1.1 Computational Setup

The 167 residue chemotaxis protein W (CheW) from E. Coli, the 105 amino acid phosphotransferase domain of chemotaxis protein A (CheA) from Thermotoga Maritima, and the 73 residue cold shock protein (CSP) from Thermus Thermophilus were chosen as model systems. Initial coordinates were taken from Protein Data Bank (PDB), specifically PDB IDs 2HO9[249] for the structure of CheW, 1TQG[250] for CheA, and 3A0J[251] for the CSP structure. All non-protein atoms, including crystallographic waters and structural hydrogens, were removed. Protonation states of histidine side chains were predicted with H++ (http://biophysics.cs.vt.edu/H++).[252–254] Using TLEAP, protein hydrogens and neutralizing ions were added. For the explicitly solvated systems, a cubic box with TIP3P[255,256] water molecules with a 9Å buffer region was used giving 6436 water molecules in CheW, 4432 in CheA, and 3833 in CSP. The system solutions were energy minimized for two cycles using the SANDER module of AMBER 14. The first cycle held the protein constant with harmonic restraints to minimize poor steric positions of the solvent. The restraints were lifted, and a second cycle of minimization allowed the entire solution to adjust to a local minima.

For explicit solvent CheW, CheA , and CSP systems, simulations were performed with the GPU-accelerated PMEMD module of AMBER 14[257] using the ff99SB force field,[258] periodic boundary conditions, and SHAKE[259] to constrain the lengths of bonds that include hydrogen atoms to their equilibrium lengths. A 9Å cutoff was chosen for the direct sum, non-bonded interactions, and the Particle Mesh Ewald method[260–262] was used to calculate long-range electrostatics. After energy minimization, solvent molecules were heated gradually to 300K over a 100ps period at constant volume. The entire system was equilibrated for 5ns at constant temperature and pres-

sure, allowing the volume to change to adjust the system's density. Simulations were performed at constant temperature (300K) and pressure (1atm) by employing the Langevin thermostat and Berendsen barostat, respectively, with a 2fs time step, and structural conformations saved every 10ps.

The implicit solvation simulation of CSP also utilized the GPU-accelerated PMEMD module[263] with the ff99SB force field, SHAKE, and Langevin thermostat, but with no cutoff for non-bonded interactions. System was again energy minimized for two cycles, heated to 300K, and equilibrated for 5ns. Production simulation was carried out at 300K, with a 2fs time step, and structural coordinates were written every 10ps.

4.1.2 Clustering

The cluster command within the CPPTRAJ module of Amber[264] was used to cluster all simulation trajectories. The agglomerative, average-linkage algorithm was employed.[265] In summary, each structure begins in its own cluster, and pairs of clusters are combined only if the average distance between all points of inter-cluster elements are less than the preselected distance cut-off (d_c). From this clustering module, it was possible to quantify the rate of new cluster discovery, and cluster occupancy as a function of simulation time. For clustering and PCA based analysis, we use cartesian representation. This is justified for folded state studies where decoupling of internal motion and rotational motion is possible, unlike folding-unfolding simulations where internal coordinates are preferred.[266]

4.1.3 Convergence inspection protocol

In the cluster counting criteria termed as SDVG, a trajectory is considered converged as the rate of discovery of new clusters approaches zero.[224] The LZ method

imposes further constraints by demanding the first half of the trajectory's probability distribution must agree with that derived from the second half of the trajectory.[220] For monitoring the constancy of the cluster entropy (CCE), we compute $\sum_i p_i \log(p_i)$, where p_i is the probability of the i^{th} found cluster, as a function of simulation time.

To more rigorously assess sample quality, two methods from the Zuckerman group were used to estimate a simulation's decorrelation time. The first method partitions a trajectory by constructing a histogram from randomly chosen reference states, and from subsamples taken from the trajectory, a variance is calculated from the histogram's bin populations observed in the subsamples. By comparison to the variance expected if the histogram bins consisted of fully independent and uncorrelated structures, a decorrelation time (τ_d) is estimated.[222] The second method again partitions the trajectory, but then finds physical states by merging histogram bins with high rates of exchange.[248] From physical state probabilities within blocks, an effective sample size, and therefore a decorrelation time (τ_d') is calculated.[267] Additionally, we utilize the Block Covariance Overlap Method (BCOM) of Romo and Grossfield[223] which calculates the average covariance overlap of Principal Component Analysis (PCA) results from individual contiguous trajectory blocks of a given size to the PCA results of the full trajectory. The covariance overlap metric has a range of values from zero (completely dissimilar fluctuations) to one (identical fluctuations). For well-sampled simulations, larger trajectory blocks should possess similar fluctuations. Therefore, the average overlap is expected to approach unity for increasing block sizes. The average covariance overlap for each block size is normalized by the covariance value if the blocks consisted of fully uncorrelated structures found by bootstrapping the trajectory frames. The inverse of the normalized ratio is expected to decay to one as the block size is increased. All of the aforementioned assessment tools (τ_d, τ_d', and BCOM) were calculated with the LOOS convergence package.[268]

Systems were initially simulated for one microsecond, and the SDVG, LZ, and CCE convergence tests were imposed. For the SDVG criteria to pass, we demanded that no new clusters were discovered in final 500ns of simulation, and for the CCE metric to be satisfied, we wanted the largest percent difference in entropy in the same 500ns block, to not be greater than 5%. If this dual criteria was not met, trajectories were extended for an additional 500ns, and again tested for convergence. For this work, we will analyze explicit solvation simulations of CSP (12μs), CheW (7μs), and CheA (2μs), and an implicit solvated 20μs trajectory of CSP, in addition to 1 ms of BPTI simulation in explicit solvent.[247] With our in-house hardware, running on an individual graphics card, we achieve nearly 450ns/day on the GPU-accelerated AMBER code for the implicit CSP system. For comparison, running the same system simulation on an 8-core CPU machine, we record 11ns/day.

4.2 Results and Discussion

4.2.1 Convergence of the cluster probability distribution can be highly demanding

We critically assess the convergence of a simulation trajectory to determine the time scales required to adequately sample realistic globular proteins.

Explicit solvent simulation of CheW

First, we present the 7μs explicit solvent trajectory of CheW, following the simulation methodology outlined above. Per the convergence inspection protocol, and clustering with a 1.5Å cutoff, the simulation did satisfy the SDVG criteria, but comparison of the first half to second half cluster probability distributions show a large

disparity in the sampling of the discovered states, and therefore not passing the LZ criteria (Figure 4.1, left and middle). Also, the trajectory does not pass the BCOM criteria, as we expect the values of the BCOM analysis to approach those from the bootstrapped BCOM, and the ratio of bootstrapped BCOM-to-BCOM to approach unity at larger block sizes (Figure 4.2).

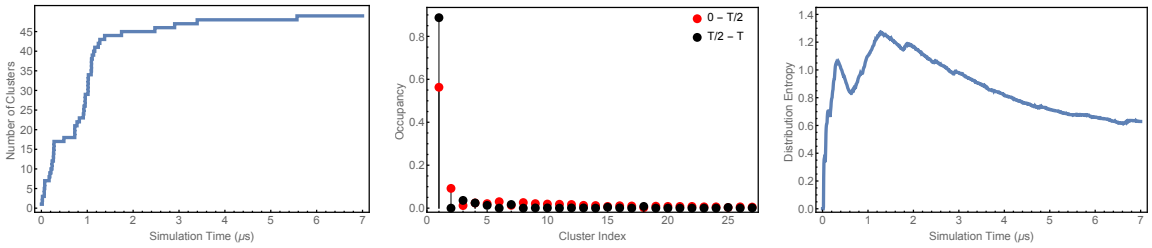

Figure 4.1: Results of convergence tests for the CheW explicit solvent simulation at7μs. Left panel shows number of clusters versus simulation time (SDVG method), where clustering was calculated with $d_c = 1.5$Å. Middle panel shows histogram of cluster occupancy of the most occupied clusters (LZ method). The first half (red) to second half (black) occupancy comparisons show large discrepancy between cluster populations. The CCE assessment is displayed in the right panel.

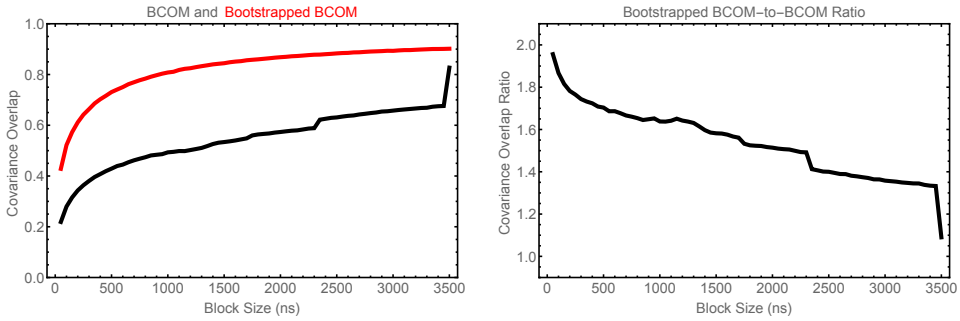

Figure 4.2: Results of Block Covariance Overlap Method (BCOM) analysis of CheW 7μs trajectory. Left panel is direct calculation of covariance overlap using contiguous blocks (black curve) and bootstrapping (red curve) at given block sizes. Converged trajectories would show the contiguous BCOM approaching values from bootstrapped BCOM in the large block size limit. Right panel shows the ratio of bootstrapped BCOM to blocked BCOM, and this curve is expected to decay to one as block size is increased.

To gain further insights to the lack of convergence, we investigate the relative stability of the distribution by monitoring the cluster probability distribution entropy

($\sum_i p_i \log(p_i)$, where p_i is the probability of state i) as a function of the simulation time. The rationale is, if the underlying configurational distribution is not changing substantially, the entropy plot should be stable and constant between $T - t$ and T, with T being the total simulation time. We call this constancy of cluster entropy (CCE) criteria. Here, the cluster entropy exhibits a significant decrease between 1.25 and 7μs (see Figure 4.1, right). A continuous decrease in the entropy indicates: i) the simulation is yet to settle to a stabilized state population, and ii) the distribution is getting heavily biased towards one state with the progression of the simulation. Examination of the cluster probability distribution shows not only a substantial difference in the distributions, but in the second half of the simulation, one particular state accounted for a very large portion ($\approx 90\%$) of the 3.5μs of sampling. Also, by plotting the state probabilities as a function of simulation time (see Figure 4.3), we see decaying occupancies in all but one state, verifying the bias.

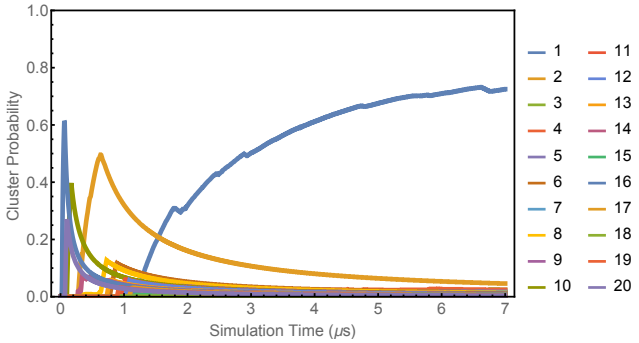

Figure 4.3: Probabilities of the top 20 most occupied clusters as a function of simulation time for the 7μs simulation of CheW. The number of clusters, and their respective populations, are found using a $d_c = 1.5$Å. Cluster 1(the most populated state) was approximately the only sampled state starting at 1.1μs.

To ensure there is no dependency on the choice of the clustering distance cutoff (d_c), we clustered and analyzed the 7μs trajectory for three unique values of d_c, and found the qualitative features of the results remain unchanged. These graphs are shown in Figures A.1 and A.2.

Would the presence of traps in a trajectory be associated with long decorrelation times? It is natural to expect decorrelation time estimation would perhaps be able to detect these traps. To this end, we estimate the decorrelation times, and therefore, an error bar by calculating the effective sample size from the ratio of the total trajectory length and the decorrelation time.[222,223,267] We use two separate algorithms to determine the two decorrelation times, τ_d and τ_d', where the latter is based on the kinetic exchange partitioning[248,267] (see Methods for details). Both of these estimates report a modest error of 12-19% (see Table A.1) while the trajectory itself does not pass self-consistency check criteria using LZ, BCOM, and CCE. Clearly, the lack of convergence and the presence of traps are not necessarily detected by decorrelation times. Unlike other SCC tests, monitoring CCE can quickly identify whether a simulation is being heavily biased towards one state, and can therefore serve as a potential decision maker to choose between continuing the simulation, or aborting and starting a fresh simulation.

Explicit solvent simulation of CheA

To further investigate the advantage of monitoring CCE metric, we next consider the trajectory of the phosphotransferase domain of chemotaxis protein A (CheA) in explicit solvent. Simulation and convergence inspections are identical to the CheW simulation above. Within 500ns, the number of found clusters has reached a plateau, and the system proceeds for another 500ns (total of 1μs) without the discovery of a new state (Figure 4.4, left panel). However, similar to CheW, the cluster distribution entropy plot shows large changes throughout (Figure 4.4, right panel), implying the configuration distribution is not settled, and the trajectory has failed to pass the CCE criteria. Not surprisingly, the cluster occupancy histogram shows large discrepancies between the first and second-half of the trajectory (Figure 4.4, center panel).

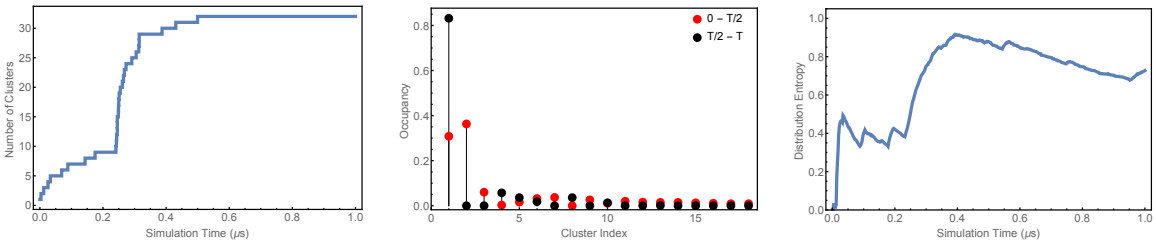

Figure 4.4: Results of convergence test for the CheA simulation after 1μs. Left panel shows number of clusters at $d_c = 1.5$Å versus simulation time (SDVG), and right panel shows entropy evolution (CCE). The middle panel shows the histogram of cluster probabilities of found clusters (LZ). The first half (red) to second half (black) occupancy comparisons show large discrepancy between cluster populations.

For the longest CheA simulation (2μs), the SDVG criteria is again easily satisfied (Figure 4.5, left). The cluster entropy now shows stabilization in the last 0.5μs, indicating stabilization of the distribution in the last 500 nanoseconds out of a total of 2μs (Figure 4.5, right). The results of the LZ criteria show an improvement in cluster occupancies for the most occupied cluster in the trajectory (Figure 4.5, center). However, the large discrepancies in the occupancy in the next few most populated clusters remain, and this simulation would not satisfy the LZ criteria. Both of the 1 and 2μs trajectories were subjected to the self-consistency criteria test using BCOM (Figure A.3). The trajectory fails to satisfy BCOM criteria, although it slightly improved for the 2μs trajectory when compared to 1μs simulation. The analysis reiterates the strategy that one must first ensure passing of CCE and SDVG before imposing more rigorous tests. Decorrelation times and error bars are reported in Table A.1.

Implicit solvent simulation of Cold Shock Protein

The test cases presented above show the need for longer simulations in order to investigate the more stringent self consistency checks. Compared to explicit solvent simulations, large scale sampling results can be obtained in shorter times using im-

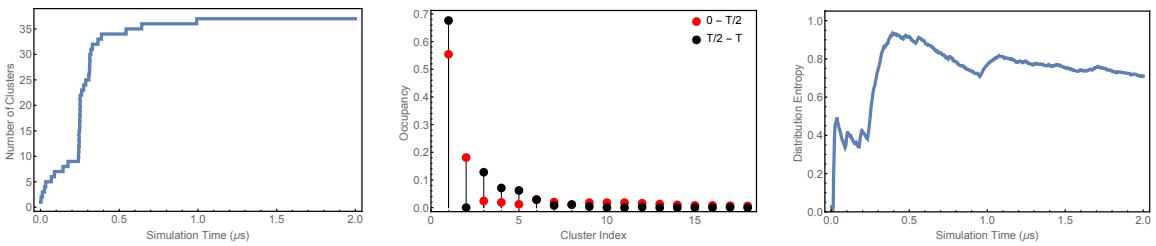

Figure 4.5: Results of convergence test for the CheA explicit simulation at 2μs. Left panel shows number of clusters versus simulation time (SDVG). Center panel shows histogram of cluster occupancy (LZ method) of found clusters, and their comparison between the first (red) and the second half (black). Right panel shows time evolution of the cluster entropy (CCE) which is relatively constant in the last 0.5μs.

plicit solvent simulations. Motivated by this, and its relatively smaller size, the cold shock protein (CSP) from Thermus Thermophilus was simulated in implicit solvent to serve as a third test case, different from the two explicit solvent systems presented in the preceding sections. For the implicit CSP system, we see after 1.3μs, the system has found all unique clusters (Figure 4.6, left), and the simulation continues to 2.5μs without the discovery of any new clusters, easily satisfying the constant clusters protocol of SDVG. However, when we compare the cluster population between the first and the second-half of the trajectory according to LZ criteria (Figure 4.6, middle), we see the simulation is far from convergence. The trajectory also shows a pronounced decrease in the cluster entropy, dropping 25% in the last 1μs (see Figure 4.6, right), reiterating an unsettled distribution, and a simulation that has not reached convergence.

How much longer must the trajectory proceed to satisfy the CCE criteria? The simulation was repeatedly extended in hopes the probability distribution would stabilize, and the CCE criteria would be satisfied. From the previous 2.5μs time stamp, the simulation required a total simulation time of 6μs to converge in both SDVG and CCE methods, but did not satisfy the LZ method. Figure 4.7 shows the analysis of the three SCC criteria from the 6μs trajectory. The cluster entropy (Figure 4.7, right)

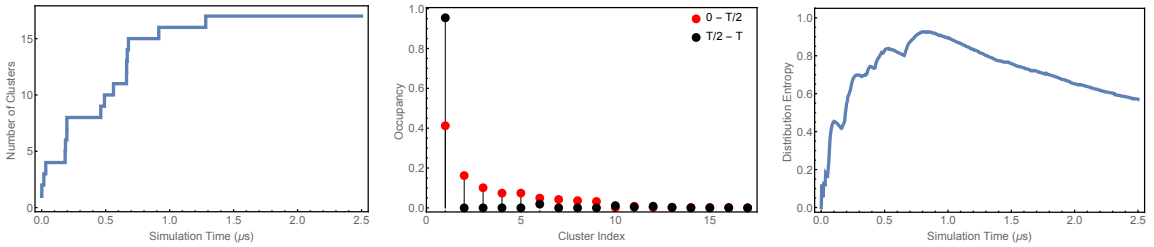

Figure 4.6: Results of convergence test for the CSP implicit simulation at 2.5μs. Left panel shows the number of clusters versus simulation time with a clustering $d_c = 1.5$Å. After 1.3μs, no new clusters are found, and the simulation is considered converged in the SDVG criteria. Middle panel shows histogram of cluster occupancy and their discrepancy between the first half (red) to second half (black) of the trajectory. Right panel plots the entropy of cluster distribution with time indicating unsettled cluster probability distribution.

reflects the ongoing changes to the probability distribution up to 4μs, and afterwards, approached a constant.

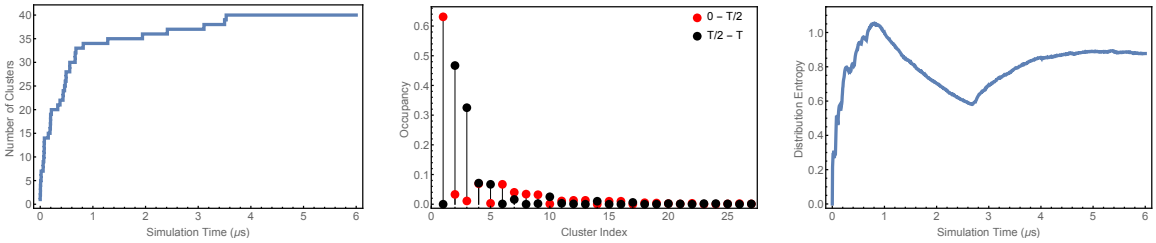

Figure 4.7: Convergence tests for a trajectory of 6μs of CSP. The left panel shows the number of clusters is constant near the end of the simulation, satisfying the SDVG method, but the center histogram shows the discrepancy in cluster occupancy between the first-half and second-half of the simulation. The right plot displays the entropy of the cluster probability distribution as a function of simulation time which stabilized after 4μs.

It is important to note, that we cannot guarantee the distribution will remain unchanged if simulation is extended. We can say, assuming the 6μs simulation is all the information we possess, the trajectory has discovered all of its clusters, and is sampling the cluster probability distribution without any trap for the last 500 nanoseconds. Results of BCOM analysis for this trajectory are shown in Figure A.4. The error bars for 2.5 and 6μs long trajectories were computed by determining the decorrelation times, and are reported in Table A.1. Again, we see these decorrelation times are reasonable, and do not give us any indication regarding traps.

4.2.2 Comparison of observables at different timescales using Cold Shock Protein Implicit solvent simulation

We present a comparison between the CCE and SDVG criteria in predicting specific observables. As illustrated, the 2.5μs CSP simulation is sufficient to pass SDVG criteria, while the CCE criteria is satisfied after 6μs. We set our longest simulation of implicit CSP (20μs) as the "gold standard", and explore the difference between calculated observables from different length trajectories that passed convergence tests at different time scales, and the "gold standard". This would provide a sense of performance of different convergence criteria, particularly with SDVG and CCE.

We evaluate the performance against contact maps, or probabilities of different contacts, that are often used to gain insights.[269,270] For this analysis, a contact is defined as two C_α atoms separated less than 8Å, and we calculate the probability that a given residue pair i, j form a contact k. For three simulation times of interest (2.5μs, 6μs, and 20μs), we calculate the contact probability for all possible i, j residues pairs, where $j \geq i + 2$ to remove nearest neighbor effects. The contact probabilities are denoted by P_k for the k contact. The P_k values are sorted with respect to the "gold standard" probabilities, and plotted in Figure 4.8. The dispersion about the 20μs curve reflects the varying probabilities the simulations produce for the same contact. However, it is qualitatively evident the dispersion is substantially reduced in the 6μs data when compared to 2.5μs trajectory.

To better quantify the dispersion when comparing the shorter time scale simulations to the "gold standard", the contact probabilities (P_k) from the 2.5μs and 6μs simulations were directly compared to that of the 20μs trajectory. Ideally, this comparison should be linear plot with a slope and an R^2 both equal to one (Figure 4.9). For the 6μs long trajectory satisfying CCE criteria, the R^2 value is 0.89; and for the

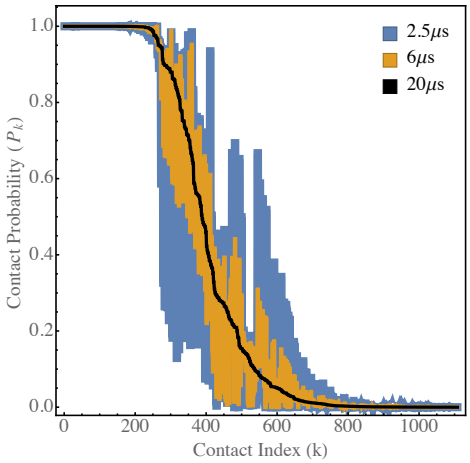

Figure 4.8: The probability of contact between amino acids i, j, indexed as k, for different simulation time scales; 20μs in black, 6μs in orange and 2.5μs in blue. The probabilities are sorted with respect to the 20μs simulation.

2.5μs long trajectory satisfying SDVG criteria, R^2 is 0.63.

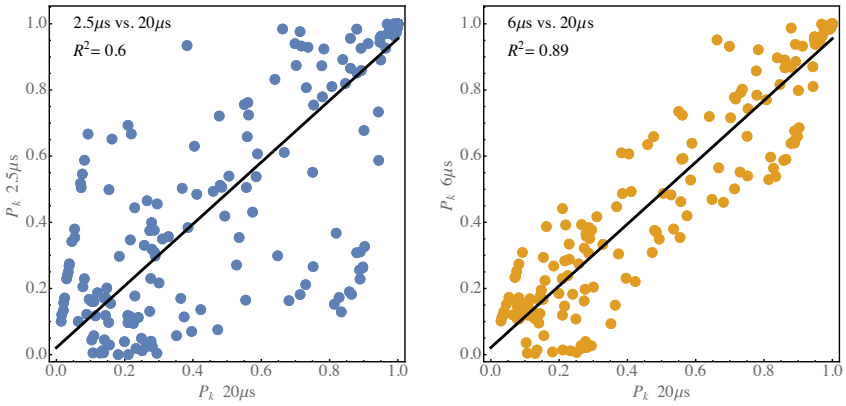

Figure 4.9: Direct comparison of contact probabilities from CSP simulations of shorter length (2.5μs, left panel, and 6μs, right panel) to those of the longest available trajectory of 20μs. Each data point represents the same i, j contact in both simulation time scales, but with differing probabilities. The reduced scatter evident in the 6μs simulation, with respect to the shorter 2.5μs trajectory, displays the improvement in the contact observable calculated from simulations passing CCE convergence criteria compared to the SDVG method.

4.2.3 Shortcomings of CCE convergence criteria

Calculating the entropy of the cluster probability distribution can help determine whether the trajectory is sampling the underlying probability distribution fairly, and in particular, it can help detecting kenetic traps. Furthermore, the above analysis of physical observables, and their comparison with the "gold standard", clearly shows the advantage of satisfying CCE criteria with respect to the SDVG method. However, the use of the CCE convergence metric has its limitations, and can be sensitive for some observables.

First, CCE is a necessary condition, but not sufficient, i.e. CCE alone cannot determine whether a trajectory is "truly" converged. Instead, this metric can only say the simulation is *not* converged. For example, after 6μs of simulation with CSP in implicit solvent, while the entropy is beginning to stabilize, the probability distribution does not pass LZ criteria (Figure 4.7), and the trajectory fails to pass the BCOM criteria (see Figure A.4).

The second caveat of the CCE method, and like many other existing convergence criteria, is that it can only provide an assessment of the presented information as whether adequate sampling has occurred. Any addition to the simulation may explore new phase space, consequently change the underlying distribution, and convergence will be lost in certain metrics. As an example, when applying the convergence test of Hess[225] in which we examine the covariance overlap (Ω) between 6μs and 20μs principal components of the implicit CSP simulation, we see a value of 0.69 (see Figure 4.10). For this metric, a value of zero reflects totally dissimilar fluctuations, and a value of unity is completely identical. Although, the metric is higher at 6μs (time scale needed to pass the CCE criteria) compared to 2.5μs where SDVG criteria passes, the addition of 14μs to the trajectory after claiming convergence by the CCE metric displays a disparity in fluctuations as shown in Figure 4.10.

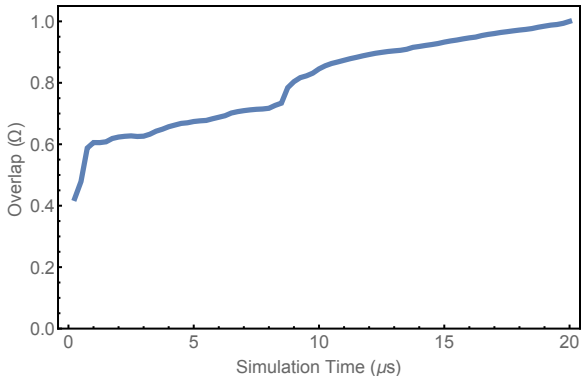

Figure 4.10: Covariance overlap[225] of the implicit CSP trajectory relative to principal components of 20μs long trajectory. The comparison between 6μs (where CCE criteria was satisfied), and the "gold standard" shows the discrepancy in fluctuations, giving $\Omega = 0.69$.

4.2.4 Long simulations show self-consistency is not equivalent to "true" convergence

As has been already alluded to above, the different convergence criteria are only a test for inherent self consistency, albeit at different levels of rigor and complexity. But the question remains, if we claim a trajectory is converged for a given simulation time, using one or *all* of these criteria, does it stay converged if we extend the simulation? If results are unchanged upon further extension, as is expected in the infinite limit, then we can claim "true" convergence. To adequately compare "self consistency check" (SCC) with "true" convergence, it is clear that very long simulations are needed. Here, the hallmark 1 millisecond simulation of native state BPTI[247] will be utilized to test SCC at multiple time scales, and test the hypothesis of "true" system convergence.

Convergence tests were carried out at microsecond iterations of simulation for the first 15μs. The trajectory was clustered with the same algorithm and cutoff ($d_c = 1.25$Å) after each addition of a microsecond to the simulation. After 15μs, convergence was checked at larger iterations, but the same clustering algorithm and cutoff were used. Therefore, we could visualize the convergence behavior of BPTI over many time scales consistently. After just 2μs of simulation, the trajectory satisfies

both SDVG and LZ criteria (Figure 4.11, left and middle panels). Also, the introduced criteria of CCE displays a constant value in the final 1μs (see Figure 4.11, right). If no further information regarding the simulation was known, one could safely claim the native state space for BPTI was adequately sampled within 2μs, i.e. a self-consistent region of the trajectory was found.

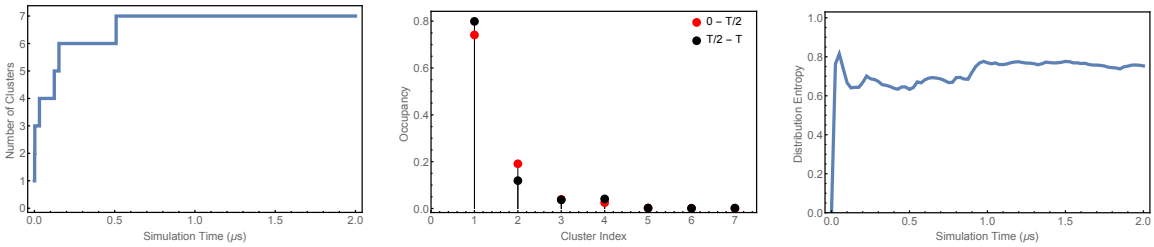

Figure 4.11: Convergence criteria satisfies at 2μs of the 1millisecond BPTI simulation. Left graph shows the evolution of the number of clusters (SDVG method), middle graph shows cluster histogram comparison (LZ criteria), and the right graph shows time dependence of the entropy of the distribution (CCE).

It is worth noting, the BPTI system satisfied both the LZ criteria seen in Figure 4.11, and the BCOM method (see Figure A.5) within a relatively short simulation time. Comparing these results to the earlier analysis of our own implicit CSP system, which is similar in size (58 amino acids in BPTI, and 73 in CSP), displays significant protein-to-protein variation of simulation, and the inability to predict the necessary simulation time scales for adequate sampling.

Interestingly, after discovering the self-consistent regime at 2 microsecond, three more time scales were discovered in which the trajectory discovered unique self-consistent regions (15μs shown in the top panel in Figure 4.12, 30μs in the middle panel in Figure 4.12, and 800μs in the lower panel). The figures show at two of these time points (15 and 30μs), both the SDVG and LZ criteria were satisfied, and the cluster entropy (CCE) also remained stable for a sufficiently long time. Furthermore, the BCOM criteria was well satisfied for both 15 and 30μs time scales (Figure A.5), and error bars were also significantly lower (see Table A.1).

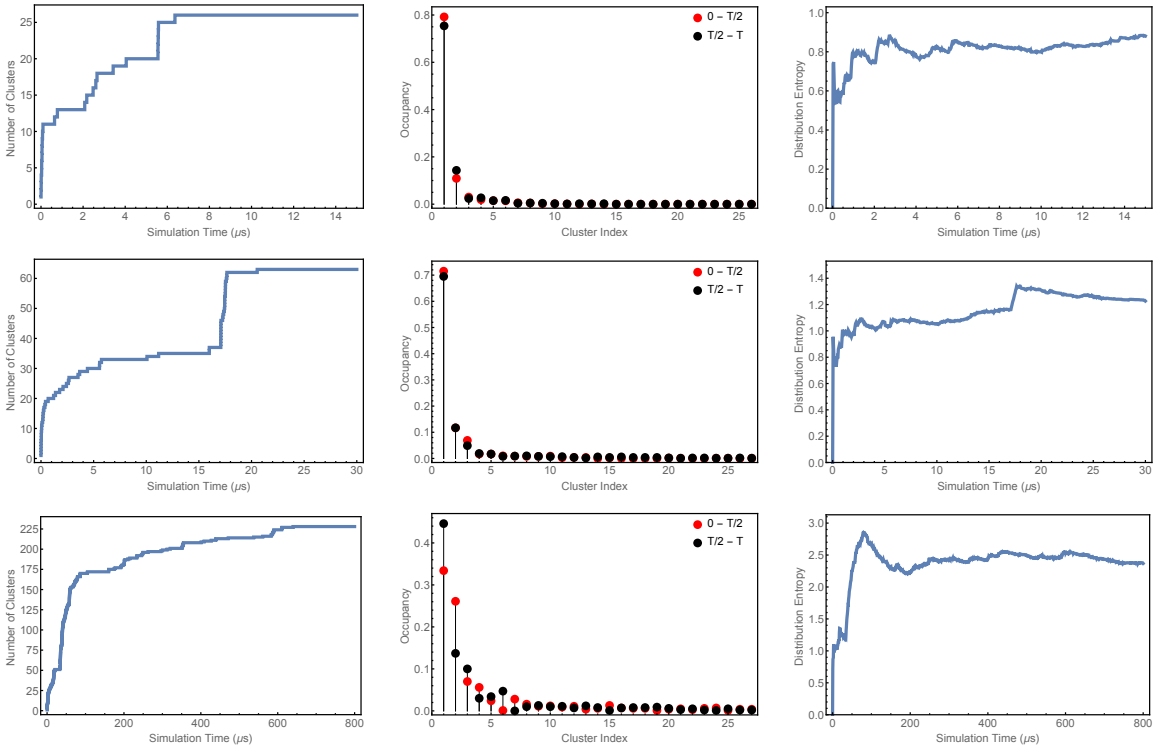

Figure 4.12: Differing time scales of the multiple self-consistent regions in the 1ms BPTI simulation. Top row: 15μs; middle row: 30μs; bottom row: 800μs. At each time scale, the SDVG criteria is satisfied, but as the time scale is increased, more unique states are found. Structural histograms for 15μs and 30μs show convergence, but 800μs is noisy. Cluster entropy at each time cutoff seems to exhibit a constant value.

The discovery of new regions of self-consistency outside of the original 2μs trajectory is alarming. Each extension of trajectory introduces new clusters, and starts exploring states with different weights. The cluster probability distribution plots (middle column of Figure 4.12) show the distribution is changing as more simulation time is considered, specifically the first most occupied state whose weight is decreasing as additional clusters are discovered. As we consider these SCC regions independently, convergence test demands are met, but these regions are not equivalent. With the addition of sampling to the trajectory, we enter into new regions of native state space, after we claimed shorter time scale simulations have converged. We observe a large and fast increase in the number of clusters after 15μs and 30μs

time stamps that could be a result of protein dynamics (see Figure 4.12, left column). For example, the first change in the chirality of the cysteine14–cysteine38 disulfide bridge just after 30μs allowed the simulation to access unexplored structural configurations[247] (see Figure A.6for the chirality dependent states). Therefore, the presence of SCC region is not a guarantee of "true" convergence, and if we had indeed achieved true convergence, no new SCC regions would have been discovered.

Is the BPTI simulation capturing motions that require long time scales to observe? Quick convergence at relatively short time scales (2μs) is indicative of the protein sampling small energy barriers that separate local minima, but as the simulation continues, the large energy barriers that are difficult to overcome in small time scales,[271,272] are passed. Therefore, we observe large increases in the number of clusters after 15 and 30μs, suggestive of sampling newly discovered space blocked by large energy barriers.

Is the observation of multiple regions of self-consistency an artifact of BPTI or a hallmark of long time simulations? Our much smaller, 12μs simulation of CSP in explicit solvent also yielded a region of self-consistent convergence that disappeared as simulation was extended. The first SCC region occurred within 1μs, where all three SCC criteria (SDVG, LZ, and CCE) were satisfied (see Figure 4.13), but not the BCOM criteria (Figure A.7). Both SDVG and CCE methods were satisfied in two other regions, 3 and 12μs, but neither the LZ, nor BCOM criteria were satisfied at these time scales. The number of unique states discovered in the first two regions, 1μs and 3μs, are comparable, but the introduction of these states greatly modifies the cluster probability distribution. If the early convergence is ignored, we again notice the difficulty of passing LZ criteria.

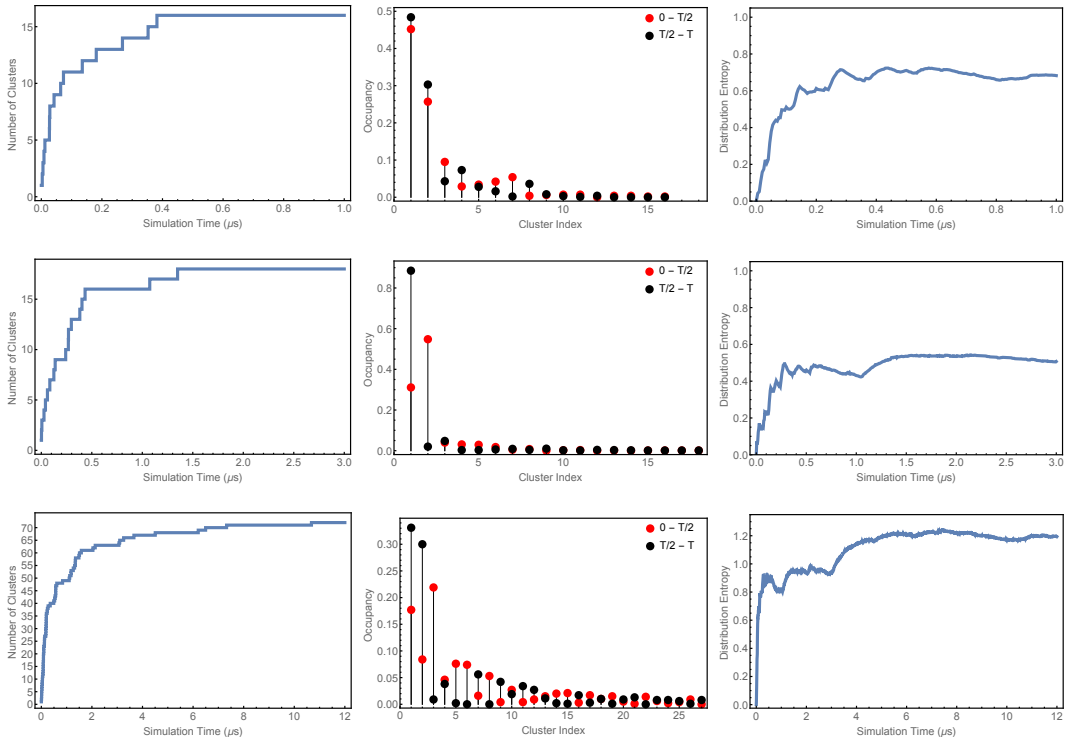

Figure 4.13: Three different SCC regions from the simulation of CSP in explicit solvent. Top row: 1μs shows satisfaction of all three convergence criteria; middle row (corresponding to 3μs of simulation); and bottom row (corresponding to 12μs of simulation) show the simulation trajectory has passed the SDVG and CCE criteria, but not LZ. Within each row, the left graph shows the time evolution of the number of clusters, and the right shows that of the cluster entropy distribution. Middle panel shows the comparison of cluster distributions between the first and the second half of the trajectory.

4.3 Conclusions

Based on our detailed study on multiple protein systems at multiple time scales analyzed using several consistency checks, we make four key conclusions. First, the simulation time required to satisfy different convergence criteria not only vary between different algorithms, but there is also a significant variation between proteins. Second, the self-consistency check criteria of LZ, requiring self similarity in cluster distribution between the first and the second half of the trajectory, and Block covariance Overlap method (BCOM) warn us that native state protein simulations can

be challenging, and can require several tens of microseconds for adequate sampling. Third, we observed that microsecond long native state simulations may encounter traps, and these traps can not be easily detected by mere determination of decorrelation times. We propose monitoring the constancy of cluster entropy (CCE) to detect such traps early, and make a decision to either continue, or abort the simulation. We suggest, in addition to monitoring the constancy of the number of clusters, one must monitor the constancy of the entropy of the cluster distribution. Once simulations have simultaneously satisfied these two conditions, more rigorous tests from the LZ and BCOM methods, and error bar estimations should be imposed as additional quality control. Finally, we reiterate convergence metrics can only ensure if the distribution is being fairly sampled given a trajectory of fixed time, therefore ensuring a self-consistency check. We cannot assess whether the distribution is correctly representing the "true" weights of the discovered states. Any extension, or new simulation of the system can explore new phase space, and self-consistent convergence can be lost. We observed this loss for the BPTI and CSP explicit solvent simulations, where assessing self-consistent convergence did not guarantee "true" convergence as evidenced by existence of multiple regions of self-consistency.

In summary, the notion of "true" convergence remains illusive. But we propose, one must at least ensure both the satisfaction of the SDVG criteria, and stability of the distribution by monitoring entropy of distribution, followed by more rigorous testing of BCOM.

Chapter 5

Elevated Temperature Simulations

While structural information is plentiful,[273] experimental thermodynamic data is scarce,[274] Therefore, in the construction of the thermophilic-mesophilic homolog pairs data set, we were able to gather a total of 13 unique pairings that satisfied our criteria (discussed in subsequent chapter), but experimental denaturing data was not available for all of the homolog pairs modeled. In fact, only three complete pairs, and only 13 of the 26 total proteins simulated have such information (see Table 5.1). Therefore, to i) verify the thermophile system in each studied pair displayed resistance to unfolding at a same increased temperature as their mesophilic homolog, and ii) of equal importance, validate the simulation program correctly modeled systems away from 300K, the homolog pairs were simulated with an elevated thermostat.

5.1 Computational Setup

Starting with coordinates generated from the two-cycle minimization of the 300K simulations, the systems were gradually heated to the elevated temperature from zero over a 250ps period. The choice of temperature was trial-and-error, but had to be

	T_m (°C)	
System	Mesophile	Thermophile
ACP	–	111.5[275]
CheA	–	99.7[276]
CheW	–	–
CheY	57.9[277]	101.0[278]
CSP	57.0[279]	–
HGCS	–	–
HPr	63.5[280]	88.3[281]
NusB	64.7[282]	–
PaiA	–	–
RNaseH	66.0[283]	86.0[283]
RNaseP	66.5[284]	–
Sigma	–	–
Thioredoxin (Oxidized)	88.8[285]	–
Thioredoxin (Reduced)	71.0[286]	–

Table 5.1: Experimental melting temperature data available for the 13 homolog pairs in this study.

shared between the mesophilic-thermophilic constituents within each pair. Initially, all system pairs were heated to 400K and simulated a minimum 200ns. After analysis, if both systems denatured, the thermostat was reduced. However, if both systems retained native-like characteristics, simulations were extended until a total simulation time of 1μs was reached. If both systems were still native-like after analysis, the thermostat was increased. Because of the higher temperature environment, the simulation time step was reduced to 1.6fs, down from 2fs for 300K modeling. Coordinates were written to trajectory files every 10ps.

5.2 Analysis

A major difference between high temperature and 300K simulations is the lack of convergence inspection. For this analysis, the choice in calculated observables needed to intuitively insist that the systems were either still folded, or the systems were denaturing. Ideally, observables calculated from the high temperature trajectories should either closely match those from 300K simulations, or the observables should sharply differ, indicative of native-like properties seen at higher temperatures, or unfolding in response to the elevated temperature environment, respectively. To this end, three separate measures (RMSD, native contact propensity, and secondary structure content) were calculated to quantify nativeness at higher temperatures, and compared to the same measurements found from folded state trajectories. For continuity in comparisons between the two temperature regimes, the coordinates generated upon completion of the 5ns equilibration phase at 300K were selected as the necessary reference state needed in the aforementioned calculations.

The Root Mean Square Deviation (RMSD) measures the average absolute distance between atoms of two optimally superimposed three dimensional protein structures. For this analysis, one of the structures is a fixed reference state to measure structural divergence over a simulation. The RMSD was calculated from trajectory coordinates as a function of simulation time, $\mathbf{r}(t)$, to the above reference state, $\mathbf{r}_{ref}$, as:

$$\mathrm{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i}^{N} ||\mathbf{r}(t)_i - \mathbf{r}_{ref,i}||^2} \tag{5.1}$$

where the summation is over N C_α atoms present in the system. The results of RMSD calculations are shown in the left panel for all system pairs. The opaque curve reflects the 300K simulation RMSD, and the bright curves are from the elevated temperature trajectories.

Discontinuity of native state contacts reflects denaturing within the trajectory. From the folded state reference coordinates mentioned above, a contact was defined as two C_α atoms within 7Å, and all contacts within the coordinates meeting this criteria represented the set of native state contacts. These defined contacts were followed and counted if present as a function of simulation time, and finally the contact count within each frame was normalized by the size of the native state set, giving a fraction Q. For each of the system pairs, this Q metric is plotted in the middle pane. The simulations capturing native states show consistency in the present contacts, while high temperature simulations that capture denaturing display a drop off in this value as contacts break.

Finally, the secondary structure content of the protein was monitored as a function of simulation time. The initial assignment of secondary structure is taken from the above reference state, and secondary structure presented during the simulation is normalized as a percentage retained compared to the reference state. The right panel of each system displays the secondary structure content for native state and elevated temperature simulations. The resultant outcomes of these three observables are plotted below at the chosen elevated thermostat setting for each system.

5.3 Results

5.3.1 ACP

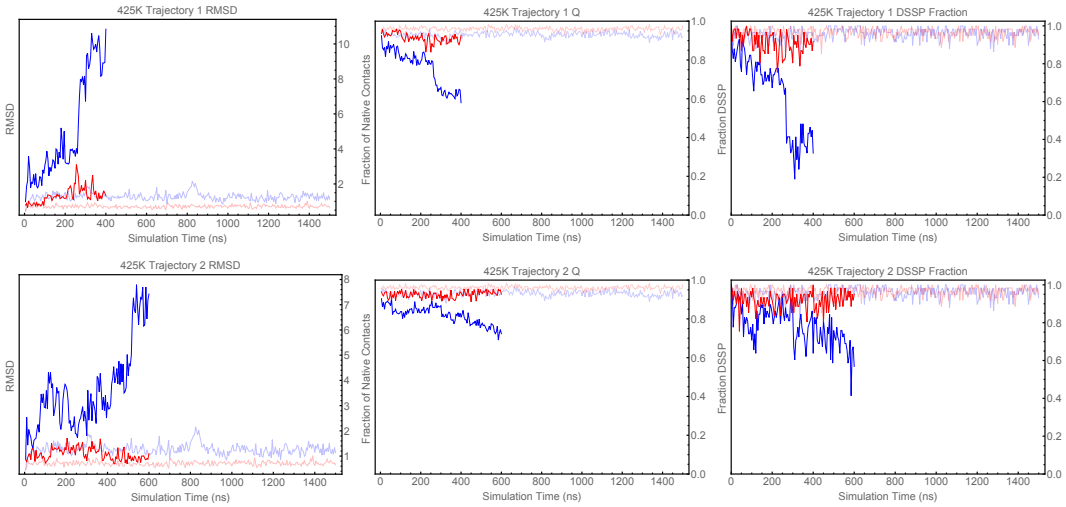

Figure 5.1: Results of 452K simulations of ACP.

5.3.2 CheA

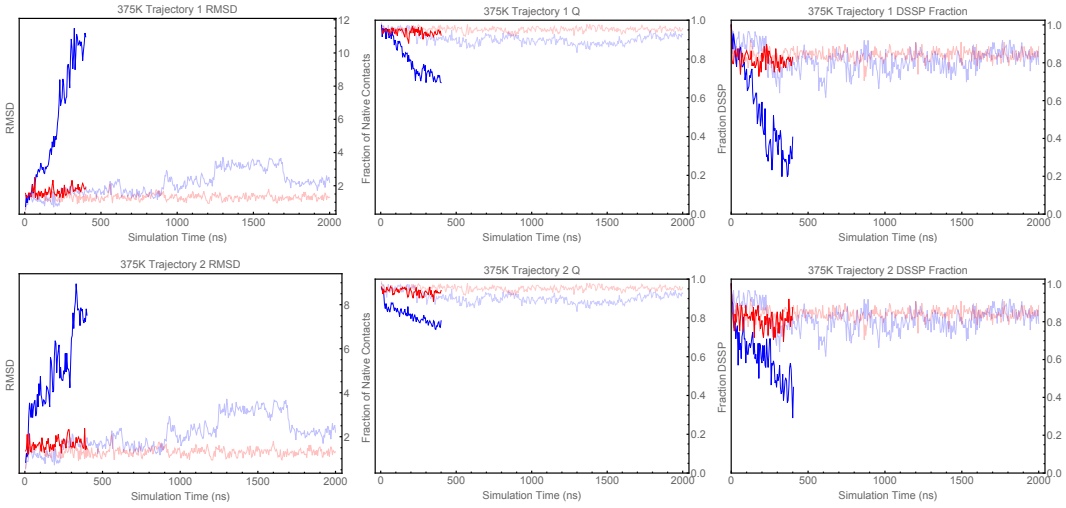

Figure 5.2: Results of 375K simulations of CheA.

5.3.3 CheW

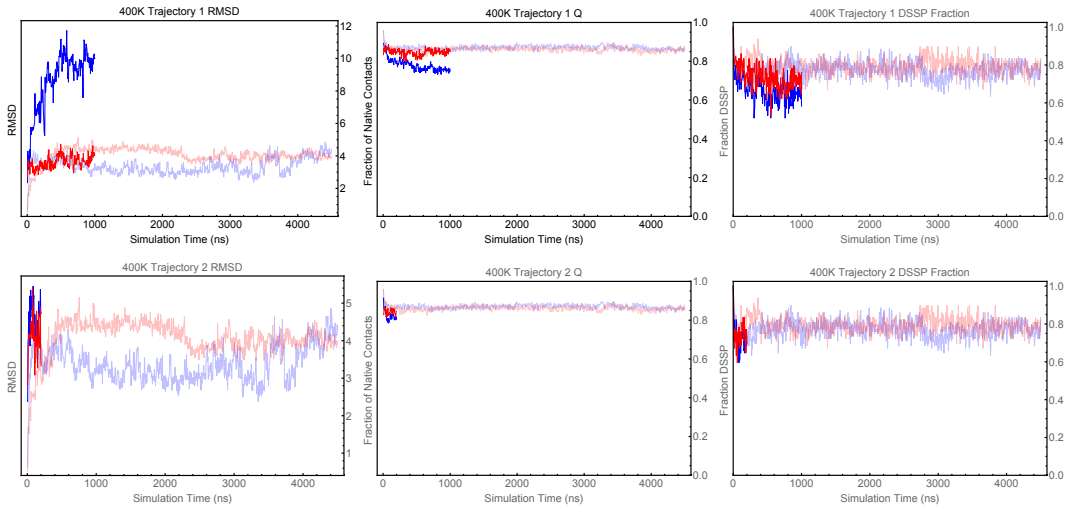

Figure 5.3: Results of 425K simulations of CheW.

5.3.4 CheY

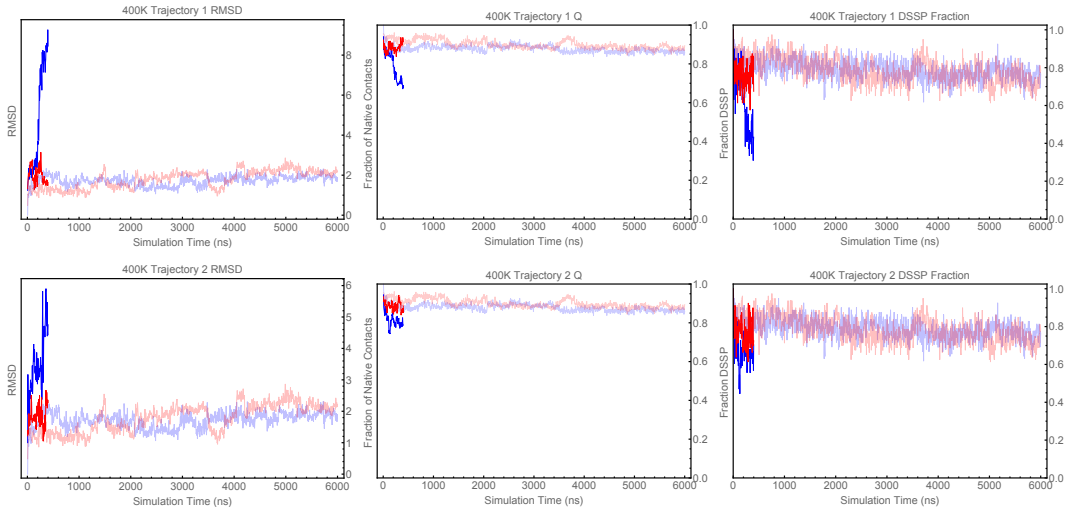

Figure 5.4: Results of 400K simulations of CheY.

5.3.5 CSP

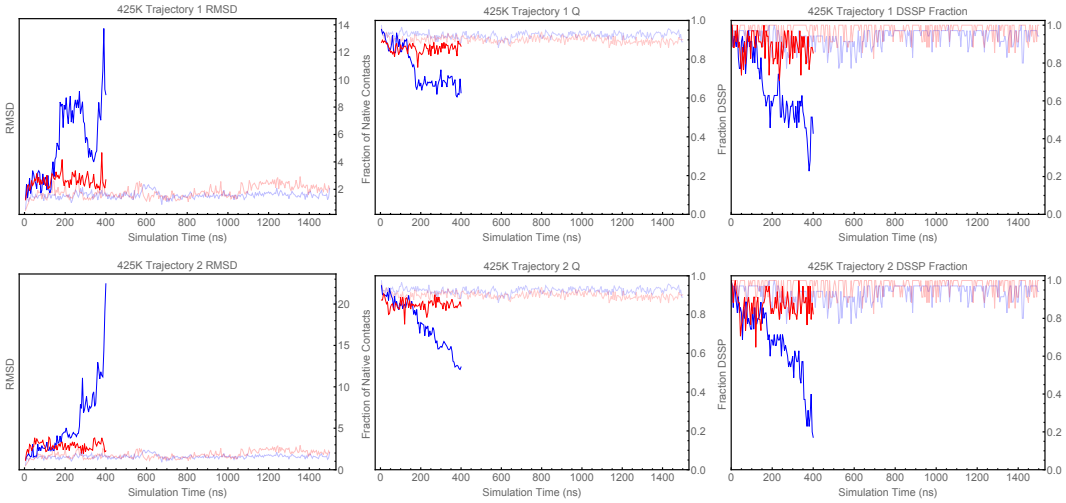

Figure 5.5: Results of 425K simulations of CSP.

5.3.6 HGCS

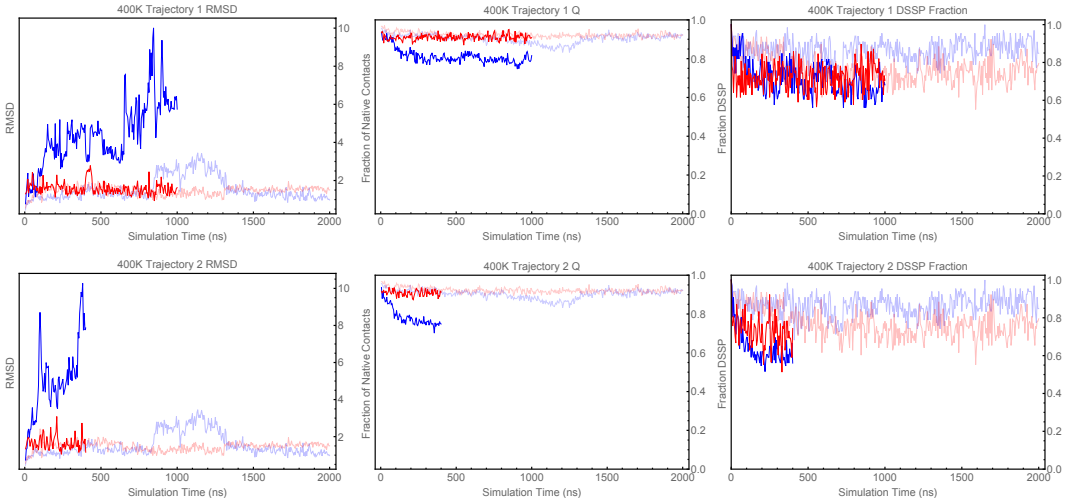

Figure 5.6: Results of 400K simulations of HGCS.

5.3.7 HPr

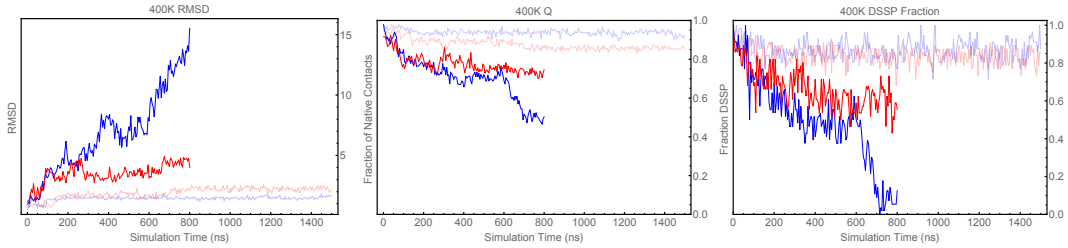

Figure 5.7: Results of 400K simulations of HPr.

5.3.8 NusB

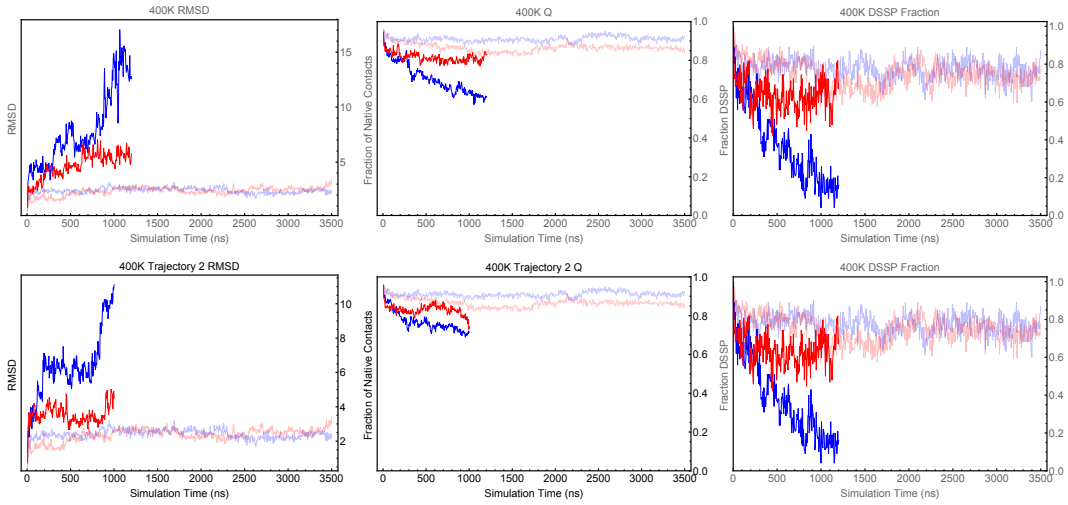

Figure 5.8: Results of 400K simulations of NusB.

5.3.9 PaiA

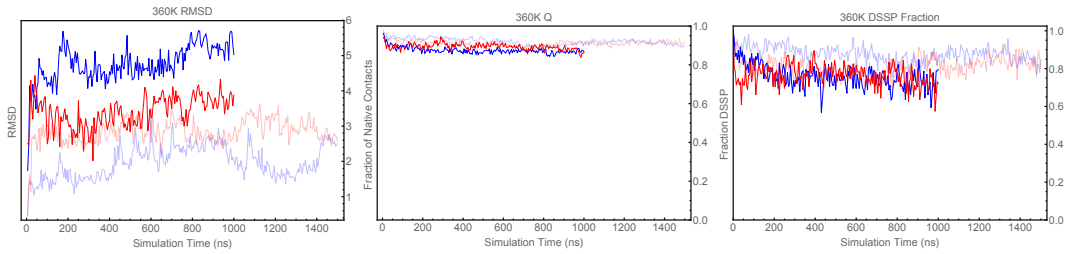

Figure 5.9: Results of 360K simulations of PaiA.

5.3.10 RNaseH

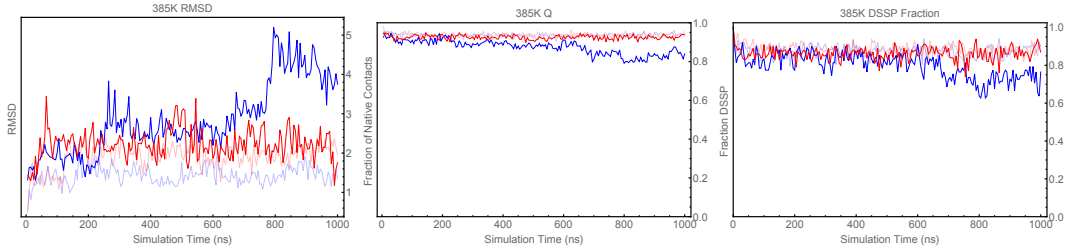

Figure 5.10: Results of 385K simulations of RNaseH.

5.3.11 RNaseP

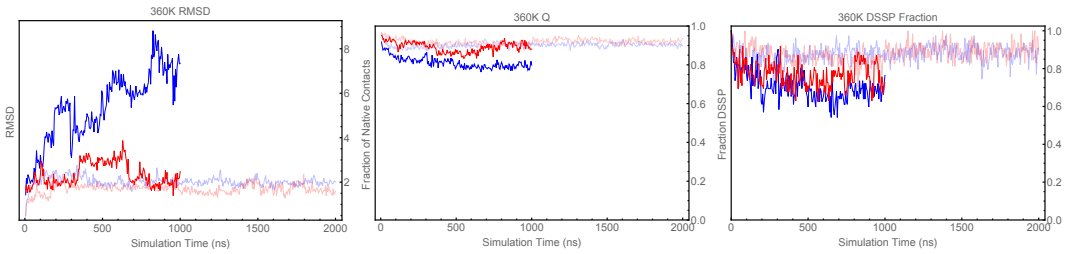

Figure 5.11: Results of 360K simulations of RNaseP.

5.3.12 Sigma

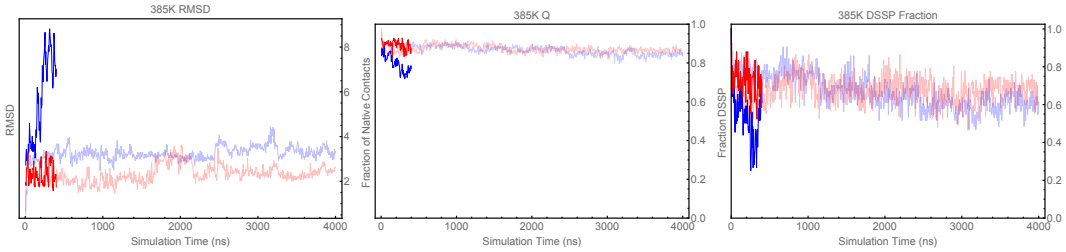

Figure 5.12: Results of 385K simulations of Anti-σ.

5.3.13 Thioredoxin (Oxidized)

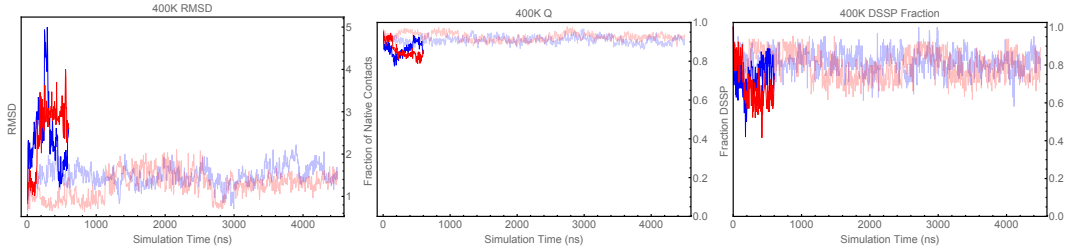

Figure 5.13: Results of 400K simulations of oxidized Thioredoxin.

5.3.14 Thioredoxin (Reduced)

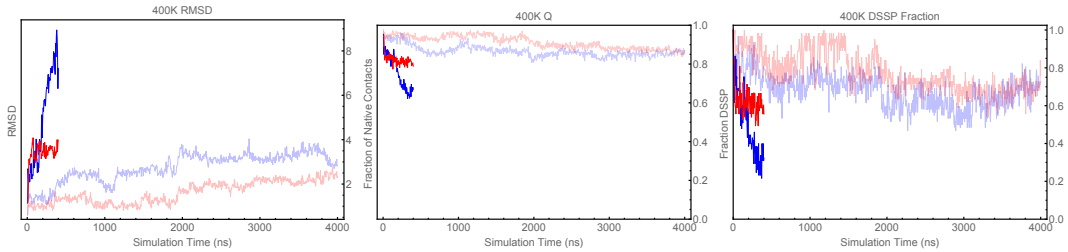

Figure 5.14: Results of 400K simulations of reduced Thioredoxin.

5.4 Conclusions

In the systems presented, the thermophilic homolog retained near native-like properties at elevated temperatures, while the mesophilic component show denaturing in a corroboration of the three observable metrics. Therefore, in the absence of experimental thermodynamic information, we are confident that the systems studied are indeed distinguishable by their response to high temperature environments. Also, the results show that the computational force field used throughout this study is capable of capturing the intrinsic mechanisms leading to resistance of temperature induced unfolding.

Chapter 6

Native State Modeling

Our analysis has first shown that thermodynamic parameters upon unfolding ($\Delta H, \Delta S$, and ΔC_p) are observed to be, on average, reduced in thermophilic proteins, leading to the logical deduction of residual structure in the unfolded state of thermophiles. And while the reduced parameters are insightful, by themselves, they do not depict any physically valuable strategy to enhanced stability. Secondly, motivated by the evidence from the thermodynamic study to the role of unfolded state, we have observed the effects of applying a polymer physics model to the sequences of homologous proteins to capture electrostatic interactions in the denatured state. Specifically, discriminating thermophilic-mesophilic sequence pairs based on exclusively their charge patterning, we have shown that thermophilic proteins tend to have a smaller unfolded state dimension when compared to their mesophilic homolog. Here, we will describe a third inquiry that is in parallel to the thermodynamic and sequence-centric analysis, but based on the structural properties observed from the protein native states.

A majority of the studies focus on particular protein pairs, making it difficult to gain insights about the universality of proposed mechanisms. In other words, to what

extent a mechanism conclusively proven to be responsible in one pair is also operative in another pair? Are there competing mechanisms and if so, is there a primary mechanism? Second, with exception of a few simulation studies,[51,54,81] previous work utilized relatively short simulation trajectories, and paid little attention to quality of sampling measures, leaving the reported observables unreliable, or artifacts of inadequate sampling. Recent work has shown that ensuring convergence of molecular dynamics simulations for the native state are quite challenging.[130]

These concerns are addressed by conducting a global analysis of 13 orthologous protein pairs subjected to long MD simulations of at least 1 microsecond, or longer, as demanded to pass the convergence criteria described in a previous chapter. The exhaustive study of protein structural dynamics between these pairs has revealed several new insights. First, short simulation times typical of previous studies can yield qualitatively different results when compared to long simulations. Also, the results from simply analyzing PDB structures can be misleading, providing further evidence for carrying out careful molecular dynamics simulation to gain correct observations. Second, the inclusion of many pairs highlights the different competing mechanisms, and showcases their relative importance, previously not possible with the one-protein-at-a-time approach. In the majority of pairs studied, thermophilic proteins have a well-connected, attractive electrostatic network, and were associated with more favorable electrostatic interaction energies when compared to their mesophilic counterparts. Furthermore, the adaptive strategy driven by electrostatics may employ several subtle mechanisms, i.e. favorable electrostatic interaction energy by increasing attractive interactions, or reducing repulsive interactions, and/or reducing the desolvation penalty by enhancing the dielectric constants. These strategies are encoded in the sequence and structure by charge composition, and more importantly, the patterning amongst the amino acids. Third, structural dynamics is a consequence

of these strategies, and consequently, thermophilic proteins may have either enhanced or suppressed fluctuations profiles compared to mesophiles, in contradiction to long-standing belief. Finally, alternate strategies of thermophilic adaptation were revealed from a few special proteins pairs where modifying electrostatics does not appear to be a viable strategy for functional reason.

6.1 Methods

6.1.1 Selection of protein pairs

For this comparative study, 13 orthologous protein pairs from mesophilic and thermophilic organisms were selected. System pairs were chosen by first identifying structures from known thermophilic organisms in the PDB, and finding a mesophilic homolog. Exact proteins resolved from differing techniques (e.g. NMR to X–Ray) have been shown to have differing flexibility and dynamics.[287,288] Therefore, we required the mesophile-thermophile protein pairs be resolved from identical techniques. Furthermore, our interest in this study is to explore the properties of independently stable, globular proteins, so pairings had to be biologically monomeric, and have no bound cofactors. Finally, to keep the time necessary for completion of this study manageable, a maximum of 200 amino acids per protein was chosen. The above constraints resulted in 13 protein pairs displayed in Table 6.1 Experimental determined melting temperatures were not available for all homolog pairs to verify heightened thermal resilience, therefore, all systems were simulated at elevated temperatures to ensure the thermophilic protein displayed high temperature resistance (see Figures 5.1 – 5.14).

System	Mesophile		Thermophile	
	PDB	Source Organsim	PDB	Source Organsim
ACP	3BR8[289]	B. Subtilis	1W2I[290]	P. Horikoshii
CheA	1I5N[291]	S. Typhimurium	1TQG[250]	T. Maritima
CheW	2HO9[249]	E. Coli	1K0S[292]	T. Maritima
CheY	3CHY[293]	E. Coli	1TMY[294]	T. Maritima
CSP	1MJC[295]	E. Coli	3A0J [251]	T. Thermophilus
HGCS	3A7L[296]	E. Coli	1ONL[297]	T. Thermophilus
HPr	1POH[298]	E. Coli	1Y4Y[299]	B. Stearothermophilus
NusB	3D3B[300]	E. Coli	1TZW[301]	T. Maritima
PaiA	1TIQ[302]	B. Subtilis	3FIX[303]	T. Acidophilum
RNaseH	2RN2[304]	E. Coli	1RIL[305]	T. Thermophilus
RNaseP	1A6F[306]	B. Subtilis	3Q1Q[307]	T. Maritima
Anti-σ	1AUZ[308]	B. Subtilis	1SBO[309]	T. Maritima
Thioredoxin	2TRX[310]	E. Coli	2YZU[311]	T. Thermophilus

Table 6.1: Homologous mesophile–thermophile protein pairs used in this study. PDB accession codes and source organism are shown.

Here, we will present the function of the selected proteins.

Acylphosphatase (ACP) Catalyzes the hydrolysis of the carboxyl-phosphate bond in acylphosphates.

Chemotaxis Proteins (CheA, CheW, CheY) Chemotaxis is the swimming behavior of bacteria in response to environmental stimuli. The overall movement is the result of alternating tumble and swim phases. In the presence of a chemical gradient bacteria will chemotax, or direct their overall motion based on the gradient. Chemical gradients are sensed through multiple transmembrane receptors, called methyl-accepting chemotaxis proteins. CheA and CheW bind to the receptor. The activation of the receptor by an external stimulus causes

autophosphorylation in CheA at a single highly conserved histidine residue. CheA in turn transfers phosphoryl groups to a conserved aspartate residues in the response regulators CheB and CheY. CheA is a histidine kinase, and does not actively transfer the phosphoryl group. The response regulator CheB takes the phosphoryl group from CheA. This mechanism of signal transduction is called a two-component system, and is a common form of signal transduction in bacteria. CheY induces tumbling by interacting with the flagellar switch protein FliM, inducing a change from counter-clockwise to clockwise rotation of the flagellum. Change in the rotation state of a single flagellum can disrupt the entire flagella bundle and cause a tumble.

Cold Shock Protein (CSP) The cold shock response in E. Coli follows an abrupt shift in growth temperature from 37°C to 10°C, inducing a lag phase in cell growth of 4-5 hours. It is accompanied by a severe reduction in protein synthesis. As a consequence of this cold shock, the relative rate of production of at least 14 cold shock proteins is increased. For 13 out of the 14 proteins, the increase is 2 to 10-fold. The synthesis of the major cold shock protein, CSPa, increases at least 100-fold, reaching a level of $> 10\%$ of total protein synthesis at 10°C. CSPa was shown to act as a transcriptional activator of the hns and gyrA genes encoding two other cold shock proteins.

Glycine Cleavage System H-protein (HGCS) The glycine cleavage system is a series of enzymes that are triggered in response to high concentrations of the amino acid glycine. The same set of enzymes is sometimes referred to as glycine synthase when it runs in the reverse direction to form glycine. The glycine cleavage system is composed of four proteins: the T, P, L, and H-protein. They do not form a stable complex, so it is more appropriate to call it a "system"

instead of a "complex". The H-protein interactions with the three other proteins, and acts as a shuttle for some of the intermediate products in glycine decarboxylation

Histidine-containing phosphocarrier protein (HPr) A key component of the phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS), a major carbohydrate transport system in bacteria. The PTS catalyses the phosphorylation of sugar substrates during their translocation across the cell membrane. The mechanism involves the transfer of a phosphoryl group from phosphoenolpyruvate (PEP) via enzyme 1 (E1) to enzyme 2 (E2) of the PTS system, which in turn transfers it to a phosphocarrier protein (HPr). A conserved histidine in the N-terminus of HPr serves as an acceptor for the phosphoryl group of E1.

N-Utilization Substance Protein B (NusB) Transcription anti-termination protein that participates with 30S ribosomal subunit protein S10 in processive transcription anti-termination. Anti-termination is the prokaryotic cell's aid to fix premature termination of RNA synthesis during transcription. Anti-termination factor enable polymerase read-through.

PaiA N-acetyltransferase activity toward polyamines, including spermine and spermidine. N-acetyltransferase is an enzyme that catalyzes the transfer of acetyl groups from acetyl-CoA to arylamines

Ribonuclease (RNaseH, RNaseP) A type of nuclease that catalyzes the degradation of RNA into smaller components. A non-specific endonuclease that catalyzes the cleavage of RNA via a hydrolytic mechanism. Endonuclease are enzymes that cleave the phosphodiester bond within a polynucleotide chain.

In DNA replication, RNaseH is responsible for removing the RNA primer, allowing completion of the newly synthesized DNA. RNaseP cleaves off an extra (precursor) sequence of RNA on tRNA molecules.

Anti-Sigma Factor Antagonist (Anti-σ) In the regulation of gene expression in prokaryotes, anti-sigma factors bind to sigma factors and inhibit transcriptional activity.

Thioredoxin Redox protein that antioxidants by facilitating the reduction of other proteins by cysteine thiol-disulfide exchange

6.1.2 Computational setup

All non-protein atoms, including crystallographic waters and structural hydrogens, were removed from the PDB structures. The CheA mesophilic structure varied from the thermophilic structure in that it contained a linker region in the form of an additional alpha helix.[250] To maintain direct comparison, this linker region was removed from the structure. The protonation state of histidine side chains were predicted with H++ (http://biophysics.cs.vt.edu/H++).[252–254] Using TLEAP,[312] protein hydrogens, a cubic box of TIP3P[255,256] water molecules with a 9Å buffer region, and neutralizing ions was added. The system solutions were energy minimized for two cycles using the SANDER module of AMBER 14.[313] The first cycle held the protein constant with harmonic restraints to minimize poor steric positions of the solvent. The restraints were lifted, and a second cycle of minimization allowed the entire solution to adjust to a local minima.

All simulations were performed with the GPU-accelerated PMEMD module of AMBER 14[257] using the ff99SB force field,[258] periodic boundary conditions, and SHAKE[259] to constrain the lengths of bonds that include hydrogen atoms to their

equilibrium lengths. A 9Å cutoff was chosen for the direct sum, non-bonded interactions, and the Particle Mesh Ewald method[260–262] was used to calculate long-range electrostatics. After energy minimization, solvent molecules were heated gradually to 300K over a 100ps period at constant volume. The entire system was equilibrated for 5ns at constant temperature and pressure, allowing the volume to change to adjust the system's density. Simulations were performed at constant temperature (300K) and pressure (1atm) by employing the Langevin thermostat and Berendsen barostat, respectively, with a 2fs time step, and structural conformations saved every 10ps.

6.1.3 Convergence inspection

To ensure our simulations returned reliable, and replicable observable measurements, a convergence criteria was implemented. Convergence was tested by monitoring the discovery of new clusters, and the stability of the cluster distribution entropy (described in the proceeding chapter)[130] over the trajectory's final one-fourth simulation time block, or 500ns, whichever was least. If, within the final time block, the trajectory found less than 5% of its total number of clusters, and the largest percent difference in cluster entropy was less than 5%, the simulation was considered converged.

The cluster command within the CPPTRAJ module of AMBER was used to cluster all simulation trajectories.[264] The agglomerative, average-linkage algorithm was employed[265] utilizing an RMS distance metric to compare structures. In summary, each structure begins in its own cluster, and pairs of clusters are combined only if the average distance (RMSD) between all points of inter-cluster elements are less than the preselected distance cutoff, d_c. For this direct comparison study, we demanded the selection of d_c match across mesophilic-thermophilic pairs.

The protein pairs were simulated for an initial minimum of 1μs, and the resulting trajectories were tested for convergence. If convergence was not satisfied, the simulations were extended for an additional 500ns, and convergence retested. This cycle of simulation and convergence testing was repeated until both systems of a given mesophilic-thermophilic pair passed our criteria with identical simulation times. The convergence was tested by following the discovery of new cluster and the stability of the cluster distribution entropy.[130]

6.1.4 Representative atoms for the charged groups, and the calculation of their interaction probabilities

For calculation of interactions between charged amino acids, representative atoms for each charged group were selected (α-amino N, α-carboxyl C, Asp Cγ, Glu Cδ, His Cϵ1, Lys Nζ, and Arg Cζ) ,[73,156] and the appropriate charge of $\pm e$ was assigned to only these atoms. These atoms best represented the geometric center of charge for these particular side chains.

The distances between all possible $N(N-1)/2$ pairings of N ionic representative atoms was calculated for each frame of the simulation trajectory using CPPTRAJ. Each pairing's distance set was analyzed to find the fraction of frames with a distance d less than, or equal to, a predefined critical distance d_c, i.e. $d \leq d_c$, giving the probability p that a particular ion pair interacted. For a particular value of p_c, the number of attractive interacting ion pairs (edges) was found and normalized by the total number of charged residues (vertices) present in the system, giving an edge-to-vertex ratio (ζ) synonymous to a measure of connectedness in the attractive ion interaction network. Figuress 6.1 and 6.2 show the ζ ratio as a function of the cut-off probability p, and demonstrates the high level of connectivity in thermophiles.

6.1.5 Calculation of protein interior dielectric constant

From the protein dipole moment fluctuations calculated from the MD simulations, and application of the Frölich–Kirkwood theory of dielectrics, the internal dielectric constant of our protein systems were determined as[86,87]

$$\frac{\langle \Delta M_p^2 \rangle}{kTr_p^3} = \frac{(2\epsilon_w + 1)(\epsilon_p - 1)}{(2\epsilon_w + \epsilon_p)} \tag{6.1}$$

where $\langle \Delta M_p^2 \rangle$ represents the time average of dipole moment deviations throughout a system simulation, r_p is the spherical radius of the protein, taken to be $\sqrt{5/3}$ of the radius of gyration, and ϵ_w and ϵ_p are the dielectrics of the solvent and protein interior, respectively. The deviation of the dipole moment is origin dependent for systems with net charge, and for uniformity, all systems in this study used the protein center of mass in their calculations. The protein dipole moment and radius of gyration for each structural frame was calculated with CPPTRAJ,[264] and analyzed with in-house software.

6.1.6 Calculation of average RMSF

Fluctuations about the structural average ($RMSF$) are a measure of localized protein flexibility, and can be easily calculated from simulation data. Using CPPTRAJ,[264] the trajectory structures are superimposed to eliminate rotational and translational effects arising from the simulation, an average structure found, and the fluctuations were calculated from the start of the simulation to an accumulation of simulation time (T) in successive intervals up to the total simulation time. The fluctuations ($RMSF_i$) of the C_α atom in the ith amino acid about its average position

were calculated as

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(C_{\alpha,i}(t) - \langle C_{\alpha,i}\rangle\right)^2} \tag{6.2}$$

giving N $RMSF_i$ values for a system of N amino acids. To remove the effect of unassigned tail regions, residues not assigned to a secondary structure at the N terminus (residue 1 up to just before the first secondary structure) and C terminus (just past the last secondary structure's final amino acid to residue N) at the start of the simulation were removed from the calculation. Loop regions connecting secondary structure regions were kept for the calculation. The remaining $RMSF_i$ values are averaged, giving $\langle RMSF \rangle$ as a function of accumulative simulation time for a given protein.

6.2 Results and Discussion

Long time scale simulations were subject to rigorous sampling criteria, and detailed analysis revealed several important differences in the native state properties between thermophilic and mesophilic protein pairs.

6.2.1 Attractive ionic networks are better connected in thermophiles

Previous studies have indicated that thermophilic protein native states have more ion pairs compared to their mesophilic counterparts.[314] However, these studies are often and primarily based on the static PDB structure, and ignore structural dynamics. Contacts identified from the PDB structures may be only transient, i.e. they may disappear, and new contacts may emerge over a simulation trajectory. Furthermore, the total number of ion-pairs does not give us insights about the *connectivity* be-

tween all ion-pairs. The ion network topology can be important, as increasing the total number of attractive ion-pairs does not necessarily enhance stability due to the competing destabilization effect arising from the desolvation penalty of charged groups.[13] The desolvation penalty is expected to increase with the total number of charged side chains, therefore, an efficient design strategy is to increase the number of ionic interactions while keeping the total number of charges relatively unaltered. This can be achieved by sharing charged residues within the ionic interactions,[13,65] leading to better *connected* networks amongst charged groups. As a result, the attractive electrostatic interaction can become more favorable, and eventually overcome the desolvation penalty. Network topology analysis of non-covalent connections has emphasized the importance of hubs in thermophilic proteins,[315] while network analysis highlighted the role of the largest rigid cluster in thermophilic proteins.[111]

To explore the role of connectivity in a quantitative manner, a simple metric ζ was defined as the ratio of the number of ion pair interactions (edges) to the number of charged residues (vertices). Higher values of ζ implied more ion pairs per charged residues, and better connectivity. Furthermore, the dynamic nature of ion-pair formation was quantified by its probability (p), and was determined by monitoring the fraction of simulation frames in which the distance (d) between two oppositely charged side chain representative atoms was within a defined critical distance d_c, i.e. $d \leq d_c$. The importance of quantifying contact probabilities in this manner have proven useful in other studies of protein folding mechanisms as well, for example, in distinguishing two-state folder from a three-state folding protein.[270] For a given p, all contacts between oppositely charged groups with a probability of at least p were considered. When $p = 1$, only the most stable contacts contribute, and in the opposite limit of very small p, all contacts, including the most unstable, were expected to contribute. The resilience of these networks and their connectedness is quantified by computing

ζ as a function of p. For the majority of the pairs (10 of 13), thermophilic proteins formed better connected, and more stable networks among oppositely charged amino acids compared to their mesophilic counterparts (see Figure 6.1). RNaseP, HGCS, and Thioredoxin are the only three pairs that does not show any significant difference between thermophile and mesophile (see Figure 6.2). PDB based analysis will not reveal such pattern due to static structure. Short simulations, on the other hand, may suffer from sampling issues yielding unreliable values of p thus highlighting the importance of careful MD simulations subjected to strict convergence test.

This finding is consistent with earlier studies that indicated the major role of electrostatics in protein stability.[18,62,65,67,114,156] Mutational studies have directly shown enhanced thermostability can be achieved by forming large networks of ionic interactions.[65] This is linked to the hypothesis that better connected ion-pair networks may impart greater resilience to thermophilic proteins at high temperature.[316] A possible consequence of greater electrostatic resilience is an increased unfolding barrier, and therefore, a slower unfolding rate which can give rise to higher stability, assuming the folding rate remains unaltered. This has been supported by the experimental observation that the unfolding rate in thermophilic rubredoxin is slower compared to its mesophilic counterpart.[317] Recent studies on trigger factor has also shown similar trend.[318] The role of salt bridges in imparting higher stability at the cost of low activity at low temperature has been demonstrated by mutagenesis experiment.[97]

6.2.2 Thermophilic proteins have more favorable electrostatic interaction energies in their native state

The results above indicate that *attractive* interactions between oppositely charged side chains are more favorable in thermophilic native states compared to their mesophilic

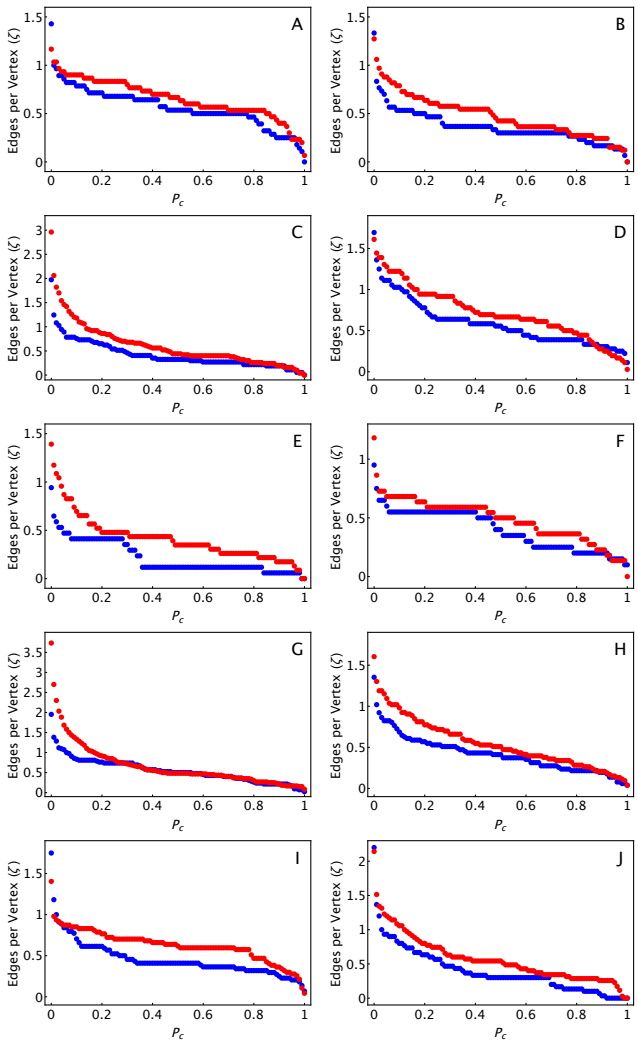

Figure 6.1: Edges per vertex (defined as ζ) as a function of cut-off probability (p_c) for 10 homologous protein pairs: (A) ACP, (B) CheA, (C) CheW, (D) CheY, (E) CSP, (F) HPr, (G) NusB, (H) PaiA, (I) RNaseH, and (J) Anti-σ. The level of connectivity (the ratio of edges to vertices) is systemically greater in the thermophile protein (red data) with respect to the mesophilic pair homolog (blue data) as the cutoff probability is increased. See Figure S1 for the proteins where there is no difference between thermophile and mesophile.

counterparts. But, what is the contribution from the *repulsive* interactions between similarly charged amino acids? Do they offset the attractive interaction, and alter the trend for the overall net electrostatic interaction shown above? Moreover, electrostatics are long-ranged, therefore, the connectivity alone does not give a precise measure of the attractive electrostatic interaction. To further quantitate this effect,

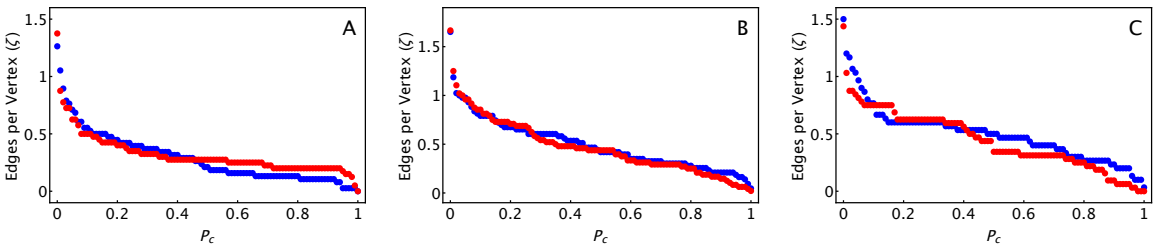

Figure 6.2: Attractive interactions per charged residue (edges per vertex, ζ) as a function of probability cut-off for (A) HGCS, (B) RNaseP, and (C) Thioredoxin. These systems did not show a discernible difference in the ζ metric.

and provide a more comprehensive picture, the total electrostatic interaction energy was computed from all charged residues in the folded state of the protein, but due to the complicated, non-spherical geometry of the folded state, DelPhi[319] was used to compute the interaction energy. The electrostatic interaction energy in the folded state was defined as

$$G^f_{int} = \frac{1}{2}\sum_k q_k \phi_k \tag{6.3}$$

where q_k is the kth test charge under consideration, and ϕ_k is the potential at the kth charge location resulting from all other charges in the system. The factor of 1/2 ensures a proper summation, i.e. no double counting. Note the self energy terms, those related to the solvation penalty have not been included in this calculation, but are discussed in detail below.

The folded state electrostatic free energy of interaction (G^f_{int}) is the summation of the coulombic energy (G^f_{coul}) of interacting protein ionic charges and the polarization energy (G^f_{pol}) arising from the charges set within the protein boundary having a different dielectric constant than that of solvent. The strength of the polarization term is determined by the distance of the charges from the protein boundary, the shape of the boundary, and the protein and solvent dielectric constants. Given a protein structure, and respective charge and radii libraries, DelPhi[319] will calculate the total coulombic energy and a corrected reaction field energy (CRFE). From our in-

house benchmarking with systems with exact analytical solutions (see Appendix), we have deduced that the CRFE represents the combination of the polarization (cross) energy and the total solvation (self) energy for the structure given. The total solvation energy of the folded state is found in an iterative manner by summing the solvation energies for each individual ionic residue in the presence of the entire protein, but with charges on all other remaining ionic amino acids turned off.[320] The difference of the CRFE and the total solvation energy produced the polarization energy for the folded state, and by combining this energy with the folded state coulombic energy, we have deduced G^f_{int}. This procedure was followed for two values of internal dielectric constants representing the dielectric constant inside the folded protein boundary. Here, we used a generalized internal dielectric of $\epsilon_{int} = 2$, and a dielectric constant calculated from the MD simulations that represents the protein polarizability $\epsilon_{int} = \epsilon_p$ (see Methods section), and a fixed dielectric constant of 78 for the external solvent. Also, all calculations were conducted utilizing a charge library that placed a single charge of $\pm e$ at the representative atoms discussed in the Methods section, with all other atoms were set to neutrality. Similar calculations were presented by Zhou[13] using spherical geometry, and Brooks2004 for non-spherical shape.[87]

For this analysis, we wanted to include information from the long time-scale simulations to account for the fluctuations in the folded state, but the number of trajectory frames representing the folded state is quite large (100,000 frames per microsecond of simulation), and thusly, the electrostatic energy calculation became quite resource demanding if we attempted to use every simulation frame. The following compromise was found. Instead of utilizing every frame from a simulation trajectory, we extracted the representative structures of the first n most occupied ith clusters that accounted for 90% of the total simulation frames ($\sum_i^n p_i > 0.90$). The electrostatic energies for each system's cluster representative structures were calculated by following the

protocol outlined above with DelPhi, weighted per each cluster's probability, and finally summed. The difference in the interaction energies between thermophile and mesophile was defined as $\Delta G^f_{int} = G^{f,thermo}_{int} - G^{f,meso}_{int}$; where $G^{f,thermo}_{int}$ is the electrostatic interaction energy of the folded state of the thermophilic protein, and $G^{f,meso}_{int}$ represented the mesophilic protein. Values of ΔG^f_{int} for the protein pair systems are reported in Table 6.2.

System	$\Delta G^f_{int}(\epsilon_{int}=2)$	$\Delta G^f_{int}(\epsilon_{int}=\epsilon_p)$	ϵ_p^{thermo}	ϵ_p^{meso}	$\Delta\Delta G_{elec}(\epsilon_{int}=\epsilon_p)$
ACP	-53.63	-6.36	16.9	20.7	-3.29
CheA	-37.28	-39.68	16.3	42.6	-10.08
CheW	-31.55	-22.21	58.1	38.7	-32.22
CheY	-71.24	-14.71	27.5	24.3	-18.70
CSP	-45.20	-8.14	40.6	23.8	-11.44
HGCS	-28.74	-19.30	18.1	24.9	-1.76
HPr	-10.52	-5.05	21.1	17.6	-7.05
NusB	-50.12	-7.07	46.0	27.9	-21.52
PaiA	-53.82	-6.60	46.7	25.1	-26.58
RNase H	-36.55	-9.06	14.8	14.3	-0.91
RNase P	21.46	23.52	46.6	37.5	17.30
Anti-σ	-48.45	-7.08	38.4	36.5	-5.18
Thioredoxin	9.70	3.91	21.6	20.9	4.12

Table 6.2: Electrostatic energies calculated from structural ensembles taken from simulation trajectories. The difference in electrostatic interaction energy (columns 2 and 3) between thermophilic and mesophilic ($\Delta G^f_{int} = G^{f,T}_{int} - G^{f,M}_{int}$) homologs at specified internal dielectrics. Dielectric constants calculated using protein dipole moment fluctuations from MD simulations (columns 4 and 5). All values of ΔG^f_{int} and $\Delta\Delta G_{elec}$ have units of kT.

For $\epsilon_{int} = 2$, the thermophilic protein native states have, for 11 of the 13 pairs, more favorable electrostatic interaction energy compared to mesophilic counterparts. The only exceptions were the RNaseP and Thioredoxin systems (see the second col-

umn of Table 6.2). The resulting calculations differed from calculations utilizing only static PDB structures,[78] and are shown in Table 6.3. The degree of quantitative difference in interaction energies calculated from MD simulation trajectories and static PDB structures is apparent by comparing Tables 6.2 and 6.3. Interestingly, for three protein systems (CheW, PaiA, and RNaseH) even a qualitative difference is observed. PDB based structural analysis for these three system pairs predicted mesophilic proteins with more favorable electrostatic interaction energy, in direct contradiction with the simulation based analysis, reiterating the need for incorporating protein dynamics in such calculations.

System	$\Delta G^f_{int}(\epsilon_{int}=2)$	$\Delta G^f_{int}(\epsilon_{int}=\epsilon_p)$	ϵ_p^{thermo}	ϵ_p^{meso}	$\Delta\Delta G_{elec}(\epsilon_{int}=\epsilon_p)$
ACP	-24.48	0.20	16.9	20.7	9.20
CheA	-2.46	-37.08	16.3	42.6	-23.27
CheW	44.21	-16.73	58.1	38.7	-26.73
CheY	-59.18	-17.79	27.5	24.3	-24.39
CSP	-38.35	-6.86	40.6	23.8	-9.66
HGCS	-45.14	-22.79	18.1	24.9	-8.89
HPr	-22.69	-10.86	21.1	17.6	-10.94
NusB	-135.44	-15.46	46.0	27.9	-27.10
PaiA	16.20	1.47	46.7	25.1	-14.93
RNaseH	15.21	1.58	14.8	14.3	14.41
RNaseP	13.54	29.54	46.6	37.5	24.81
Anti-σ	-90.70	-12.29	38.4	36.5	-12.17
Thioredoxin	51.69	9.94	21.6	20.9	9.08

Table 6.3: Calculation of electrostatic energies using strictly the PDB structure for each system. The difference in electrostatic interaction energy (columns 2 and 3) between thermophilic and mesophilic ($\Delta G^f_{int} = G^{f,T}_{int} - G^{f,M}_{int}$) homologs at specified internal dielectrics. Dielectric constants calculated using protein dipole moment fluctuations from MD simulations (columns 4 and 5). All values of ΔG^f_{int} and $\Delta\Delta G_{elec}$ have units of kT.

A similar trend is observed by introducing the dielectric constant of the protein interior (ϵ_p) estimated from the molecular dynamics simulation. Electrostatic interaction energies in the native state were more favorable in thermophilic proteins compared to their mesophilic ortholog in 11 of 13 pairs. As seen previously, the RNaseP and Thiroredoxin systems were the only two pairs that were exceptions to this finding. The discrepancy between PDB and MD based analysis is again observed, particularly with the PaiA and RNaseH systems where the trends are reversed. For the ACP pairing, PDB based calculation shows almost no difference in the electrostatic interaction energy between thermophilic and mesophilic constituents, in contrast to the MD generated trajectory based calculation (compare the third column between Tables 6.2 and 6.3). These findings are consistent with several experimental studies demonstrating the importance of charged residues and improved electrostatic interaction in protein stability.[67,73–75,321,322] Particularly, Makhatadze and colleagues have indicated the role of electrostatic interaction from charged *surface* residues.[67,71,321–323]

6.2.3 Role of solvation

It is now clear that thermophilic protein native states have more favorable electrostatic interactions with respect to their mesophilic counterparts. However, enhancement in favorable electrostatic interaction is often associated with a competing effect of increased desolvation penalty.[13,83,87] The desolvation penalty ($\Delta G_{solv} = G^f_{solv} - G^u_{solv}$) is the difference in solvation energy of the charged groups in the folded (G^f_{solv}) and the unfolded state (G^u_{solv}). In general, charged groups are better solvated in the unfolded state than in the folded state, giving rise to positive ΔG_{solv} values. The desolvation penalty is expected to increase with the number of charges in the system, and therefore, the desolvation penalty is of particular concern when favorable electrostatic interactions in the native states are achieved by an increase in the

total number of charges. Brooks and colleagues demonstrated the possibility that thermophilic proteins may lower the desolvation penalty by increasing the interior dielectric constant of the protein.[87] Elcock[83] has noticed that although the desolvation penalty for thermophiles can be higher than mesophiles at room temperature, it is significantly lower at high temperature where the dielectric mismatch between protein interior and the solvent dielectric is reduced. Motivated by these observations, the effect of desolvation among thermophilic and mesophilic protein pairs was further quantified by calculation of the solvation energy in the folded (described above) and unfolded states. The unfolded state solvation energy was first calculated by considering each amino acid surrounded by its two nearest sequential neighbors, where the charges on the nearest neighbors are turned off. Using protein sequence information, the pentamers were constructed in TLEAP,[312] and energy minimized with restraints on the backbone atoms to strictly remove bad side chain steric clashes. The electrostatic solvation energy was calculated with DelPhi[319] for each constructed pentamer, and the summation of these solvation energies represented the unfolded state solvation energy.[320]

To compare the results of summing the constructed pentameric contributions, a second method was developed to calculate the solvation energy of the unfolded state. The charged residues were again set in the center of a pentamer, but the surrounding ± 2 amino acids were selected by iterating over all possible combinations of large amino acids. The possible surrounding amino acids consisted of ASN, GLN, HIS (HIE), LEU, MET, PHE, THR, TRP, TYR, and VAL, giving 10,000 pentameric combinations. Termini contributions were found by constructing trimers that consisted of an N- or C-terminal Alanine fixed to the respective end of all possible trimeric combinations of amino acids listed above, giving 100 trimers per terminus. Construction and calculation of the solvation energy for each charged group followed the same

method as above. The solvation energies found for each charged group were averaged over all possible combinations, and are shown in Table 6.4 with their variance. The total unfolded state solvation energy is found by summing the contributions from each charged amino acid in the sequence. Unfolded state solvation energies found from these average values show less than 1% difference when compared to the energies calculated from the previous method above. With these average values, this approach offers a high throughput, but very accurate calculation of unfolded state solvation energies, and can be utilized for fast estimation of solvation energies in the unfolded state.

Charged group	$\langle U_{solv} \rangle$ $[kT]$
N-termini	-62.16 ± 0.02
ASP	-54.58 ± 0.27
GLU	-56.54 ± 0.76
ARG	-58.06 ± 0.10
LYS	-66.74 ± 0.29
HIS (HIP)	-57.29 ± 2.22
C-termini	-53.85 ± 4.21

Table 6.4: Average solvation energy of charged groups found from combinatoric constructions of pentamers (amino acids) and trimers (terminal ends). Amino acids were centered in pentamers, while terminal groups were placed at their respective end.

To quantitate the net effect of electrostatics upon folding, ΔG_{elec} was calculated by combining the electrostatic interaction energy and the desolvation penalties as $\Delta G_{elec} = G^{f}_{int} + G^{f}_{solv} - G^{u}_{solv}$. The interaction energy in the unfolded state was assumed negligible, and its contribution ignored to simplify calculations. The comparison of ΔG_{elec} values between thermophile and mesophile pair constituents were calculated as $\Delta\Delta G_{elec} = \Delta G^{thermo}_{elec} - \Delta G^{meso}_{elec}$, where superscript *thermo* and *meso* denote thermophilic and mesophilic constituent, respectively. Comparison be-

tween thermophilic and mesophilic pairs displayed a majority of the protein pairs $\Delta G_{elec}^{thermo} < \Delta G_{elec}^{meso}$ (see column 7 of Table 6.2). We conclude the beneficial effect of promoting favorable electrostatic interaction in thermophilic proteins does not get over compensated by the destabilizing effect of solvation. The only exceptions are the RNaseP and Thioredoxin systems, as identified above. However, we notice there are there other protein pairs (ACP, HGCS, and RNaseH) where the benefit of electrostatic interaction is marginal (around 3kT) after accounting for the contribution from the desolvation penalty (see the section below for a discussion of this).

While the electrostatic contribution to energetics at room temperature (300K) can be illuminating and revealing, it is possible that the role of high temperature is important to properly understand enhanced thermostability.[83] This is also consistent with the hypothesis that better connected ion-pair networks may impart greater resilience to thermophilic proteins at high temperature.[316] Therefore, for protein systems ACP, RNaseH, and HGCS, it is possible that the combined effect of solvation and electrostatic interaction becomes significantly favorable in the thermophile's natural elevated temperature environment.

6.2.4 Insights gleaned from the sequence comparison between orthologous proteins

Are there any general guidelines from sequences that can rationalize the structural observables discussed above? The global study presented here provided an opportunity to explore the relationship between biophysical observables and properties of the sequence.

First, by focusing on the role of protein net charge, and defining $|Q_{net}^{thermo}|$ and $|Q_{net}^{meso}|$ as the net charge in thermophile and mesophile, respectively, it was observed

that $|Q_{net}^{thermo}|$ is less than, or equal to $|Q_{net}^{meso}|$ for the protein pairs CheA, CheW, CheY, CSP, HGCS, HPr, NusB, PaiA, Anti$-\sigma$, and Thioredoxin (see Table 6.5).

System	$\|Q_{net}^{thermo}\|$	$\|Q_{net}^{meso}\|$	SCD_{thermo}	SCD_{meso}
ACP	6	2	-40.6	-40.5
CheA	5	16	42.6	640.6
CheW	4	11	-119.5	365.1
CheY	0	4	-60.0	-38.9
CSP	1	1	-40.7	-24.0
HGCS	18	20	943.2	1263.7
HPr	0	2	-51.3	-52.9
NusB	0	0	-176.1	-146.8
PaiA	1	5	-202.6	-31.2
RNaseH	9	4	155.0	-167.9
RNaseP	20	15	936.5	621.2
Anti-σ	1	2	-61.0	-48.3
Thioredoxin	2	4	-108.9	-60.5

Table 6.5: Net charge and calculated SCD per protein system pair. Protein net charge is calculated by summing the charge contributions from basic and acidic amino acids (+1 and -1, respectively) in sequence. SCD is found from equation 3.14.

Interestingly, with only the exception of Thioredoxin, these protein pairs have shown an electrostatics mediated adaptation strategy as discussed above, supporting the simple picture of confining too many unbalanced charges in the compact folded state can be destabilizing. Quantitive models of electrostatics based on protein net charge have been developed and used to describe pH, salt dependent stability,[141] and oxidative damage on aging proteome.[324] In this study, the only exception to this rule is the protein pair RNaseH, where the thermophile has a much higher net charge than the mesophile, and yet more favorable electrostatic energy than the mesophile, demonstrating the importance of delicate placement of charges in the structure to

gain enhanced stability. An obvious strategy to achieve this would be to selectively distribute similar charges far from each other.[114] It is also interesting to note, for proteins NusB and CSP, $|Q_{net}^{thermo}| = |Q_{net}^{meso}|$, but both of these protein pairs have more favorable electrostatic interaction energies in the thermophile compared to mesophiles. This again highlights the role of preferential charge placement[67,69] in addition to net charge composition.

Another interesting feature was observed in system pairs CheA and CheW, where the mesophilic protein have a significantly high charge imbalance compared to its thermophile, implying the thermophilic sequences have relieved repulsion to make electrostatic contribution more favorable with respect to their mesophilic sequences. Differences in electrostatic interaction energy values in Table 6.2 support this view, and is consistent with experimental studies on other proteins.[73–76]

The HGCS protein pair displayed the same imbalanced trend, but with a subtle difference. In this case, the significant charge imbalance remains in thermophilic sequence, albeit to a lesser extent than the mesophile, i.e. $|Q_{net}^{thermo}| = 18$ and $|Q_{net}^{meso}| = 20$. Consequently, both the thermophilic and mesophilic proteins have an overall repulsive electrostatic interaction energy (positive value, not shown) in the folded state, however the thermophilic protein relieves some of this repulsion. For RNaseP, the electrostatic folded state interaction is positive, consistent with the fact that RNaseP has a high charge imbalance, $|Q_{net}^{meso}| = 15$ and $|Q_{net}^{thermo}| = 20$. Furthermore, a larger increase in the charge imbalance seen in thermophilic RNaseP is consistent with the overall destabilizing electrostatic energies observed in this system (positive values in columns 2 and 3 in Table 6.2).

While the discussion above has focused on the role of charge composition, what can we learn from the patterning of the charges in the primary sequence? Recent work[129] has introduced a novel sequence charge decoration (SCD) metric that can discriminate between two sequences having the same charge composition, but different patterning (see discussion in previous chapter). For two orthologous proteins, considering *only* the difference in charge content and patterning, and assuming everything else equal, a lower value of SCD implies a more compact denatured state.[129] Using this metric on a dataset of 540 orthologous pairs, it has been shown that SCD is, on average, lower in thermophiles compared to mesophiles, implying thermophiles, in general, have a more compact denatured state compared to mesophiles and highlights the role of charge patterning in thermophilic adaptation. Consistent with this finding, we notice SCD_{thermo} is less or comparable to SCD_{meso} for the majority of the sequences, 11 out of 13 pairs explored in this work, see Table 6.5. The only exceptions are RNaseH and RNaseP for which SCD_{thermo} is significantly higher than SCD_{meso}. Both of these protein pairs show a significant charge imbalance in the thermophilic system.

We make another interesting observation between SCD and the observed favorable electrostatic interaction energies in the folded state. For two orthologous proteins, the one with lower value of SCD also has a more favorable electrostatic interaction energy. Therefore suggesting preferential placement of opposite charges to promote attraction in the native structure is related to lowering of SCD, which in turn would predict a more compact denatured state. This is consistent with the notion that specific contacts from the folded state may cause residual structure in the unfolded state. Nine protein pairs out of the 13 studied exhibit this trend, with the exceptions being ACP, HPr, RNaseH, and Thioredoxin. In the ACP and HPr systems, the SCD values between thermophile and mesophile are comparable, while

thermophilic protein has significantly more favorable electrostatic interaction in the folded state. In the Thioredoxin and RnaseH systems, the ortholog with a higher SCD was observed to have a lower electrostatic interaction energy, in contrast to the general trend. These exceptions show the subtle differences between charge decoration in the primary sequence and in the folded structure.

It is also illuminating to further investigate the sequences of ACP, HGCS, RNaseH, and RNaseP that show the combined effect of solvation and electrostatic interaction energy is comparable (ACP, HGCS, and RNaseH), or less (RNaseP) favorable in the thermophile compared to the mesophile. Interestingly, three of these pairs (ACP, HGCS, and RNaseP) have a significant enhancement in the fraction of hydrophobic residues in their thermophilic sequences. An increase in hydrophobicity can significantly enhance stability, for example an increase in hydrophobic composition fraction from 0.38 to 0.45 is sufficient to increase the melting temperature by almost 40°C.[25] Consistent with this, mutation studies on RNaseH have shown the hydrophobic core is responsible for the enhanced thermal stability in thermophile.[15] This presents an intriguing possibility that wherever charge neutrality is not an option to obtain higher stability, proteins may rely on an alternate mechanism of significantly enhancing hydrophobic fraction. Proteins with high charge and specific charge patterning have been associated with specific function,[325,326] and these functional restrictions may limit adopting electrostatics as a strategy to enhance stability in these proteins.

Finally, it is important to discuss the protein system pair of Thioredoxin, a pairing that also does not seem to follow the electrostatics strategy. Thermophilic and mesophilic thioredoxin does not show any marked difference in either charge content, or in the degree of hydrophobicity. However, thermophilic sequence of Thioredoxin has a significant proline enrichment compared to its mesophilic counterpart, and separate experiments have shown that proline substitutions can increase the melting

temperature.[171] It should also be noted that the mesophilic Thioredoxin has an unusually high melting temperature (80°C).[286] Therefore, it can not be ruled out, in the absence of experimental measurement of the melting temperature of the thermophilic counterpart, that the thermophilic and mesophilic Thioredoxin perhaps have comparable stability.

6.2.5 Thermophilic proteins do not necessarily have suppressed fluctuations in the native state

Thermophilic proteins are often believed to be more rigid compared to their mesophilic counterparts. While previous simulation studies[54,246,327,328] have already proven a lack of such trend, further exception to this rule is provided here for several protein pairs not studied before. The average fluctuation from the mean structure ($RMSF$) was computed as a function of time for a given protein as described in Methods. Figure 6.3 displays the average fluctuation ($\langle RMSF \rangle$) comparison between thermophile and mesophile constituents for each homolog pair. This observation shows thermophiles do not necessarily have suppressed fluctuations compared to mesophiles. This may seem to contradict that thermophilic enzymes have lower activity at room temperature. However, low activity may be due to high activation barrier of the chemical step, and not necessarily due to structural rigidity.[54,329] It is also important to note, thermophilic proteins may have greater resilience at high temperature, due to better connected ion pairs, which is not synonymous to greater rigidity.[316] Finally, the time scales of reported fluctuations were limited to the simulation time, and therefore, there is no knowledge of large scale motion, or their relative differences between mesophilic and thermophilic proteins.

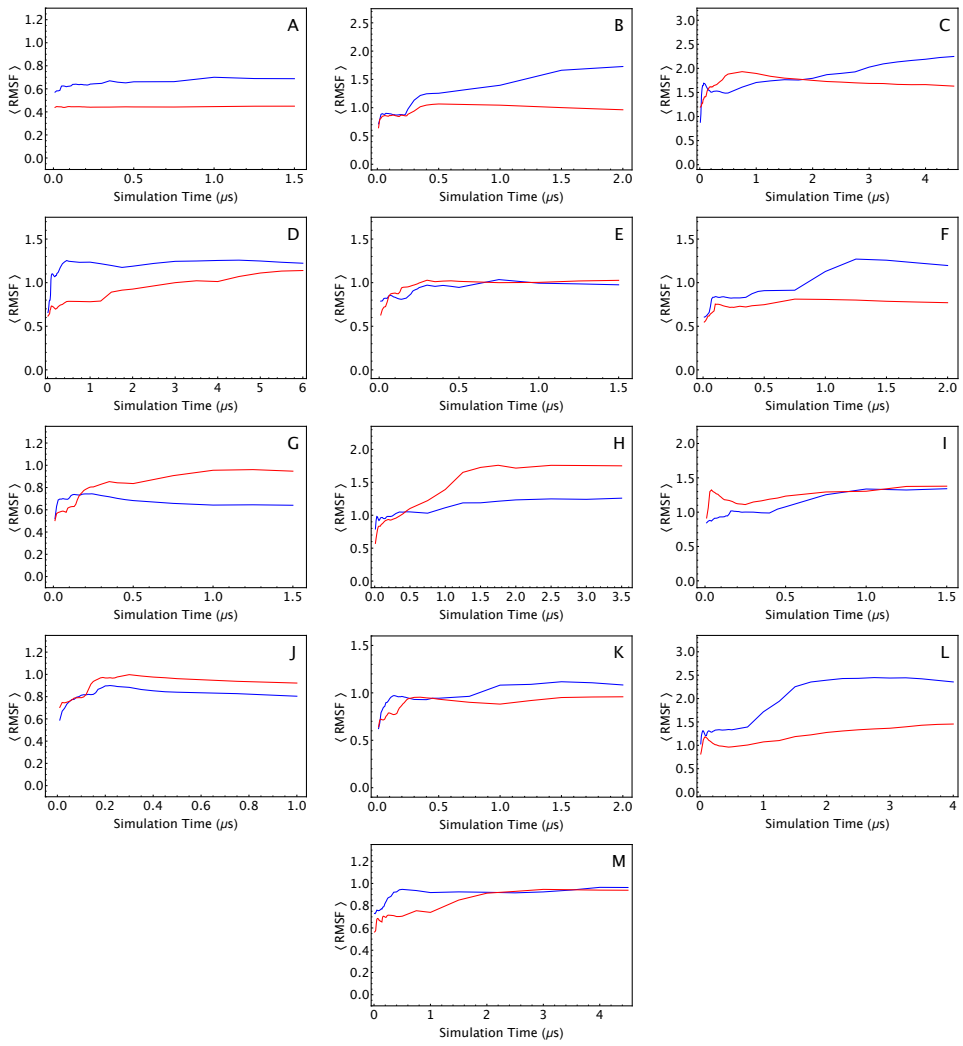

Figure 6.3: Cumulative $\langle RMSF \rangle$ as a function of simulation time for thermophile-mesophile (A) ACP, (B) CheA, (C) CheW, (D) CheY, (E) CSP, (F) HGCS, (G) HPr, (H) NusB, (I) PaiA, (J) RNaseH, (K) RNaseP, (L) Anti-σ, and (M) Thioredoxin. The crossing of curves show the instability in the calculation for short time scales.

The time evolution of $\langle RMSF \rangle$ for each homologous pair,Figure 6.3, makes another important point. For multiple pairs, it is observed that the comparison of fluctuations between thermophile and mesophile derived from short trajectories can lead to drastically different conclusions compared to long time scale simulations, as is evident from multiple crossings of the time evolution graphs between thermophile and mesophile in these systems. The importance of long-time simulation and sampling

issues have been pointed out in the flexibility study of Trigger-factor[240] as well. This once again highlights the importance of long time simulation subjected to careful convergence checks.

6.2.6 Presence of hydrogen bonds

From pairwise thermophilic-mesophilic homolog structure comparisons, thermophilic proteins have been shown to have an increased hydrogen bonding.[44,45] Furthermore, a correlation has been observed between an increased hydrogen bond content and thermal stability[1,39,43,47–49] . Results of MD simulations have shown that the thermophile formed not just more protein-protein hydrogen bonds, but also displayed an affinity to form these bonds with the solvent,[46] and additional hydrogen bonds help maintain the native states of thermophilic systems during high temperature simulations, while the mesophilic homolog denatures.[50,51] However, several contradictory studies have demonstrated that there is no difference in the amount of hydrogen bonds between homologous groupings,[52–55] leaving the role of these bonds toward increased thermal stability a disputed topic.

The results of our all-atom MD simulations allowed the detailed calculation of hydrogen bonds existent in the native state of our system pairs, and the contribution these bonds offer to enhanced thermal stability. Specifically, we can examine whether thermophilic proteins make additional h-bonds, and not from a static structural representation, but from a distribution of dynamic native state frames.

Methods

Here, we will examine the presence of four categories of hydrogen bonds classified as: i) any donor-receptor atom pair within the protein structure (labeled U-U), ii) strictly protein backbone donor atom to backbone receptor atom bonds (B-B), iii)

protein atom to solvent h-bonds (U-V), and iv) hydrogen bond bridges that occur when a solvent molecule is shared between two separate protein atoms (U-V-U). All h-bonds were calculated with CPPTRAJ,[264] and analyzed with in-house software.

Results

We define the difference in the number of hydrogen bonds of a given type as the number presented in the thermophile minus those present in the mesophilic counterpart. Therefore, values greater than zero reflect the system pairs with the thermophile having a larger count of a given type, and values less than zero show the mesophilic protein with a larger hydrogen bond count. The difference in counts are shown in Table 6.6. As seen from the above analysis, the thermophilic component from the homologous pairs have a tendency for increased hydrogen bonds within the protein structure (second column of Table 6.6). Two systems display a minimal increase (less than two more) in the mesophilic constituent over the thermophilic (HPr and RNaseP), and the lone significant outlier is the PaiA system, where the mesophile has, on average, six more structural h-bonds. Each hydrogen bond will enhance the interaction free energy to a more favorable value by approximately $2kT$,[38] so we can claim that the additional bondings will help stabilize the thermophilic proteins. But, in accounting for the energetics of donor-acceptor groups, are the solvation and polarization energies going be substantially altered such that they destabilize the system? The energy calculations to directly answer this question are a focus of near future work, but a quick discussion will be presented here. We would need to account for the donor-acceptor groups in DelPhi to calculate the energies that arise from additional hydrogen bonds. However, we know from Table B.4 that the polarization and solvation energies found in the folded state of CSP by using all-atom and the ionic representative atom are in good agreement. That is, by accounting for more than

System	Δ(U-U)	Δ(B-B)	Δ(U-V)	Δ(U-V-U)
ACP	3.4	4.0	-13.5	-0.3
CheA	6.8	5.6	-25.0	-3.6
CheW	2.7	-1.1	5.3	-2.2
CheY	4.1	-2.8	-16.3	-3.1
CSP	5.7	1.8	7.8	2.3
HGCS	1.6	-2.2	-22.2	-4.3
HPr	-1.6	-3.5	-8.0	-2.2
NusB	9.7	-1.0	10.2	-0.2
PaiA	-6.2	-4.2	-36.8	-2.1
RNaseH	4.3	2.9	1.1	-2.0
RNaseP	-0.6	1.0	-17.8	-5.9
Anti-σ	3.4	-1.3	18.4	2.8
Thioredoxin	5.5	2.3	3.3	1.4

Table 6.6: The average hydrogen bond content calculated from MD native state simulations. The four classifications of bonds are U-U (all protein-protein), B-B (strictly protein backbone-backbone), U-V (protein-solvent), and U-V-U (bridges between two protein sites and a solvent molecule).

just the representative ionic atoms, we see little change in solvation and polarization energies in the folded state. The nuance will be presented in the unfolded state calculations, where the methodology used to calculate electrostatic solvation energies will not suffice. However, by inspection of each homolog pair's sequences, we observe all thermophiles have less polar amino acids than their mesophilic partner. Assuming the backbone donor-acceptor groups are equal in solvation contributions, thermophiles would pay a smaller desolvation penalty than mesophiles with respect to polar side chain interactions with solvent. The conclusion would be that mesophilic solvation energies from polar amino acids would be more unfavorable than thermophilic due to the enrichment of these types of amino acids interacting with solvent in the un-

folded state. Also, the thermophilic component of PaiA has such a larger favorable electrostatic free energy compared to the mesophile, that the average difference of six hydrogen bonds is not enough to change the observed stability claims.

6.3 Conclusion

Based on the presented global study of multiple protein pairs, five key conclusions are to be made. First, for the majority of the protein pairs, the electrostatic interaction energy is more favorable in the native state of thermophilic proteins compared to mesophiles, consistent with several earlier experimental studies.[67,71,73–75,321–323] The trend remained unaltered even with the inclusion of the desolvation penalty upon folding. However, with the inclusion of the desolvation penalty, three protein pairs (ACP, HGCS, and RNaseH) found the favorable electrostatic gain became marginal, less than $3kT$. Second, within the simulation time scale, the widely held view that thermophilic proteins are preferentially suppressed motion than their mesophilic counterparts is not supported. Third, long simulation trajectories are necessary to avoid sampling artifacts, as is evident from the observed behavior based on early simulation times (short time scale) that can conflict with long simulation results. This reiterates recent work that has shown the convergence of native state simulations can be challenging, and may require several microseconds long simulations.[130] Fourth, estimating electrostatic contribution to energetics solely based on PDB structures can give incorrect qualitative and quantitative predictions. Meaningful conclusions can only be drawn from simulation trajectories that include native state dynamics. Finally, identifying alternate mechanisms other than electrostatics for few protein pairs may demonstrate different available strategies to enhance thermostability when modifying electrostatics is not possible for functional reasons.

Chapter 7

Organismal growth rate calculation and comparison with experiment

Understanding the link between the molecular effects of protein mutations and cellular fitness is key to understanding evolutionary dynamics. The work of Shamoo et al has shown that by replacing the gene for Adenylate Kinase (ADK) in a thermophilic organism with the gene from a mesophile, the modified thermophile expressed the mesophile enzyme, but organismal fitness was greatly reduced due to enzyme heat inactivation.[126–128] Furthermore, Shakhnovich and coworkers observed that by incorporating destabilizing mutations in a core metabolic enzyme dihydrofolate reductase (DHFR), the growth rate of E. Coli was substantially reduced.[330] Both of these experimental works demonstrate the quantitative connection of the molecular effects of protein mutations to cellular fitness.

The physical limitations of cells is dominated by the cell's collection of proteins, the proteome, and cells live on the edge of a proteome catastrophe with respect to temperature.[331,332] Based on our thermodynamic analysis, the protein stability can be modeled as a function of chain length using parameters of the *Ideal Thermal*

Protein.[141,331] This model was extended to calculate the stability distribution of entire proteomes, and estimate organismal growth rates. As will be shown, alterations to protein thermodynamics effect the growth rate of the organism.

7.1 Model

For many organisms, the chain length distribution can be modeled as a gamma distribution[333]

$$P(N) = \frac{N^{\alpha-1}\exp\left(-N/\theta\right)}{\Gamma(\alpha)\theta^{\alpha}} \tag{7.1}$$

where, α and θ are two parameters. This approach has been previously employed to model the proteome stability distribution for different organismal growth rates, such as E. Coli, Yeast, C. Elegans.[331] Using the proteome chain length distribution, $P(N)$, and free energy equations (2.7 and 2.8), the free energy distribution $P(\Delta G)$ of the entire proteome was calculated to estimate the growth rate, $r(T)$, of several mesophilic and thermophilic organism. As before, for a given proteome, we take

$$r(T) = r_0 \exp(-\Delta H^{\dagger}/RT) \prod_{i=1}^{\Gamma} \frac{1}{1+\exp(-\Delta G_i/RT)} \tag{7.2}$$

where, r_0 is an intrinsic rate, and $\Delta H^{\dagger}$ represents an Arrhenius activation barrier for a metabolic reaction rate.[331,334,335] The product term describes the stabilities of proteins $i = 1, 2, 3, \ldots, \Gamma$, where Γ represents the number of essential proteins that are necessary for the growth rate. The expression above assumes fitness depends on all the essential proteins and their propensity to be in the folded state. This is motivated by the fact that compromising the stability of any of these essential proteins is lethal to the organism, and therefore explains the product notation in equation 7.2. Furthermore, it assumes growth rate is related to fitness. Equation 7.2 has already

been successfully used to model growth rates in different organisms.[331,334,335] Taking the logarithm of the rate gives equation 7.2 as

$$\log r(T) = \log r_0 - \frac{\Delta H^\dagger}{RT} - \sum_{i=1}^{\Gamma} \log(1 + \exp(-\Delta G_i/RT)) \tag{7.3}$$

The summation was approximated as an integral over the entire proteome free energy distribution, $P(\Delta G)$ and express average rate[334] as

$$\langle \log r(T) \rangle = \log r_0 - \frac{\Delta H^\dagger}{RT} - \Gamma \int \log(1 + \exp(-\Delta G/RT)) P(\Delta G) d\Delta G \tag{7.4}$$

The expression above requires the stability distribution $P(\Delta G)$. We estimate this distribution by using the proteome length distribution equation 7.1 and equation 2.7 (for mesophiles) or equation 2.8 (for thermophiles). Equation 7.4 predicts that cellular growth rates increase with temperature at low temperature due to the assumed activated process. It predicts maximum growth at an optimum growth temperature, and growth rates that decrease at high temperatures due to proteome denaturation (see Figure 7.1). These curves are highly asymmetrical near the temperature of maximum growth, and the model predicts this well. For this calculation, the model requires two free parameters $\Delta H^\dagger$ and Γ, which were determined by fitting the experimental data on several mesophilic and thermophilic organisms (see Figure 7.1). Values of the fitted parameters are reported in Table 7.1 using corresponding expressions of ΔG respective of the organism.

7.2 Results

Quantitative agreement between experimental thermal growth data, and our proteome modeling based on *Ideal Mesophilic Protein* and *Ideal Thermophilic Protein*

Mesophilic Species	$\Delta H^{\dagger}$	Γ	T_{opt} [K]	$\overline{L}$
B. Megaterium[336]	19.7	26	315.2	272
B. Subtilis[336]	16.9	30	314.5	294
E. Coli[337]	27.1	51	314.1	301
L. Monocytogenes[338]	25.9	223	309.7	303
P. Aeruginosa[336]	21.6	101	311.9	328
P. Fluorescens[336]	28.9	166	311.7	340
Thermophilic Species	$\Delta H^{\dagger}$	Γ	T_{opt} [K]	$\overline{L}$
B. Acidocaldarius[339]	15.6	107	339.2	306
M . Thermoautotrophicus[340]	29.8	81	341.2	281
T. Aquaticus[336]	19.1	3	347.0	284
T. Brockii[341]	48.8	22	345.3	304
T. Neopolitana[342]	28.4	1	348.0	313
T. Thermophilus[343]	10.6	2	346.9	299

Table 7.1: Fitted Parameters from Proteome Analysis: $\Delta H^{\dagger}$, Γ, and T_{opt} are the activation barrier, essential proteins, and optimum growth temperature, respectively, calculated from fitting equation 7.4 to experimental data. $\overline{L}$ is the average length of all proteins found within a given proteome.

parameters provide additional support to our approach and finding. Growth rates computed from equation 7.4, and protein thermodynamic data, captured the optimal growth temperature, as well as the asymmetric temperature dependent growth curves across many thermophilic and mesophilic organisms. Furthermore, it should be noted, only mesophilic (thermophilic) ideal protein parameters could be used to fit mesophilic (thermophilic) organisms. Unphysical parameter values and poor fit were observed while using mesophilic protein parameters to fit thermophilic organism growth data, and vice versa.

The extreme sensitivity of thermodynamic parameters to model growth data, and successful modeling only upon proper selection of protein thermodynamic parameters,

suggests engineering protein stability is perhaps one of the key factors organisms utilize to adapt to elevated temperatures. It should also be noted that application of growth rate (equation 7.2) to model fitness may not always be accurate and depends on nutrient conditions.[345,346]

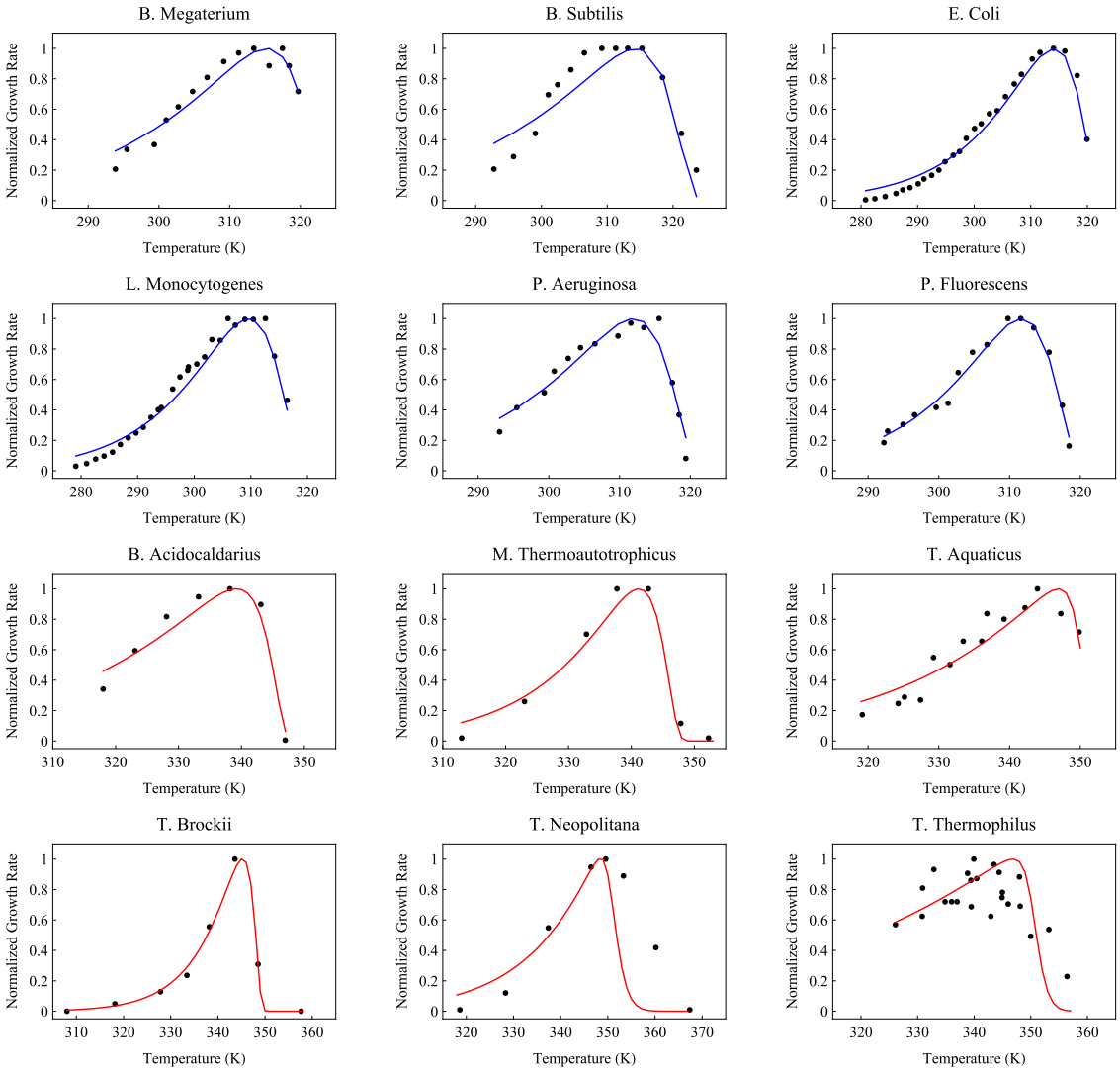

Figure 7.1: Black circles denote experimental determined growth rate as function of temperature for species listed. In the y–axis, normalized growth rate with respect to maximum growth rate is plotted, while the x-axis displays temperature. Solid lines are fit to data from equation 7.4, and the name of the fitted species are given in the graph. Blue curves denote mesophilic organisms, while red graphs are for thermophiles. The sources of the growth-rate data for the different organisms are given in Table 7.1 next to the species names. Proteome chain length information about each of the organisms were obtained from Genbank[344]

Chapter 8

Concluding Remarks

Attempting to understand the molecular strategies of enhanced stability in thermophilic proteins has presented many challenges. Here, we have presented the results of many studies in an attempt to gain insightful knowledge of possible mechanisms. From a study of thermodynamic data, we have observed that thermophilic proteins, on average, have a reduced change in thermodynamic parameters (ΔH, ΔS, and ΔC_p), but only the reduced entropy change had a favorable effect to the stability curve. The reduced parameters allowed us to deduce that the presence of residual structure in the unfolded state of thermophiles may be a potential mechanism adopted to thrive in high temperature environments. Further study of the unfolded state demonstrated that the constitution and pattering of charged amino acids at the sequence level played a large role, and the effect of electrostatic interactions lead to a reduced size in thermophilic proteins, verifying the residual structure hypothesis. Furthermore, the role of electrostatic interactions from the folded state dynamics study highlighted the contributions charged amino acids make in forming well connected ionic interaction networks in thermophilic proteins. Compared to their mesophilic counterparts, and accounting for desolvation penalties, thermophilic proteins benefit from a more

favorable electrostatic free energy from their ionic constituents. As observed, thermophilic proteins tend to enhance the attractive ionic interactions, and relieve the repulsive contributions, and this method should be an avenue of consideration when attempting to engineer proteins with heightened thermal resistance.

The observations we have gathered to date include the aforementioned dynamics study performed strictly at room temperature. Although this study was insightful by demonstrating the interactions of ionic amino acids in the folded state, do these interactions lead directly to enhanced stability? An in-depth examination of the unfolding simulation trajectories displayed in Figures 5.1-5.14 could illuminate potential unfolding pathways in mesophilic proteins due to high temperature stresses. Further analysis of these simulations could highlight the "weak spots" in mesophilic proteins, and comparison to thermophilic information would offer insights into the modifications thermophilic proteins have adopted to resist high temperature denaturing.

Bibliography

[1] Vogt, G.; Argos, P. *Folding and Design* **1997**, *2, Supplement 1*, S40 – S46.

[2] Bruins, M.; Janssen, A.; Boom, R. *Applied Biochemistry and Biotechnology* **2001**, *90*, 155–186.

[3] Turner, P.; Mamo, G.; Karlsson, E. N. *Microbial Cell Factories* **2007**, *6*, 1–23.

[4] Kaushik, S.; Cuervo, A. M. *Nature Medicine* **2015**, *21*, 1406–1415.

[5] Labbadia, J.; Morimoto, R. I. *F1000Prime Reports* **2014**, *6*.

[6] Balch, W. E.; Morimoto, R. I.; Dillin, A.; Kelly, J. W. *Science* **2008**, *319*, 916–919.

[7] Powers, E. T.; Morimoto, R. I.; Dillin, A.; Kelly, J. W.; Balch, W. E. *Annual Review of Biochemistry* **2009**, *78*, 959–991.

[8] Nojima, H.; Ikai, A.; Oshima, T.; Noda, H. *Journal of Molecular Biology* **1977**, *116*, 429 – 442.

[9] Rees, D. C.; Adams, M. W. *Structure* **1995**, *3*, 251 – 254.

[10] Alexander, P.; Fahnestock, S.; Lee, T.; Orban, J.; Bryan, P. *Biochemistry* **1992**, *31*, 3597–3603.

[11] McCrary, B. S.; Edmondson, S. P.; Shriver, J. W. *Journal of Molecular Biology* **1996**, *264*, 784 – 805.

[12] Beadle, B. M.; Baase, W. A.; Wilson, D. B.; Gilkes, N. R.; Shoichet, B. K. *Biochemistry* **1999**, *38*, 2570–2576.

[13] Zhou, H.-X. *Biophysical Journal* **2002**, *83*, 3126 – 3133.

[14] Ratcliff, K.; Corn, J.; Marqusee, S. *Biochemistry* **2009**, *48*, 5890–5898.

[15] Robic, S.; Guzman-Casado, M.; Sanchez-Ruiz, J.; Marqusee, S. *Proc. Natl. Acad. Sci.* **2003**, *100*, 11345–11349.

[16] Fu, H.; Grimsley, G.; Scholtz, J.; Pace, C. *Protein Science* **2010**, *19*, 1044–1052.

[17] Chan, C.-H.; Yu, T.-H.; Wong, K.-B. *PLoS One* **2011**, *6*.

[18] Kumar, S.; Tsai, C.-J.; Nussinov, R. *Biochemistry* **2001**, *40*, 14152–14165.

[19] Sawle, L.; Ghosh, K. *Biophysical Journal* **2011**, *101*, 217 – 227.

[20] Prabhu, N. V.; Sharp, K. A. *Annual Review of Physical Chemistry* **2005**, *56*, 521–548.

[21] Dill, K. A. *Biochemistry* **1990**, *29*, 7133–7155.

[22] Creighton, T. E. *Current Opinion in Structural Biology* **1991**, *1*, 5 – 16.

[23] Chakravarty, S.; Varadarajan, R. *FEBS Letters* **2000**, *470*, 65–69.

[24] Gromiha, M. M.; Pathak, M. C.; Saraboji, K.; Ortlund, E. A.; Gaucher, E. A. *Proteins: Structure, Function, and Bioinformatics* **2013**, *81*, 715–721.

[25] Dill, K. A.; Alonso, D. O. V.; Hutchinson, K. *Biochemistry* **1989**, *28*, 5439–5449.

[26] Shih, P.; Kirsch, J. F.; Holland, D. R. *Protein Science* **1995**, *4*, 2050–2062.

[27] Serrano, L.; Kellis, J. T.; Cann, P.; Matouschek, A.; Fersht, A. R. *Journal of Molecular Biology* **1992**, *224*, 783 – 804.

[28] Anderson, D. E.; Hurley, J. H.; Nicholson, H.; Baase, W. A.; Matthews, B. W. *Protein Science* **1993**, *2*, 1285–1290.

[29] Pace, C. N.; Fu, H.; Fryar, K. L.; Landua, J.; Trevino, S. R.; Shirley, B. A.; Hendricks, M. M.; Iimura, S.; Gajiwala, K.; Scholtz, J. M.; Grimsley, G. R. *Journal of Molecular Biology* **2011**, *408*, 514 – 528.

[30] Puchkaev, A. V.; Koo, L. S.; de Montellano, P. R. O. *Archives of Biochemistry and Biophysics* **2003**, *409*, 52 – 58.

[31] van den Burg, B.; Enequist, H. G.; van der Haar, M. E.; Eijsink, V. G.; Stulp, B. K.; Venema, G. *Journal of Bacteriology* **1991**, *173*, 4107–4115.

[32] Van den Burg, B.; Dijkstra, B.; Vriend, G.; Van der Vinne, B.; Venema, V., Gand Eijsink *European Journal of Biochemistry* **1994**, *220*, 981–985.

[33] Serrano, L.; Bycroft, M.; Fersht, A. R. *Journal of Molecular Biology* **1991**, *218*, 465 – 475.

[34] Strub, C.; Alies, C.; Lougarre, A.; Ladurantie, C.; Czaplicki, J.; Fournier, D. *BMC Biochemistry* **2004**, *5*, 1–6.

[35] Matsuura, Y.; Takehira, M.; Joti, Y.; Ogasahara, K.; Tanaka, T.; Ono, N.; Kunishima, N.; Yutani, K. *Scientific Reports* **2015**, *5*.

[36] Prevost, M.; Wodak, S. J.; Tidor, B.; Karplus, M. *Proceedings of the National Academy of Sciences* **1991**, *88*, 10880–10884.

[37] Priyakumar, U. D. *Journal of Biomolecular Structure and Dynamics* **2012**, *29*, 961–971.

[38] Hubbard, R. E.; Kamran Haider, M. *eLS*; John Wiley & Sons, Ltd, 2001.

[39] Myers, J.; Pace, C. *Biophysical Journal* **1996**, *71*, 2033 – 2039.

[40] Baldwin, R. L. *Protein Folding Handbook*; Wiley-VCH Verlag GmbH, 2008; pp 127–162.

[41] Rose, G. D.; Wolfenden, R. *Annual Review of Biophysics and Biomolecular Structure* **1993**, *22*, 381–415.

[42] Poland, D.; Scheraga, H. *Theory of helix-coil transitions in biopolymers: statistical mechanical theory of order-disorder transitions in biological macromolecules*; Molecular biology; Academic Press, 1970.

[43] Vogt, G.; Woell, S.; Argos, P. *Journal of Molecular Biology* **1997**, *269*, 631 – 643.

[44] Wallon, G.; Kryger, G.; Lovett, S. T.; Oshima, T.; Ringe, D.; Petsko, G. A. *Journal of Molecular Biology* **1997**, *266*, 1016 – 1031.

[45] Dalhus, B.; Saarinen, M.; Sauer, U. H.; Eklund, P.; Johansson, K.; Karlsson, A.; Ramaswamy, S.; Bjørk, A.; Synstad, B.; Naterstad, K.; Sirevåg, R.; Eklund, H. *Journal of Molecular Biology* **2002**, *318*, 707 – 721.

[46] Melchionna, S.; Sinibaldi, R.; Briganti, G. *Biophysical Journal* **2006**, *90*, 4204 – 4212.

[47] Shirley, B. A.; Stanssens, P.; Hahn, U.; Pace, C. N. *Biochemistry* **1992**, *31*, 725–732.

[48] Takano, K.; Scholtz, J. M.; Sacchettini, J. C.; Pace, C. N. *Journal of Biological Chemistry* **2003**, *278*, 31790–31795.

[49] Pace, C. N. et al. *Protein Science* **2014**, *23*, 652–661.

[50] Colombo, G.; Merz, K. M. *Journal of the American Chemical Society* **1999**, *121*, 6895–6903.

[51] Singh, B.; Bulusu, G.; Mitra, A. *The Journal of Physical Chemistry B* **2015**, *119*, 392–409.

[52] Wang, X.; He, X.; Yang, S.; An, X.; Chang, W.; Liang, D. *Journal of Bacteriology* **2003**, *185*, 4248–4255.

[53] Szilágyi, A.; Závodszky, P. *Structure* **2000**, *8*, 493 – 504.

[54] Kalimeri, M.; Rahaman, O.; Melchionna, S.; Sterpone, F. *J. Physical Chemistry B* **2013**, *117*, 13775–13785.

[55] Manjunath, K.; Sekar, K. *Journal of Chemical Information and Modeling* **2013**, *53*, 2448–2461.

[56] Taylor, T. J.; Vaisman, I. I. *BMC Structural Biology* **2010**, *10*, 1–10.

[57] Yip, K.; Stillman, T.; Britton, K.; Artymiuk, P.; Baker, P.; Sedelnikova, S.; Engel, P.; Pasquo, A.; Chiaraluce, R.; Consalvi, V.; Scandurra, R.; Rice, D. *Structure* **1995**, *3*, 1147 – 1158.

[58] Jaenicke, R.; Schurig, H.; Beaucamp, N.; Ostendorp, R. *Enzymes and Proteins from Hyperthermophilic Microorganisms*; Advances in Protein Chemistry; Academic Press, 1996; Vol. 48; pp 181 – 269.

[59] Korndörfer, I.; Steipe, B. S.; Huber, R.; Tomschy, A.; Jaenicke, R. *Journal of Molecular Biology* **1995**, *246*, 511 – 521.

[60] Perutz, M.; Raidt, H. *Nature* **1975**, *255*, 256–259.

[61] Jónsdóttir, L. B.; Ellertsson, B. Ö.; Invernizzi, G.; Magnúsdóttir, M.; Thorbjarnardóttir, S. H.; Papaleo, E.; Kristjánsson, M. M. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2014**, *1844*, 2174 – 2181.

[62] Kumar, S.; Ma, B.; Tsai, C.-J.; Nussinov, R. *Proteins: Structure, Function, and Bioinformatics* **2000**, *38*, 368–383.

[63] Missimer, J. H.; Steinmetz, M. O.; Baron, R.; Winkler, F. K.; Kammerer, R. A.; Daura, X.; van Gunsteren, W. F. *Protein Science* **2007**, *16*, 1349–1359.

[64] Kumar, S.; Nussinov, R. *ChemBioChem* **2002**, *3*, 604–617.

[65] Lebbink, J. H.; Knapp, S.; van der Oost, J.; Rice, D.; Ladenstein, R.; de Vos, W. M. *Journal of Molecular Biology* **1999**, *289*, 357 – 369.

[66] Spassov, V. Z.; Karshikoff, A. D.; Ladenstein, R. *Protein Science* **1994**, *3*, 1556–1569.

[67] Loladze, V. V.; Ibarra-Molero, B.; Sanchez-Ruiz, J. M.; Makhatadze, G. I. *Biochemistry* **1999**, *38*, 16419–16423.

[68] Xiao, L.; Honig, B. *Journal of Molecular Biology* **1999**, *289*, 1435 – 1444.

[69] Sanchez-Ruiz, J. M.; Makhatadze, G. I. *Trends in Biotechnology* **2001**, *19*, 132 – 135.

[70] Alsop, E.; Silver, M.; Livesay, D. R. *Protein Engineering* **2003**, *16*, 871–874.

[71] Gribenko, A. V.; Patel, M. M.; Liu, J.; McCallum, S. A.; Wang, C.; Makhatadze, G. I. *Proceedings of the National Academy of Sciences* **2009**, *106*, 2601–2606.

[72] Dao-Pin, S.; Anderson, D. E.; Baase, W. A.; Dahlquist, F. W.; Matthews, B. W. *Biochemistry* **1991**, *30*, 11521–11529.

[73] Grimsley, G. R.; Shaw, K. L.; Fee, L. R.; Alston, R. W.; Huyghues-Despointes, B. M.; Thurlkill, R. L.; Scholtz, J. M.; Pace, C. N. *Protein Science* **1999**, *8*, 1843–1849.

[74] Perl, D.; Mueller, U.; Heinemann, U.; Schmid, F. X. *Nature Structural & Molecular Biology* **2000**, *7*, 380–383.

[75] Spector, S.; Wang, M.; Carp, S. A.; Robblee, J.; Hendsch, Z. S.; Fairman, R.; Tidor, B.; Raleigh, D. P. *Biochemistry* **2000**, *39*, 872–879.

[76] Perl, D.; Schmid, F. X. *Journal of Molecular Biology* **2001**, *313*, 343 – 357.

[77] de Bakker, P. I.; Hünenberger, P. H.; McCammon, J. *Journal of Molecular Biology* **1999**, *285*, 1811 – 1830.

[78] Danciulescu, C.; Ladenstein, R.; Nilsson, L. *Biochemistry* **2007**, *46*, 8537–8549.

[79] Huang, X.; Zhou, H.-X. *Biophysical Journal* **2006**, *91*, 2451 – 2463.

[80] Lee, K.-J. *Journal of Chemical Information and Modeling* **2012**, *52*, 7–15.

[81] Chen, L.; Li, X.; Wang, R.; Fang, F.; Yang, W.; Kan, W. *Journal of Biomolecular Structure and Dynamics* **2016**, *0*, 1–14.

[82] Kumar, S.; Nussinov, R. *Journal of Molecular Biology* **1999**, *293*, 1241 – 1255.

[83] Elcock, A. H. *Journal of Molecular Biology* **1998**, *284*, 489 – 502.

[84] Hendsch, Z. S.; Tidor, B. *Protein Science* **1994**, *3*, 211–226.

[85] Karshikoff, A.; Ladenstein, R. *Trends in Biochemical Sciences* **2001**, *26*, 550 – 557.

[86] Simonson, T.; Brooks, C. L. *Journal of the American Chemical Society* **1996**, *118*, 8452–8458.

[87] Dominy, B. N.; Minoux, H.; Brooks, C. L. *Proteins: Structure, Function, and Bioinformatics* **2004**, *57*, 128–141.

[88] Vieille, C.; Zeikus, J. G. *Trends in Biotechnology* **1996**, *14*, 183 – 190.

[89] Eisenmesser, E.; Millet, O.; Labeikovsky, W.; Korzhnev, D.; Wolf-Watz, M.; Bosco, D.; Skalicky, J.; Kay, L.; Kern, D. *Nature* **2005**, *438*, 117–21.

[90] Teilum, K.; Olsen, J. G.; Kragelund, B. B. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2011**, *1814*, 969 – 976.

[91] Jaenicke, R. *European Journal of Biochemistry* **1991**, *202*, 715 – 728.

[92] Varley, P. G.; Pain, R. H. *Journal of Molecular Biology* **1991**, *220*, 531 – 538.

[93] Somero, G. N. *Annual Review of Physiology* **1995**, *57*, 43–68.

[94] Zavodszky, P.; Kardos, J.; Svingor, A.; Petsko, G. A. *Proceedings of the National Academy of Sciences* **1998**, *95*, 7406–7411.

[95] Wolf-Watz, M.; Thai, V.; Henzler-Wildman, K.; Hadjipavlou, G.; Eisenmesser, E. Z.; Kern, D. *Nature Structural & Molecular Biology* **2004**, *11*, 945–949.

[96] Feller, G. *Journal of Physics: Condensed Matter* **2010**, *22*, 323101.

[97] Lam, S.; Yeung, R.; Yu, T.-H.; Sze, K.-H.; Wong, K.-B. *PLoS Biol* **2011**, *9*, e1001027.

[98] Matsumura, M.; Signor, G.; Matthews, B. W. *Nature* **1989**, *342*, 291–293.

[99] Clarke, J.; Fersht, A. R. *Biochemistry* **1993**, *32*, 4322–4329.

[100] Matthews, B.; Nicholson, H.; Becktel, W. J. *Proceedings of the National Academy of Sciences of the United States of America* **1987**, *84*, 6663–6667.

[101] Mansfeld, J.; Vriend, G.; Dijkstra, B. W.; Veltman, O. R.; Van den Burg, B.; Venema, G.; Ulbrich-Hofmann, R.; Eijsink, V. G. H. *Journal of Biological Chemistry* **1997**, *272*, 11152–11156.

[102] Wrba, A.; Schweiger, A.; Schultes, V.; Jaenicke, R.; Zavodszky, P. *Biochemistry* **1990**, *29*, 7584–7592.

[103] Sekhar, A.; Lam, H. N.; Cavagnero, S. *Protein Science* **2012**, *21*, 1489–1502.

[104] Hernández, G.; Jenney, F. E.; Adams, M. W. W.; LeMaster, D. M. *Proceedings of the National Academy of Sciences* **2000**, *97*, 3166–3170.

[105] Jaenicke, R. *Proceedings of the National Academy of Sciences* **2000**, *97*, 2962–2964.

[106] Hollien, J.; Marqusee, S. *Proceedings of the National Academy of Sciences* **1999**, *96*, 13674–13678.

[107] Fitter, J.; Heberle, J. *Biophysical Journal* **2000**, *79*, 1629 – 1636.

[108] Fitter, J.; Herrmann, R.; Hauss, T.; Lechner, R.; Dencher, N. *Physica B* **2001**, *301*, 1–7.

[109] Querol, E.; Perez-Pons, J. A.; Mozo-Villarias, A. *Protein Engineering* **1996**, *9*, 265–271.

[110] Jacobs, D. J.; Rader, A.; Kuhn, L. A.; Thorpe, M. *Proteins: Structure, Function, and Bioinformatics* **2001**, *44*, 150–165.

[111] Rader, A. J. *Physical Biology* **2010**, *7*, 016002.

[112] Lazaridis, T.; Lee, I.; Karplus, M. *Protein Science* **1997**, *6*, 2589–2605.

[113] Rathi, P. C.; Jaeger, K.-E.; Gohlke, H. *PLoS ONE* **2015**, *10*, e0130289.

[114] Karshikoff, A.; Nilsson, L.; Ladenstein, R. *FEBS Journal* **2015**, *282*, 3899–3917.

[115] Lazaridis, T.; Lee, I.; Karplus, M. *Protein Science* **1997**, *6*, 2589–2605.

[116] Grottesi, A.; Ceruso, M.-A.; Colosimo, A.; Di Nola, A. *Proteins: Structure, Function, and Bioinformatics* **2002**, *46*, 287–294.

[117] Leszczynski, J.; Rose, G. *Science* **1986**, *234*, 849–855.

[118] Thompson, M. J.; Eisenberg, D. *Journal of Molecular Biology* **1999**, *290*, 595 – 604.

[119] Zhou, H.-X. *Accounts of Chemical Research* **2004**, *37*, 123–130.

[120] Nagi, A. D.; Regan, L. *Folding and Design* **1997**, *2*, 67 – 75.

[121] Viguera, A.-R.; Serrano, L. *Nature Structural & Molecular Biology* **1997**, *4*, 939 – 946.

[122] Ladurner, A. G.; Fersht, A. R. *Journal of Molecular Biology* **1997**, *273*, 330 – 337.

[123] Grantcharova, V. P.; Riddle, D. S.; Baker, D. *Proceedings of the National Academy of Sciences* **2000**, *97*, 7084–7089.

[124] Hardy, F.; Vriend, G.; Vinne, B. v. d.; Frigerio, F.; Grandi, G.; Venema, G.; Ejjsink, V. G. *Protein Engineering* **1994**, *7*, 425–430.

[125] Daggett, V.; Levitt, M. *Journal of Molecular Biology* **1993**, *232*, 600 – 619.

[126] Couñago, R.; Chen, S.; Shamoo, Y. *Molecular Cell* **2006**, *22*, 441 – 449.

[127] Couñago, R.; Wilson, C. J.; Pea, M. I.; Wittung-Stafshede, P.; Shamoo, Y. *Protein Engineering Design and Selection* **2008**, *21*, 19–27.

[128] Miller, C.; Davlieva, M.; Wilson, C.; White, K. I.; Couñago, R.; Wu, G.; Myers, J. C.; Wittung-Stafshede, P.; Shamoo, Y. *Biophysical Journal* **2010**, *99*, 887 – 896.

[129] Sawle, L.; Ghosh, K. *J. Chem. Phys.* **2015**, *143*, 085101.

[130] Sawle, L.; Ghosh, K. *Journal of Chemical Theory and Computation* **2016**, *12*, 861–869.

[131] Kumar, S.; Nussinov, R. *Cell. Mol. Life. Sci.* **2001**, *58*, 1216–1233.

[132] Razvi, A.; Scholtz, J. *Protein Science* **2006**, *15*, 1569–1578.

[133] England, J.; Shakhnovich, B.; Shakhnovich, E. *Proc. Natl. Acad. Sci.* **2003**, *100*, 8727–8731.

[134] Berezovsky, I.; Shakhnovich, E. *Proc. Natl. Acad. Sci.* **2005**, *102*, 12742–12747.

[135] Ladenstein, R. *Biotechnology & Biotechnological Equipment* **2008**, *22*, 612–619.

[136] Jaenicke, R.; Bohm, G. *Current Opinion in Structural Biology* **1998**, *8*, 738–748.

[137] Singleton, R.; Amelunxen, R. *Bacteriological Reviews* **1973**, *37*, 320–342.

[138] Razvi, A.; Scholtz, J. *Biochemistry* **2006**, *45*, 4084–4092.

[139] Graziano, G. *Journal of Theoretical Analysis and Calorimetry* **2008**, *93*, 429–438.

[140] Robertson, A.; Murphy, K. *Chemical Reviews* **1997**, *97*, 1251–1267.

[141] Ghosh, K.; Dill, K. *Proc. Natl. Acad. Sci.* **2009**, *106*, 10649–10654.

[142] Privalov, P.; NN, K. *J. Mol. Biol.* **1974**, *86*, 665–684.

[143] Privalov, P. *Adv. Prot. Chem.* **1979**, *33*, 167–241.

[144] Baldwin, R. *Proc. Natl. Acad. Sci.* **1986**, *83*, 8069–8072.

[145] Doig, A.; William, D. *Biochemistry* **1992**, *31*, 9371–9375.

[146] Murphy, K.; Privalov, P.; Gill, S. *Science* **1990**, *247*, 559–561.

[147] Fu, L.; Freire, E. *Proc. Natl. Acad. Sci.* **1992**, *89*, 9335–9338.

[148] Murphy, K.; Gill, S. *J. Mol. Biol.* **1991**, *222*, 699–709.

[149] Kumar, S.; Nussinov, R. *Biophysical Chemistry* **2004**, *111*, 235–246.

[150] Kumar, S.; Tsai, C.; Nussinov, R. *Biochemistry* **2003**, *42*, 4864–4873.

[151] Shiraki, K.; Nishikori, S.; Fujiwara, S.; Hashimoto, H.; Kai, M., Yand Takagi; Imanaka, T. *European Journal Of Biochemistry* **2001**, *268*, 4144–4150.

[152] Wallgreen, M.; Aden, J.; Pylypenko, O.; Mikaelsson, T.; Johansson, L.; Rak, A.; Wolf-Watz, M. *J. Mol. Biol.* **2008**, *379*, 845–858.

[153] Chen, C.-H.; Roth, L.; MacColl, R.; Berns, D. *Biophysical Chemistry* **1994**, *50*, 313–321.

[154] Clarke, J.; Hounslow, A.; Bond, C.; Fersht, A.; Daggett, V. *Protein Science* **2000**, *9*, 2394–2404.

[155] Rosato, V.; Pucello, N.; Giuliano, G. *TRENDS in Genetics* **2002**, *18*, 278–281.

[156] Pace, C.; Alston, R.; Shaw, K. *Protein Science* **2000**, *9*, 1395–1398.

[157] Cho, J.-H.; Sato, S.; Raleigh, D. *J. Mol. Biol.* **2004**, *338*, 827–837.

[158] Ge, M.; Xia, X.-Y.; Pan, X.-M. *J. Biol. Chem.* **2008**, *283*, 31690–31696.

[159] Liu, C.; Yang, Y.; LiCata, V. *Biophysical J* **2009**, *96*, 330a–330a.

[160] Fitter, J.; Haber-Pohlmeier, S. *Biochemistry* **2004**, *43*, 9589–9599.

[161] Shortle, D.; Chan, H.; Dill, K. *Prot. Sci.* **1992**, *1*, 201–215.

[162] Dill, K.; Shortle, D. *Ann. Rev. Biochem.* **1991**, *60*, 795–825.

[163] Miyazawa, S.; Jernigan, R. *Protein Engineering* **1994**, *7*, 1209–1220.

[164] Shortle, D. *FASEB Journal* **1996**, *10*, 27–34.

[165] Shortle, D.; Ackerman, M. *Science* **2001**, *293*, 487–489.

[166] Lindorff-Larsen, K.; Kristjansdottir, S.; Teilum, K.; Fieber, W.; Dobson, C.; Poulsen, F.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 3291–3299.

[167] Bowler, B. *Molecular Biosystems* **2007**, *3*, 88–99.

[168] Pace, C.; Huyghues-Despointes, B.; Fu, H.; Takano, K.; Scholtz, J.; Grimsley, G. *Protein Science* **2010**, *19*, 929–943.

[169] Scott, K.; Alonso, D.; Sato, S.; Fersht, A.; Daggett, V. *Proc. Natl. Acad. Sci.* **2007**, *104*, 2661–2666.

[170] Watanabe, K.; Suzuki, Y. *J Mol. Cat* **1998**, *4*, 167–180.

[171] Trevino, S.; Schaefer, S.; Scholtz, J.; Pace, C. *J. Mol. Biol.* **2007**, *373*, 211–218.

[172] Nemethy, G.; Leach, S.; Scheraga, H. *J. Phys. Chem.* **1966**, *70*, 998–1004.

[173] Berezovsky, I.; Chen, W.; Choi, P.; Shakhnovich, E. *PLOS Comp. Bio.* **2005**, *1*, 0322–0332.

[174] Livesay, D.; Jacobs, D. *Proteins-Structure Function and Bioinformatics* **2006**, *62*, 130–143.

[175] Anfinsen, C. *Science* **1973**, *181*, 4096.

[176] Tompa, P. *Current opinion in structural biology* **2011**, *21*, 419–425.

[177] Tompa, P. *Trends Biochem Sci.* **2012**, *37*, 509–516.

[178] Uversky, V. *The International Journal of Biochemistry & Cell Biology* **2011**, *43*, 1090–1103.

[179] Das, A.; Sin, B.; Mohazab, A.; Plotkin, S. *J. Chem. Phys.* **2013**, *139*, 121925.

[180] Lindorff-Larsen, K.; Trbovic, N.; Maragakis, P.; Piana, S.; Shaw, D. *J. Am. Chem. Soc.* **2012**, *134*, 3787–3791.

[181] Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Current Opinion in Structural Biology* **2009**, *19*, 120–127.

[182] Voelz, V.; Singh, V.; Wedenmeyer, W.; Lapidus, L.; Pande, V. *J. Am. Chem. Soc.* **2010**, *132*, 4702–4709.

[183] Sweet, J.; Nowling, R.; Cickovski, T.; Sweet, C.; Pande, V.; Izaguirre, J. *J. Chem. Theory. Comput* **2013**, *9*, 3267–3281.

[184] Vitalis, A.; Pappu, R. *Annual Reports in Computational Chemistry* **2009**, *5*, 49–76.

[185] Das, R.; Pappu, R. *Proc. Natl. Acad. Sci.* **2013**, *110*, 13392–13397.

[186] Edwards, S.; Singh, P. *J. Chem. Soc. Faraday Transaction 2* **1979**, *75*, 1020–1029.

[187] Muthukumar, M. *J. Chem. Phys.* **1987**, *86*, 7230–7235.

[188] Ha, B.; Thirumalai, D. *Macromolecules* **1995**, *28*, 577–581.

[189] Ha, B.; Thirumalai, D. *Physical Review A* **1992**, *46*, R3012–R3015.

[190] Ghosh, K.; Carri, G.; Muthukumar, M. *J. Chem. Phys.* **2001**, *115*, 4367–4375.

[191] Ghosh, K.; Muthukumar, M. *J. Polymer Science, B Polymer Physics* **2001**, *39*, 2644–2652.

[192] Higgs, P.; Joanny, J. *J. Chem. Phys.* **1991**, *94*, 1543–1554.

[193] Chan, H.; Dill, K. *J. Chem. Phys.* **1994**, *100*, 9238–9257.

[194] Socci, N.; Onuchic, J. *J. Chem. Phys.* **1995**, *103*, 4732–4744.

[195] Shakhnovich, E. *Phys. Rev. Lett.* **1994**, *72*, 3907–3910.

[196] Chakraborty, A. *Physics Reports* **2001**, *342*, 1–61.

[197] Ha, B.; Thirumalai, D. *J. De. Physique II* **1997**, *7*, 887–902.

[198] Gutin, A.; Shakhnovich, E. *Physical Review E* **1994**, *50*, R3322.

[199] Dobrynin, A.; Colby, R.; Rubinstein, M. *J. Polym. Sci. B* **2004**, *42*, 3513–3538.

[200] Dobrynin, A.; Rubinstein, M. *Prog. Polym. Sci.* **2005**, *30*, 1049–1118.

[201] Srivastava, D.; ; Muthukumar, M. *Macromolecules* **1996**, *29*, 2324–2326.

[202] Wallace, D.; Dill, K. *Biopolymers* **1996**, *39*, 115–127.

[203] Dill, K. A. *Biochem* **1985**, *24*, 1501–1509.

[204] Miyazawa, S.; Jernigan, R. *Macromolecules* **1985**, *18*, 534–552.

[205] Kohn, J.; Millett, I.; Jacob, J.; Zagrovic, B.; Dillon, T.; Cingel, N.; Dothager, R.; Seifert, S.; Thiyagarajan, P.; Sosnick, T.; Hasan, M.; Pande, V.; Ruczinski, I.; Doniach, S.; Plaxco, K. *Proc. Natl. Acad. Sci.* **2004**, *101*, 12491–12496.

[206] Ghosh, K.; Ozkan, S. B.; Dill, K. A. *Journal of the American Chemical Society* **2007**, *129*, 11920–11927.

[207] Ghosh, K.; Dill, K. A. *Journal of the American Chemical Society* **2009**, *131*, 2306–2312.

[208] Rustad, M.; Ghosh, K. *J. Chem. Phys.* **2012**, *137*, 205104.

[209] Doi, M.; Edwards, S. *The theory of polymer dynamics*; Oxford Science Publications: Oxford, 1986.

[210] Muthukumar, M. *J. Chem. Phys.* **2004**, *120*, 9343.

[211] Dill, K.; Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology*; Garland Science: New York, 2003.

[212] Muller-Spath, S.; Soranno, A.; V, H.; Hofmann, H.; Ruegger, S.; Reymond, L.; D, N.; Schuler, B. *Proc. Natl. Acad. Sci.* **2010**, *107*, 14609–14614.

[213] Mao, A.; Crick, S.; Vitalis, A.; Chicoine, C.; Pappu, R. *Proc. Natl. Acad. Sci.* **2010**, *107*, 8183–8188.

[214] Myers, J.; Pace, C.; Scholtz, J. *Protein Science* **1995**, *4*, 2138–2148.

[215] Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. *Proc. Natl. Acad. Sci.* **2012**, *109*, 16155–16160.

[216] Aznauryan, M.; Nettles, D.; Holla, A.; Hofmann, H.; Schuler, B. *J. Am. Chem. Soc.* **2013**, *135*, 14040–14043.

[217] Li, Y.; Middaugh, C. R.; Fang, J. *BMC Bioinformatics* **2010**, *11*, 1–11.

[218] Robinson-Rechavi, M.; Godzik, A. *Structure* **2005**, *13*, 857 – 860.

[219] Grossfield, A.; Zuckerman, D. M. *Annual reports in computational chemistry* **2009**, *5*, 23–48.

[220] Lyman, E.; Zuckerman, D. *Biophysical Journal* **2006**, *91*, 164–172.

[221] Genheden, S.; Ryde, U. *Physical Chemistry Chemical Physics* **2012**, *14*, 8662–8677.

[222] Lyman, E.; Zuckerman, D. M. *The Journal of Physical Chemistry B* **2007**, *111*, 12876–12882.

[223] Romo, T. D.; Grossfield, A. *Journal Chemical Theory and Computation* **2011**, *7*, 2464–2472.

[224] Smith, L. J.; Daura, X.; van Gunsteren, W. F. *Proteins: Structure, Function, and Bioinformatics* **2002**, *48*, 487–496.

[225] Hess, B. *Phys. Rev. E* **2002**, *65*, 031910.

[226] Faraldo-Gómez, J. D.; Forrest, L. R.; Baaden, M.; Bond, P. J.; Domene, C.; Patargias, G.; Cuthbertson, J.; Sansom, M. S. *Proteins: Structure, Function, and Bioinformatics* **2004**, *57*, 783–791.

[227] Brigo, A.; Lee, K. W.; Mustata, G. I.; Briggs, J. M. *Biophysical Journal* **2005**, *88*, 3072 – 3082.

[228] Grossfield, A.; Feller, S. E.; Pitman, M. C. *Proteins: Structure, Function, and Bioinformatics* **2007**, *67*, 31–40.

[229] Arcangeli, C.; Cantale, C.; Galeffi, P.; Rosato, V. *Journal of Structural Biology* **2008**, *164*, 119 – 133.

[230] Balasubramanian, C.; Chillemi, G.; Abbate, I.; Capobianchi, M. R.; Rozera, G.; Desideri, A. *Journal of Biomolecular Structure and Dynamics* **2012**, *29*, 879–891.

[231] Arcangeli, C.; Cantale, C.; Galeffi, P.; Gianese, G.; Paparcone, R.; Rosato, V. *Journal of Biomolecular Structure and Dynamics* **2008**, *26*, 35–47.

[232] Zou, T.; Risso, V.; Gavira, J.; Ruiz, J.; Ozkan, S. B. *Molecular Biology and Evolution* **2015**, *23*.

[233] Kumar, A.; Glembo, T. J.; Ozkan, S. B. *Biophysical Journal* **2015**, *109*, 1273–1281.

[234] Lin, E.; Shell, S. M. *Journal Chemical Theory and Computation* **2009**, *5*, 2062.

[235] Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K. *Biophysical Journal* **1997**, *72*, 1568 – 1581.

[236] Pang, Y.-P. *Proteins: Structure, Function, and Bioinformatics* **2004**, *57*, 747–757.

[237] Papaleo, E.; Fantucci, P.; ; Gioia, L. D. *Journal Chemical Theory and Computation* **2005**, *1*, 1286–1297.

[238] Gsponer, J.; Ferrara, P.; Caflisch, A. *Journal of Molecular Graphics and Modelling* **2001**, *20*, 169 – 182.

[239] Legge, F.; Budi, A.; Treutlein, H.; Yarovsky, I. *Biophysical Chemistry* **2006**, *119*, 146–157.

[240] Thomas, A. S.; Mao, S.; Elcock, A. H. *Biophysical Journal* **2013**, *105*, 732–744.

[241] Karplus, M.; McCammon, J. A. *Nature Structural Biology* **2002**, *9*, 646–652.

[242] Karplus, M.; Kuriyan, J. *Proceedings of the National Academy of Sciences* **2005**, *102*, 6679–6685.

[243] Isralewitz, B.; Gao, M.; Schulten, K. *Current Opinion in Structural Biology* **2001**, *11*, 224 – 230.

[244] Isralewitz, B.; Baudry, J.; Gullingsrud, J.; Kosztin, D.; Schulten, K. *Journal of Molecular Graphics and Modelling* **2001**, *19*, 13 – 25.

[245] Basu, S.; Sen, S. *Journal of Chemical Information and Modeling* **2013**, *53*, 423–434.

[246] Wintrode, P. L.; Zhang, D.; Vaidehi, N.; Arnold, F. H.; III, W. A. G. *Journal of Molecular Biology* **2003**, *327*, 745 – 757.

[247] Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. *Science* **2010**, *330*, 341–346.

[248] Keller, B.; Daura, X.; van Gunsteren, W. F. *Journal of Chemical Physics* **2010**, *132*.

[249] Jin, C.; Li, Y. DOI: 10.2210/pdb2ho9/pdb, PDB ID: 2HO9.

[250] Quezada, C. M.; Grădinaru, C.; Simon, M. I.; Bilwes, A. M.; Crane, B. R. *Journal of Molecular Biology* **2004**, *341*, 1283 – 1294.

[251] Miyazaki, T.; Mega, R.; Kim, K.; Shikai, A.; Masui, R.; Kuramitsu, S.; Nakagawa, N. DOI: 10.2210/pdb3a0j/pdb, PDB ID: 3A0J.

[252] Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. *Nucleic Acids Research* **2012**, *40*, W537–W541.

[253] Myers, J.; Grothaus, G.; Narayanan, S.; Onufriev, A. *Proteins: Structure, Function, and Bioinformatics* **2006**, *63*, 928–938.

[254] Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. *Nucleic Acids Research* **2005**, *33*, W368–W371.

[255] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *Journal of Chemical Physics* **1983**, *79*, 926–935.

[256] Neria, E.; Fischer, S.; Karplus, M. *Journal of Chemical Physics* **1996**, *105*, 1902–1921.

[257] Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Grand, S. L.; Walker, R. C. *Journal Chemical Theory and Computation* **2013**, *9*, 3878–3888.

[258] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Structure, Function, and Bioinformatics* **2006**, *65*, 712–725.

[259] Miyamoto, S.; Kollman, P. A. *Journal of Computational Chemistry* **1992**, *13*, 952–962.

[260] Darden, T.; York, D.; Pedersen, L. *Journal of Chemical Physics* **1993**, *98*, 10089–10092.

[261] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *Journal of Chemical Physics* **1995**, *103*, 8577–8593.

[262] Hockney, R. W.; Eastwood, J. W. *Computer Simulation Using Particles*; Taylor & Francis, Inc.: Bristol, PA, USA, 1988.

[263] Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Grand, S. L.; Walker, R. C. *Journal Chemical Theory and Computation* **2012**, *8*, 1542–1555.

[264] Roe, D. R.; Thomas E. Cheatham, I. *Journal Chemical Theory and Computation* **2013**, *9*, 3084–3095.

[265] Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. *Journal Chemical Theory and Computation* **2007**, *3*, 2312–2334.

[266] Sittel, F.; Jain, A.; Stock, G. *Journal of Chemical Physics* **2014**, *141*, 014111.

[267] Zhang, X.; Bhatt, D.; Zuckermann, D. *Journal Chemical Theory and Computation* **2010**, *6*, 3048–3057.

[268] Romo, T.; Grossfield, A. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2009.

[269] Vendruscolo, M.; Kussell, E.; Domany, E. *Folding and Design* **1997**, *2*, 295–306.

[270] Chen, T.; Chan, H. *PLoS Comput Biol* **2015**, *11*, e1004260.

[271] Romo, T. D.; Grossfield, A. *Biophysical Journal* **2014**, *106*, 1553 – 1554.

[272] Neale, C.; Hsu, J. C.; Yip, C. M.; Pomés, R. *Biophysical Journal* **2014**, *106*, L29 – L31.

[273] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Research* **2000**, *28*, 235–242.

[274] Kumar, M. D. S.; Bava, K. A.; Gromiha, M. M.; Prabakaran, P.; Kitajima, K.; Uedaira, H.; Sarai, A. *Nucleic Acids Research* **2006**, *34*, D204–D206.

[275] Cheung, Y.-Y.; Lam, S. Y.; Chu, W.-K.; Allen, M. D.; Bycroft, M.; ; Wong, K.-B. *Biochemistry* **2005**, *44*, 4601–4611.

[276] Quezada, C. M.; Grdinaru, C.; Simon, M. I.; Bilwes, A. M.; Crane, B. R. *Journal of Molecular Biology* **2004**, *341*, 1283 – 1294.

[277] Filimonov, V. V.; Prieto, J.; Martinez, J. C.; Bruix, M.; Mateo, P. L.; Serrano, L. *Biochemistry* **1993**, *32*, 12906–12921.

[278] Deutschman, W. A.; ; Dahlquist, F. W. *Biochemistry* **2001**, *40*, 13107–13113.

[279] Petrosian, S. A.; Makhatadze, G. I. *Protein Science* **2000**, *9*, 387–394.

[280] Nicholson, E. M.; Scholtz, J. M. *Biochemistry* **1996**, *35*, 11369–11378.

[281] Razvi, A.; ; Scholtz, J. M. *Biochemistry* **2006**, *45*, 4084–4092.

[282] Burmann, B. M.; Luo, X.; Rösch, P.; Wahl, M. C.; Gottesman, M. E. *Nucleic Acids Research* **2010**, *38*, 314–326.

[283] Hollien, J.; Marqusee, S. *Biochemistry* **1999**, *38*, 3831–3836.

[284] Niranjanakumari, S.; Kurz, J. C.; Fierke, C. A. *Nucleic Acids Research* **1998**, *26*, 3090–3096.

[285] Perez-Jimenez, R.; Inglés-Prieto, A.; Zhao, Z.-M.; Sanchez-Romero, I.; Alegre-Cebollada, J.; Kosuri, P.; Garcia-Manyes, S.; Kappock, T. J.; Tanokura, M.; Holmgren, A.; Sanchez-Ruiz, J. M.; Gaucher, E. A.; Fernandez, J. M. *Nature Structural & Molecular Biology* **2011**, *18*, 592–596.

[286] Hiraoki, T.; Brown, S. B.; Stevenson, K. J.; Vogel, H. J. *Biochemistry* **1988**, *27*, 5000–5008.

[287] Fan, H.; Mark, A. E. *Proteins: Structure, Function, and Bioinformatics* **2003**, *53*, 111–120.

[288] Cox, K.; Bond, P.; Grottesi, A.; Baaden, M.; Sansom, M. *European Biophysics Journal* **2008**, *37*, 131–141.

[289] Hu, J.; Li, D.; Su, X.; Jin, C.; Xia, B. DOI: 10.2210/pdb3br8/pdb, PDB ID: 3BR8.

[290] Cheung, Y.-Y.; Lam, S. Y.; Chu, W.-K.; Allen, M. D.; Bycroft, M.; ; Wong, K.-B. *Biochemistry* **2005**, *44*, 4601–4611.

[291] Mourey, L.; Da Re, S.; Pdelacq, J.-D.; Tolstykh, T.; Faurie, C.; Guillet, V.; Stock, J. B.; Samama, J.-P. *Journal of Biological Chemistry* **2001**, *276*, 31074–31082.

[292] Griswold, I.; Zhou, H.; Matison, M.; Swanson, R.; McIntosh, L.; Simon, M.; Dahlquist, F. *Nature Structural Biology* **2002**, *9*, 121–125.

[293] Volz, K.; Matsumura, P. *Journal of Biological Chemistry* **1991**, *266*, 15511–15519.

[294] Usher, K. C.; De La Cruz, A. F. A.; Dahlquist, F. W.; James Remington, S.; Swanson, R. V.; Simon, M. I. *Protein Science* **1998**, *7*, 403–412.

[295] Schindelin, H.; Jiang, W.; Inouye, M.; Heinemann, U. *Proceedings of the National Academy of Sciences of the United States of America* **1994**, *91*, 5119–5123.

[296] Fujiwara, K.; Maita, N.; Hosaka, H.; Okamura-Ikeda, K.; Nakagawa, A.; Taniguchi, H. *Journal of Biological Chemistry* **2010**, *285*, 9971–9980.

[297] Nakai, T.; Ishijima, J.; Masui, R.; Kuramitsu, S.; Kamiya, N. *Acta Crystallographica Section D* **2003**, *59*, 1610–1618.

[298] Jia, Z.; Quail, J. W.; Waygood, E. B.; Delbaere, L. T. *Journal of Biological Chemistry* **1993**, *268*, 22490–501.

[299] Sridharan, S.; Razvi, A.; Scholtz, J. M.; Sacchettini, J. C. *Journal of Molecular Biology* **2005**, *346*, 919 – 931.

[300] Luo, X.; Hsiao, H.-H.; Bubunenko, M.; Weber, G.; Court, D. L.; Gottesman, M. E.; Urlaub, H.; Wahl, M. C. *Molecular Cell* **2008**, *32*, 791 – 802.

[301] Bonin, I.; Robelek, R.; Benecke, H.; Urlaub, H.; Bacher, A.; Richter, G.; Wahl, M. C. *Biochemical Journal* **2004**, *383*, 419–428.

[302] Forouhar, F.; Lee, I.-S.; Vujcic, J.; Vujcic, S.; Shen, J.; Vorobiev, S. M.; Xiao, R.; Acton, T. B.; Montelione, G. T.; Porter, C. W.; Tong, L. *Journal of Biological Chemistry* **2005**, *280*, 40328–40336.

[303] Filippova, E.; Shuvalova, L.; Minasov, G.; Kiryukhina, O.; Zhang, Y.; Clancy, S.; Radhakrishnan, I.; Joachimiak, A.; Anderson, W. *Proteins: Structure, Function, and Bioinformatics* **2011**, *79*, 2566–2577.

[304] Katayanagi, K.; Miyagawa, M.; Matsushima, M.; Ishikawa, M.; Kanaya, S.; Nakamura, H.; Ikehara, M.; Matsuzaki, T.; Morikawa, K. *Journal of Molecular Biology* **1992**, *223*, 1029 – 1052.

[305] Ishikawa, K.; Okumura, M.; Katayanagi, K.; Kimura, S.; Kanaya, S.; Nakamura, H.; Morikawa, K. *Journal of Molecular Biology* **1993**, *230*, 529 – 542.

[306] Stams, T.; Niranjanakumari, S.; Fierke, C. A.; Christianson, D. W. *Science* **1998**, *280*, 752–755.

[307] Reiter, N. J.; Osterman, A.; Torres-Larios, A.; Swinger, K. K.; Pan, T.; Mondragon, A. *Nature* **2010**, *468*, 784–789.

[308] Kovacs, H.; Comfort, D.; Lord, M.; Campbell, I. D.; Yudkin, M. D. *Proceedings of the National Academy of Sciences* **1998**, *95*, 5067–5071.

[309] Etezady-Esfarjani, T.; Placzek, W. J.; Herrmann, T.; Wthrich, K. *Magnetic Resonance in Chemistry* **2006**, *44*, S61–S70.

[310] Katti, S. K.; LeMaster, D. M.; Eklund, H. *Journal of Molecular Biology* **1990**, *212*, 167 – 184.

[311] Ebihara, A.; Yanai, H.; Yokoyama, S.; Kuramitsu, S. DOI: 10.2210/pdb2yzu/pdb, PDB ID: 2YZU.

[312] Schafmeister, C. E. A. F.; Ross, W. S.; Romanovski, W. S. LEAP. 1995.

[313] Case, D. et al. AMBER 2014. 2014.

[314] Kumar, S.; Nussinov, R. *Biophysical Journal* **2002**, *83*, 1595–1612.

[315] Brinda, K.; Vishveshwara, S. *Biophysical J.* **2005**, *89*, 4159–4170.

[316] Aguilar, C. F.; Sanderson, I.; Moracci, M.; Ciaramella, M.; Nucci, R.; Rossi, M.; Pearl, L. H. *Journal of Molecular Biology* **1997**, *271*, 789–802.

[317] Cavagnero, S.; Debe, D.; Zhou, Z.; Adams, M.; Chan, S. *Biochemistry* **1998**, *37*, 3369–3376.

[318] Struvay, C.; Negro, S.; Matagne, A.; Feller, G. *Biochemistry* **2013**, *52*, 2982–2990.

[319] Li, L.; Li, C.; Sarkar, S.; Zhang, J.; Witham, S.; Zhang, Z.; Wang, L.; Smith, N.; Petukh, M.; Alexov, E. *BMC Biophysics* **2012**, *5*, 1–11.

[320] Dominy, B.; Perl, D.; Schmid, F.; CL, B. *J. Mol. Biol.* **2002**, *319*, 541–554.

[321] Loladze, V.; Makhatadze, G. *Protein Sci.* **2002**, *11*, 174–177.

[322] Strickler, S.; Gribenko, A.; Gribenko, A.; Keiffer, T.; Tomlinson, J.; Reihle, T.; Loladze, V.; Makhatadze, G. *Biochemistry* **2006**, *45*, 2761–2766.

[323] Tzul, F.; Schweiker, K.; Makhatadze, G. *Proc. Natl. Acad, Sci.* **2015**, *112*, E259–266.

[324] De Graff, A.; Hazoglou, M.; Dill, K. *Structure* **2016**, *24*, 1–8.

[325] Karlin, S.; Blaisdell, B.; Brendel, V. *Methods in Enzymology* **1990**, *183*, 388–402.

[326] Karlin, S. *Current Opinion in Structural Biology* **1995**, *5*, 360–371.

[327] Motono, C.; Gromiha, M.; Kumar, S. *Proteins: Structure, Function, and Bioinformatics* **2007**, *71*, 655–669.

[328] Merkley, E.; Parson, W.; Daggett, V. *Protein Engineering, Design and Selection* **2010**, *23*, 327–336.

[329] Roca, M.; Liu, H.; Messer, B.; Warshel, A. *Biochemistry* **2007**, *46*, 15076–15088.

[330] Bershtein, S.; Choi, J.-M.; Bhattacharyya, S.; Budnik, B.; Shakhnovich, E. *Cell Reports* **2015**, *11*, 645 – 656.

[331] Ghosh, K.; Dill, K. *Biophysical Journal* **2010**, *99*, 3996–4002.

[332] Dill, K. A.; Ghosh, K.; Schmit, J. D. *Proceedings of the National Academy of Sciences* **2011**, *108*, 17876–17882.

[333] Zhang, J. *Trends in Genetics* **2000**, *16*, 107–109.

[334] Chen, P.; Shakhnovich, E. *Biophysical Journal* **2010**, *98*, 1109–1118.

[335] Ratkowsky, D.; Olley, J.; Ross, T. *Journal of Theoretical Biology* **2005**, *233*, 351–362.

[336] Mohr, P.; Krawiec, S. *Journal of General Microbiology* **1980**, *121*, 311–317.

[337] Salter, M.; Ross, T.; McMeekin, T. *Journal of Applied Microbiology* **1998**, *85*, 357–364.

[338] Nichols, D.; Presser, K.; Olley, J.; Ross, T.; McMeekin, T. *Applied and Environmental Microbiology* **2002**, *68*, 2809–2813.

[339] Darland, G.; Brock, T. *Journal of General Microbiology* **1971**, *67*, 9–15.

[340] Zeikus, J.; Wolfe, R. *Journal of Bacteriology* **1972**, *109*, 707–&.

[341] Zeikus, J.; Hegge, P.; Anderson, M. *Archives of Microbiology* **1979**, *122*, 41–48.

[342] Belkin, S.; Wirsen, C.; Jannasch, H. *Applied and Environmental Microbiology* **1986**, *51*, 1180–1185.

[343] Oshima, T.; Imahori, K. *International Journal of Systematic Bacteriology* **1974**, *24*, 102–112.

[344] Benson, D.; Karsch-Mizrachi, I.; Lipman, D.; Ostell, J.; Wheeler, D. *Nucleic Acids Research* **2005**, *33*.

[345] Dietz, K. *Bioessays* **2005**, *27*, 1097–1101.

[346] Demetrius, L.; Legendre, S.; Harremoees, P. *Bulletin of Mathematical Biology* **2009**, *71*, 800–818.

Appendix A

Convergence

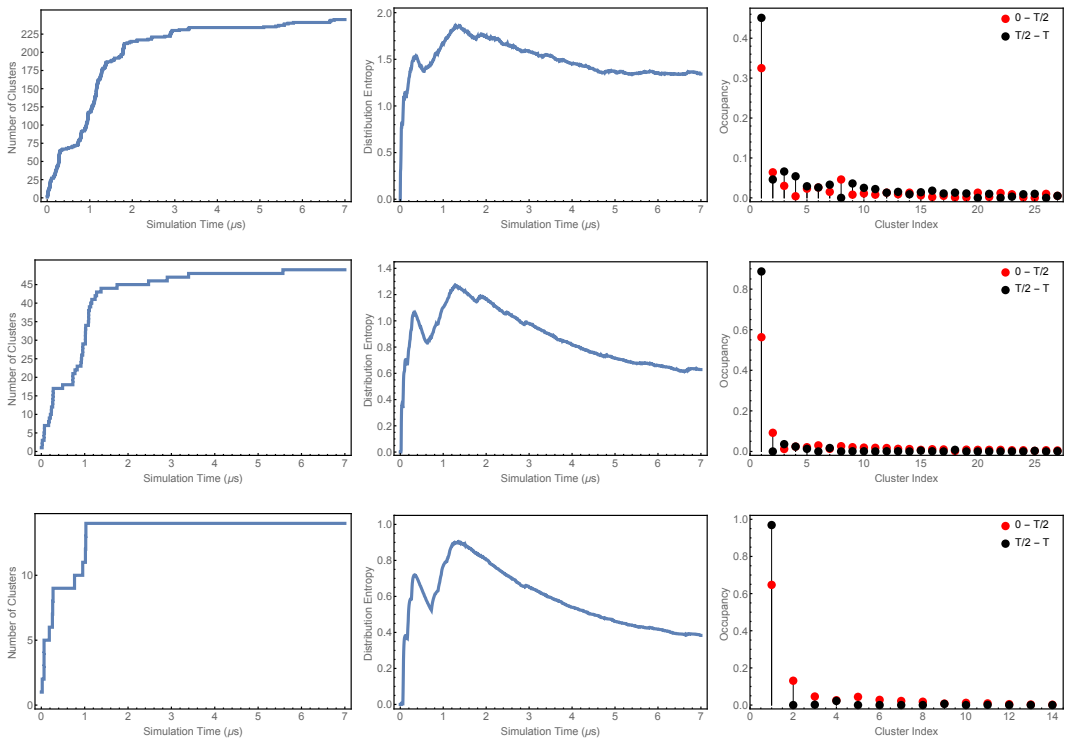

Figure A.1: The number of found clusters, cluster distribution entropy, and cluster distribution histogram for different choices of d_c used in clustering. Top row was found with a d_c choice of 1.25Å, the middle row used 1.5Å (identical to Figure 4.1), and $d_c = 1.75$Å in the bottom row. As seen, the three cutoff qualitatively produce similar outcomes.

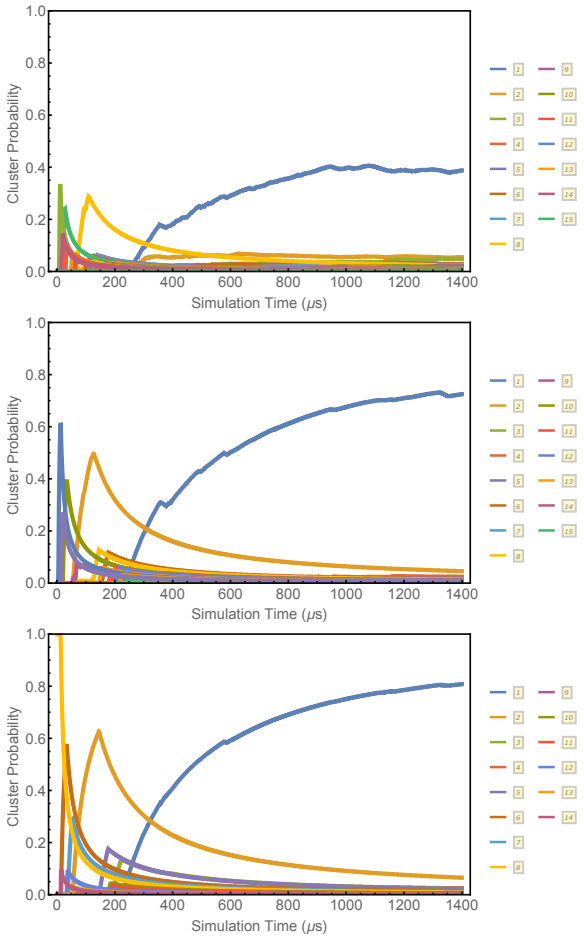

Figure A.2: Probabilities of the top 20 most occupied clusters as a function of simulation time for the 7μs simulation of CheW. The number of clusters, and their respective populations, are found using a $d_c = 1.25$Å (top), 1.5Å (middle, and identical to Figure 4.3, and 1.75Å (bottom).

System	τ_d (ns)	% error	τ_d' (ns)	% error
CheW 7μs	106–171	12.3–15.6	141.7	14.2
CheA 1μs	46–91	21.4–30.2	42.7	20.7
CheA 2μs	47–87	15.3–20.9	77.0	19.6
CSP Imp 2.5μs	143–163	23.9–25.5	99.6	19.9
CSP Imp 6μs	223–358	19.3–24.4	227.2	19.5
CSP Imp 20μs	606–1015	17.4–22.5	825.0	20.3
BPTI 2μs	11-16	7.4–8.9	3.7	4.3
BPTI 15μs	21-53	3.7–5.9	5.6	1.9
BPTI 30μs	26–53	2.9–4.2	31.5	3.2
CSP Exp 1μs	67–95	25.9–30.8	34.1	18.5
CSP Exp 3μs	77–120	16.0–20.0	59.7	14.1
CSP Exp 15μs	498–607	18.2–20.1	588.9	19.8

Table A.1: Both calculations of decorrelation time included all protein atoms except hydrogens. τ_d range is from subsample sizes of 2, 3, and 4, and utilized 25 repetitions. τ_d' is found from the average total sample size calculated from 100 iterations and 20 bins, and clustering did not always reduce to top two-states.

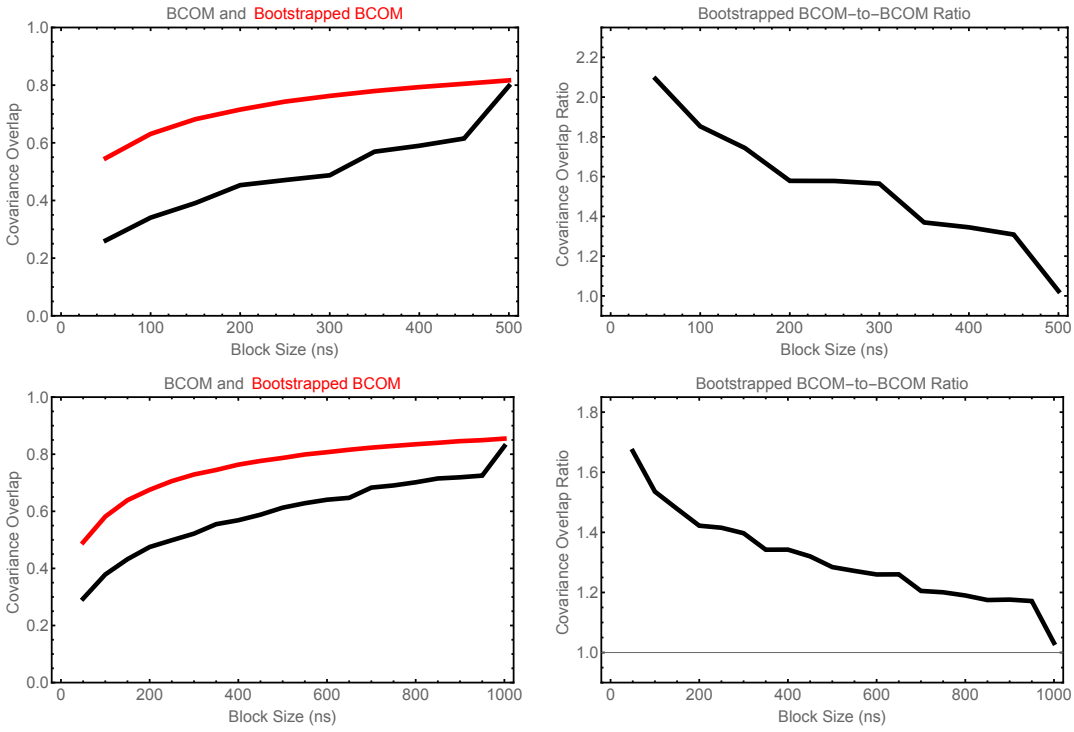

Figure A.3: Results of Block Covariance Overlap Method (BCOM) analysis of CheA 1μs (top) and 2μs (bottom) trajectories. Left panel is direct calculation of covariance overlap using contiguous blocks (black curve) and bootstrapping (red curve) at given block sizes. Right panel shows the ratio of bootstrapped BCOM to blocked BCOM.

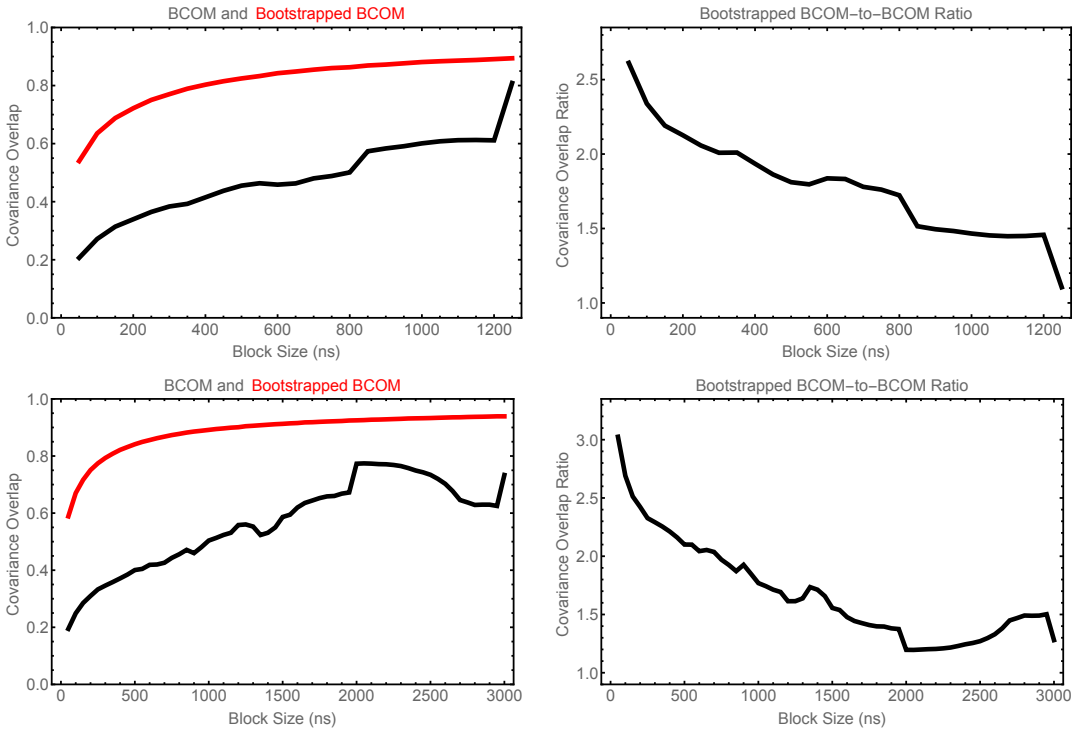

Figure A.4: Results of Block Covariance Overlap Method (BCOM) analysis of implicit CSP 2.5μs (top) and 6μs (bottom) trajectories. Left panel is direct calculation of covariance overlap using contiguous blocks (black curve) and bootstrapping (red curve) at given block sizes. Right panel shows the ratio of bootstrapped BCOM to blocked BCOM.

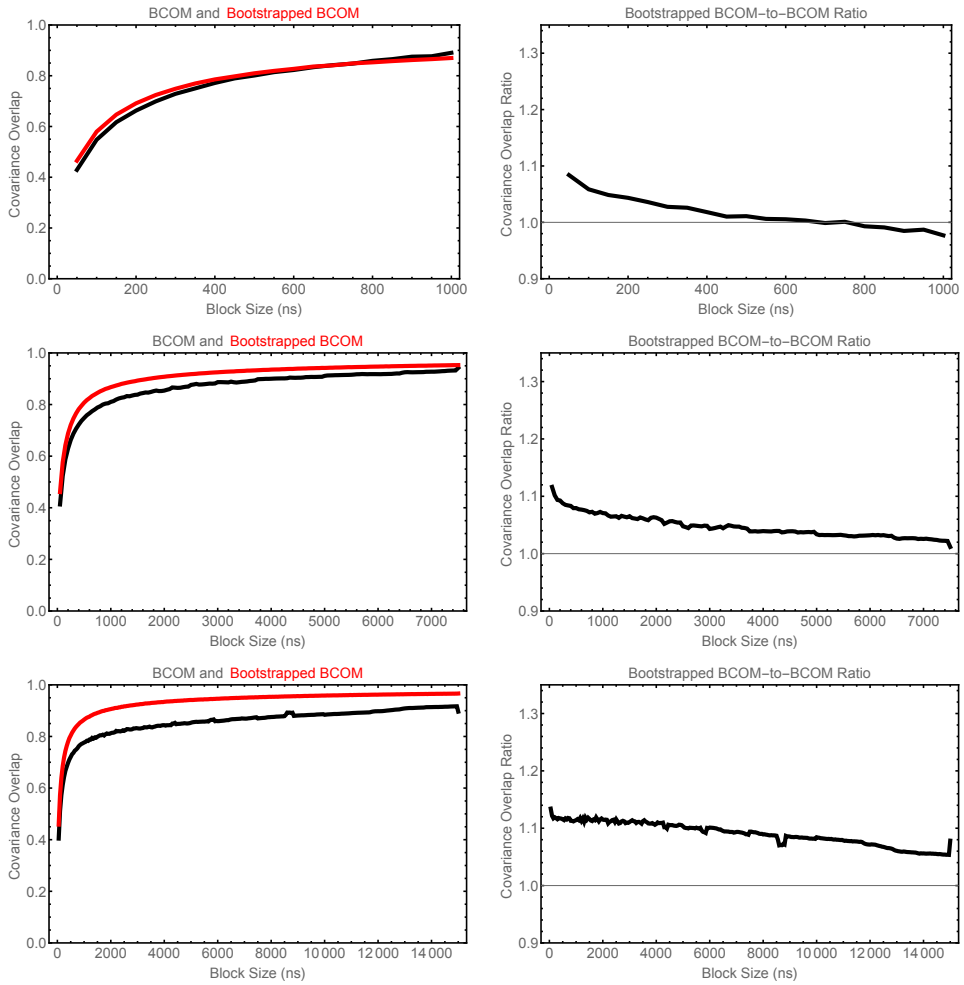

Figure A.5: Results of Block Covariance Overlap Method (BCOM) analysis of BPTI 2μs (top), 15μs (middle), and 30μs (bottom) trajectories. Left panel is direct calculation of covariance overlap using contiguous blocks (black curve) and bootstrapping (red curve) at given block sizes. Right panel shows the ratio of bootstrapped BCOM to blocked BCOM.

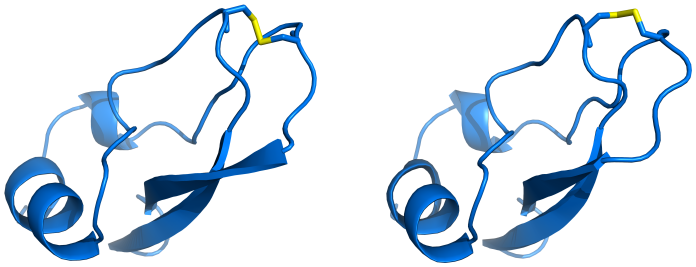

Figure A.6: Representative states for the two most occupied clusters ($d_c = 1.25$Å) from the 800μs BPTI simulation. These two clusters account for almost 60% (left representative, 20%, and right, 39%) of all possible trajectory frames. The most noticeable difference is in the chirality of the shown Cys14-Cys38 disulfide bridge (yellow sticks).

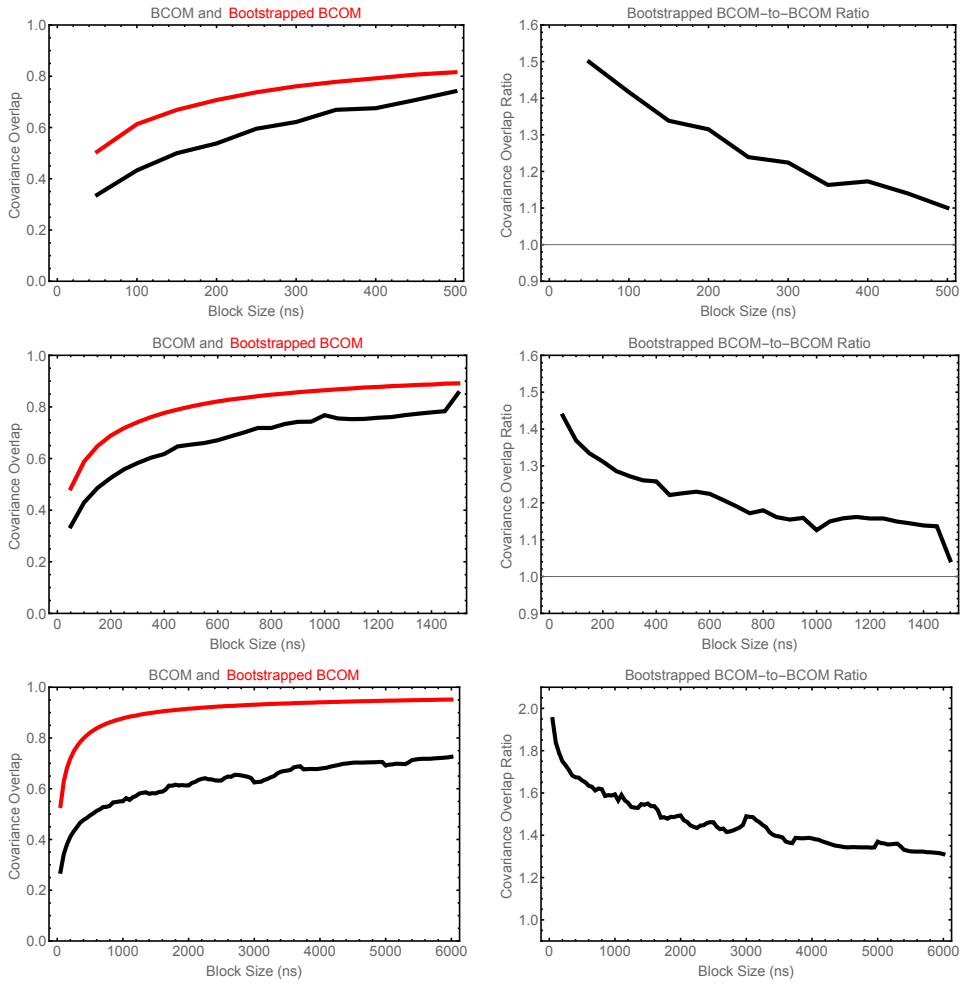

Figure A.7: Results of Block Covariance Overlap Method (BCOM) analysis of Explicit CSP 1μs (top), 3μs (middle), and 12μs (bottom) trajectories. Left panel is direct calculation of covariance overlap using contiguous blocks (black curve) and bootstrapping (red curve) at given block sizes. Right panel shows the ratio of bootstrapped BCOM to blocked BCOM.

Appendix B

DelPhi Benchmarking

To gather information on electrostatic properties, DelPhi utilizes a finite-difference method to solve the Poisson-Boltzmann equation. Benchmarking will validate the "black box" calculation, and will also serve to confirm the construction and choice of partial charge and radii libraries DelPhi demands to operate.

Comparison with Analytical Solutions

First, let's examine a system where an exact, analytical solution exist. Specifically, a collection of charges buried in a spherical cavity, of fixed radius and low dielectric, embedded within a medium of higher dielectric. Contributions to the solvation free energy (self energy) resulting from an individual charge q_i set a radial distance s_i inside a cavity of radius R and dielectric ϵ_{int}, and an exterior medium with dielectric ϵ_{ext} is expressed as:

$$U_i^{solv} = -166\frac{q_i^2}{R}\left(\frac{1}{\epsilon_{int}} - \frac{1}{\epsilon_{ext}}\right)\sum_{l=0}^{\infty}\frac{l+1}{l+1+l(\epsilon_{int}/\epsilon_{ext})}\left(\frac{s_i}{R}\right)^{2l} \tag{B.1}$$

Given the presence of a second charge, q_j, separated from q_i by a distance d_{ij} within the cavity, the free energy contributions from pairwise coulombic and polarization interactions can be written, respectively, as

$$U_{ij}^{coul} = 332\frac{q_i q_j}{d_{ij}\epsilon_{int}} \tag{B.2}$$

$$U_{ij}^{pol} = -332\frac{q_i q_j}{R}\left(\frac{1}{\epsilon_{int}} - \frac{1}{\epsilon_{ext}}\right)\sum_{l=0}^{\infty}\frac{l+1}{l+1+l(\epsilon_{int}/\epsilon_{ext})}\left(\frac{s_i s_j}{R^2}\right)^l P_l(cos\gamma) \tag{B.3}$$

where $P_l(cos\gamma)$ are the Legendre polynomials with an angle γ between q_i and q_j.

The total electrostatic energy of a system consisting of N charges is defined as the summation of solvation (self) energies and pairwise interaction energies (coulombic and polarization terms):

$$U^{total} = \sum_{i=1}^{N}\left(U_i^{solv} + \sum_{i<j}^{N} U_{ij}^{coul} + U_{ij}^{pol}\right) \tag{B.4}$$

Two Ion Model

Given the above exact solutions, we can construct a model system within DelPhi to test the viability and accuracy of the program's calculations. Specifically, the system will consist of two ions of opposite charge, $q_i = +e$ and $q_j = -e$, each placed on the x-axis a distance 10Å from the origin, but in opposing directions, giving $s_i = s_j = 10$Å, and $d_{ij} = 20$Å, inside of sphere of radius $R = 16$Å and $\epsilon_{int} = 2$. The external dielectric constant, ϵ_{ext}, was set to 78 to replicate a water solvent. This model system is visualized in Figure B.1.

Execution of DelPhi with the constructed two ion system, and necessary charge and radii libraries, returns a coulombic energy equivalent to U^{coul} (Eq. B.2), and a corrected reaction field energy (CRFE) term corresponding to the total solvation free

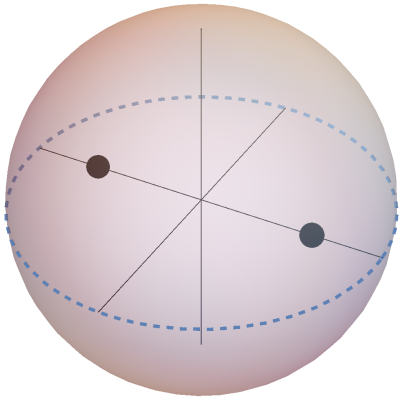

Figure B.1: An example model system where an exact solution exists. Two opposing charges (black spheres) placed on the x-axis equidistant from the origin within a spherical cavity of radius R and internal dielectric ϵ_{int}. Space external of the sphere has a dielectric ϵ_{ext}.

energy and pairwise polarization energy contributions. The total solvation free energy is calculated as the sum of all individual ion solvation free energy contribution modeled from Equation B.1. From aforementioned values of spherical radius, burial distance, charge, and dielectric constants, and the analytical expression of U^{solv} (Eq. B.1), the solvation free energy of an individual ion gives a value of $-13.9779\ kT$, an almost identical value from DelPhi calculation, $-13.9809\ kT$. The polarization free energy contribution was deduced by subtracting the total solvation energy summation from the CRFE output, Equation B.5. Free energies calculated for the two ion system from analytical expressions and DelPhi are shown in Table B.1.

Method	U^{solv} $[kT]$	U^{coul} $[kT]$	U^{pol} $[kT]$
Analytical	-27.9557	-14.0678	12.3753
DelPhi	-27.9618	-14.0499	12.3483

Table B.1: Contributions to the electrostatic free energy of the two ion model system shown in Figure B.1. U^{solv} is calculated from the summation of individual ion contributions (Eq. B.1). The U^{pol} value from DelPhi is found by subtracting U^{solv} from the CRFE output, Eq. B.5.

In summary, from comparison to analytical expressions, contributions to the electrostatic free energy calculated from DelPhi show very strong agreement. The pairwise coulombic energy, U^{coul}, is directly calculated, but to assess the pairwise polarization energy, U^{pol}, the outputted CRFE term needs to be reduced by the system's total solvation energy, which is simply the summation of all individual ion contributions in the absence of all other system ion charges. Therefore, a working algorithm utilizing DelPhi to calculate the electrostatic free energy contributions would proceed as follows:

1. For a system of N charges, calculate each charge's (q_i) solvation energy with all others ($q_{j\neq i}$) set to neutral, giving N solvation energy contributions.

2. From the N charge system, calculate the CRFE with all charges "turned on". This single execution of DelPhi will provide the pairwise coulombic energy, U^{coul}, as well.

3. Calculate the system's pairwise polarization energy by subtracting the total solvation energy of the N charges from the CRFE found in step 2.

$$U^{pol} = \text{CRFE} - \sum_{i}^{N} U_i^{solv} \tag{B.5}$$

Four Ion Model

As an extension to the two ion model system discussed above, two additional charges were added, giving four charges in the x-y plane equidistant from the origin (see Figure B.2), and all sharing the same charge, $q_{1-4} = +e$. All other parameters are kept from the two ion model, i.e., $R = 16$Å, $\epsilon_{int} = 2$, and $\epsilon_{ext} = 78$.

The working algorithm using DelPhi was implemented, and analytical solutions were solved using equations B.1-B.3, and B.5. Results from these two methods are

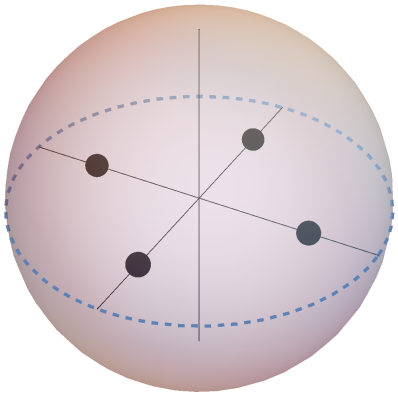

Figure B.2: Four ions of same charge (black spheres) placed in the x-y plane, each equidistant from the origin, and inside a sphere of radius R and internal dielectric ϵ_{int}. Space external of the sphere has a dielectric ϵ_{ext}.

shown in Table B.2, and again, show very strong agreement, providing more evidence towards reliable and accurate calculations from DelPhi.

Method	U^{solv} $[kT]$	U^{coul} $[kT]$	U^{pol} $[kT]$
Analytical	-55.9114	107.7150	-88.6658
DelPhi	-55.9236	107.5782	-88.5047

Table B.2: Contributions to the electrostatic free energy of the four ion model system shown in Figure B.2. U^{solv} is calculated from the summation of individual ion contributions (Eq. B.1). The U^{pol} value from DelPhi is found by subtracting U^{solv} from the CRFE output (Eq. B.5).

Predicting Energies from Potential Mapping

Given a system of charges, DelPhi allows the calculation of coulombic and reaction potentials at a specified site, from which, coulombic and polarization energies can be predicted.

Four Ion Model

For this analysis, the four ion model with charges in x-y plane was utilized, and potentials calculated at a site on the z-axis 10Å from the origin. The system is displayed in Figure B.3.

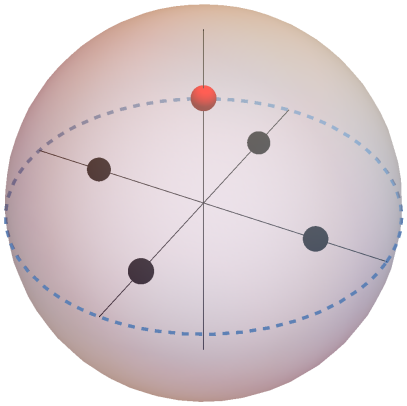

Figure B.3: Four ions of same charge (black spheres) placed in the x-y plane, each equidistant from the origin. The red sphere on the positive z-axis is the position of interest for calculation of site potentials.

DelPhi calculations returned a coulombic potential of $\phi^{coul} = 79.4784\ kT/e$, and reaction potential $\phi^{react} = -63.8188\ kT/e$ at the test position due to the charges in the x-y plane. From the previously tested four ion model electrostatic free energy contributions (Table B.2), and assuming a charge of $-e$ was inserted into the test position, the pairwise coulombic and polarization energies of a five ion system with

the geometry described were predicted as:

$$\begin{aligned} U^{coul}_{5\ ions} &= U^{coul}_{4\ ions} + q_e\phi^{coul} \\ &= 107.7150\ kT + (-e)(79.4784\ kT/e) \\ &= 28.0998\ kT \end{aligned}$$

$$\begin{aligned} U^{pol}_{5\ ions} &= U^{pol}_{4\ ions} + q_e\phi^{react} \\ &= -88.5047\ kT + (-e)(-63.8188\ kT/e) \\ &= -24.6859\ kT \end{aligned}$$

The five ion model show in Figure B.3 was constructed by assigning the test site (red sphere) a charge of $-e$, and keeping all other physical parameters constant from the four ion model discussed above. Calculations from DelPhi and analytical expressions for this model, and their comparison to predicted coulombic and polarization energies from potential mapping are show in Table B.3. Prediction of solvation self energies is not possible from potentials. As seen, the coulombic energy is an exact match to Delphi, and the polarization energy is in strong agreement with both the analytical and DelPhi results.

Method	U^{solv} $[kT]$	U^{coul} $[kT]$	U^{pol} $[kT]$
Analytical	-69.8893	28.1356	-24.7507
DelPhi	-69.9045	28.0998	-24.6966
Predicted from potentials	–	28.0998	-24.6859

Table B.3: Electrostatic free energies for a five ion model predicted from site potentials calculated from the four ion model system shown in Figure B.3, compared to direct calculations from DelPhi and analytical expressions.

Real Proteins

Departing from model systems in which exact solutions exist, a real protein system (the Cold Shock Protein (CSP) from Thermus Thermophilus) was analyzed. Instead of utilizing an all-atom partial charge library, the ionic side chains were modeled with a representative ion of charge $\pm e$, and all remaining atoms were set to neutral. The choice of single ion over all-atom representation will be discussed below, but is worth mentioning here that only small differences in calculated solvation and polarization energies were seen between usage of the two charge libraries.

For first inspection in the calculation of site potentials, four ionic residues (Arg20, Lys36, Glu43, and Asp50) were selected from CSP. Iteratively, each ion was neutralized and assigned the site for potential calculations, while the remaining three ions remained charged. As an example, Figure B.4 displays Lys36 (red sphere) neutralized, and the remaining three ions charged (black spheres) for two different view angles. The potentials at each of the four sites due to the remaining three charges were found,

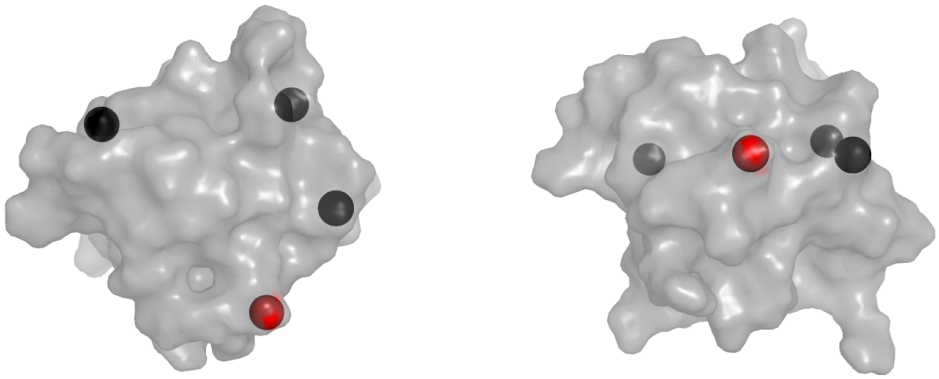

Figure B.4: Dual angle view of the CSP four ion model. Black spheres have their respective charges assigned. The red sphere ion is neutral, but will serve as the site for potential calculations.

and the nomenclature was adopted of ϕ_i^{coul} and ϕ_i^{react} representing the coulombic and reaction potentials, respectively, at site i due to the remaining three ions. The total coulombic and polarization energies of this "real" system were calculated with q_i as

the charge at site i, e.g., $q_{\text{Lys36}} = +e$ and $q_{\text{Glu43}} = -e$:

$$U^{pol} = \frac{1}{2}\sum_i q_i \ \phi_i^{react} = 46.3839 \ kT$$
$$U^{coul} = \frac{1}{2}\sum_i q_i \ \phi_i^{coul} = -48.2466 \ kT$$

With the charge of all four ions "turned on", and the individual ion solvation energies calculated, the total polarization energy was found from DelPhi and Equation B.5 as 46.3838 kT. The coulombic energy contribution from this DelPhi calculation was an exact match to U^{coul}.

Comparison with Generalized Born Model

Another method for calculation of electrostatic free energies invokes the Generalized Born model to approximate the Poisson-Boltzmann equation. Here, we depart from the spherical treatment given in expressions B.1-B.3, and remodel the expressions to capture the pairwise polarization energy from more realistic protein geometries.

$$U_{GB}^{pol} \approx -\left(\frac{1}{\epsilon_{int}} - \frac{1}{\epsilon_{ext}}\right)\sum_{i<j}\frac{q_i q_j}{\sqrt{d_{ij}^2 + r_i^{GB} r_j^{GB} \ \exp\left(-\gamma \frac{d_{ij}^2}{r_i r_j}\right)}} \tag{B.6}$$

where d_{ij} is separation distance between charges q_i and q_j, $\gamma = 1/4$, and r_i^{GB} represents the effective Born radius calculated from the solvation (self) energy of ion i:

$$r_i^{GB} = -\frac{1}{2}\left(\frac{1}{\epsilon_{int}} - \frac{1}{\epsilon_{ext}}\right)\frac{q_i^2}{U_{solv,i}} \tag{B.7}$$

The effective Born radius, r_i^{GB}, captures the degree of burial of ion i, and attempts to quantify the amount of electrostatic screening. The smaller a Born radius for an ion, the larger the screening as if the charge were in a medium with a higher dielectric constant. The calculation of Born radius required the solvation energy for each ion within the system of charges, and was calculated with DelPhi.

Real Proteins

The CSP system described in the Potential Mapping section will be used here. Specifically, the CSP system contains 23 ionic groups, allowing the calculation of $(23 \times 22)/2 = 253$ possible pairwise polarization energies that were calculated with DelPhi (Eq. B.5), and then separately with expressions B.6-B.7. Plotting polarization energy found from U_{GB}^{pol} (Eq. B.6) versus U^{pol} found from DelPhi (Eq. B.5) for each ion pair displays an almost $y = x$ relationship (Figure B.5, left). Furthermore, the ACP protein was analyzed following the same protocol, giving $(30 \times 29)/2 = 435$ pairwise combinations, and the same near $y = x$ relationship is seen when comparing GB polarization energies to those calculated from DelPhi (Figure B.5, right).

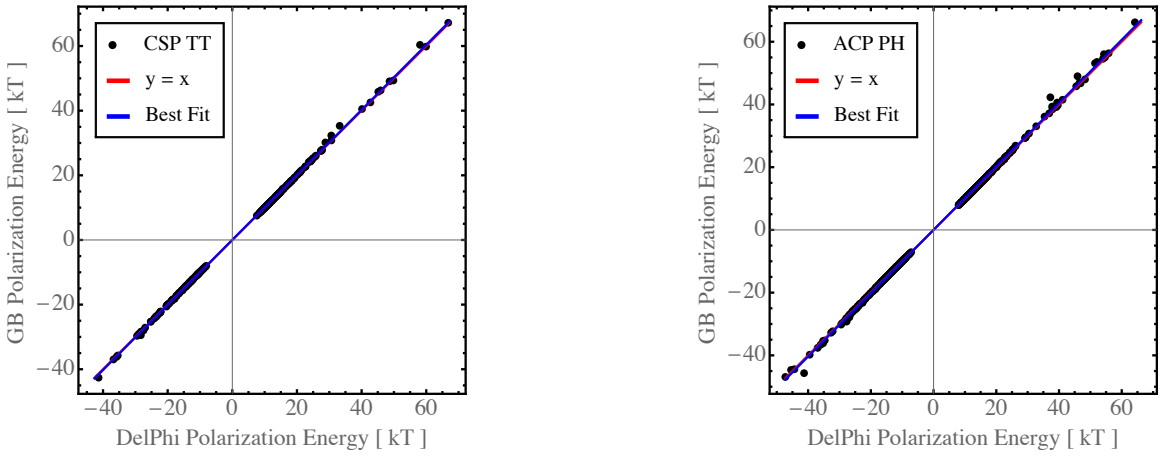

Figure B.5: Direct comparison of GB polarization energy (y-axis) versus DelPhi calculated polarization energy (x-axis) for all possible ionic pairs from CSP (left) and ACP (right).

The summation of all pairwise polarization energies represents the total polarization energy within each system. Compared directly to DelPhi output when all ions are in system, strong agreement is again seen. Specifically, for CSP, $\sum U_{GB}^{pol} = 455.831\ kT$, and from DelPhi and Equation B.5, $U^{pol} = 450.274\ kT$, and for ACP, $\sum U_{GB}^{pol} = 307.969\ kT$, and $U^{pol} = 296.998\ kT$. This is expected given the almost $y = x$ dependence.

Representative Atom Charge Library

For the DelPhi calculations of real protein systems, we chose to construct the charge libraries using a representative atom for charged groups rather than all-atom partial charges. Specifically, instead of every atom carrying some value of charge, we chose a specific atom within each of the seven charged groups (α-amino N, α-carboxyl C, Asp Cγ, Glu Cδ, His (Hip) Cϵ1, Lys Nζ, and Arg Cζ) to carry a charge of $\pm e$, and all remaining atoms were assigned to neutrality. An obvious question arises, how well do these representative charge results compare to those using an all-atom partial charge library? To ensure the results gathered are trustworthy, a brief comparative analysis and discussion is included.

Electrostatic properties from the CSP system were calculated with DelPhi using the all-atom partial charge and ionic group representation libraries, the internal dielectric of the system set to values of 2 and 40.6 (our estimate), and following the working algorithm outlined above to calculate the total solvation (U^{solv}) and polarization (U^{pol}, Eq. B.5) energies. As seen in Table B.4, calculation of total solvation and polarization energies are in good agreement. Examination of the U^{solv} term in either internal dielectric constant case shows the ionic group representation results are within almost 10% of the value found from the all-atom library. Similar results

	Charge Library	U^{solv} $[kT]$	U^{coul} $[kT]$	U^{pol} $[kT]$
$\epsilon_{int} = 2$	All-atom	-1390.9977	-19776.6639	433.778
	Ionic groups	-1227.9558	-514.5706	450.274
$\epsilon_{int} = 40.6$	All-atom	-27.8133	-974.2199	9.0009
	Ionic groups	-25.4347	-25.3483	9.7577

Table B.4: Comparison of electrostatic free energy contributions between all-atom partial charge and ionic group representation atom libraries. Calculations performed at two dielectric constants show the solvation and polarization energies are in good agreement between both libraries.

are seen in comparison of the polarization energies. However, comparison of total coulombic energies shows the value returned from all-atom charges is nearly 40 times that given from ionic group representation. Recall, the DelPhi calculation of coulombic energy is a summation of all pairwise coulombic interactions (Eq. B.2) set in the interior of the proteins's boundary. For the CSP system modeled here, 1144 atoms are assigned a partial charge in the all-atom case, but only 23 charges in the ionic group representation. This large discrepancy in assigned charges, and therefore in the number of i, j pairwise interactions in the summation, could account for the large difference seen in coulombic energies.

In summary, the use of an ionic group representative atom carrying a charge of $\pm e$ versus usage of all-atom partial charge libraries show strong agreement in the total solvation and polarization energies calculated for a given system. The large discrepancy seen in the total coulombic energy stems from the over representation of partial charges in the all-atom case. Since the present research goals are on the effects of ionic sidechain charge–charge interactions, it is recommended to utilize the representative atom library for current DelPhi calculations.