

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Understanding the unfolding mechanism and origins of extreme cooperativity in  $\alpha$ -lytic protease through molecular dynamics unfolding simulations

**Permalink**

<https://escholarship.org/uc/item/5x2557xt>

**Author**

Salimi, Neema

**Publication Date**

2009

Peer reviewed|Thesis/dissertation

Understanding the unfolding mechanism and origins of extreme cooperativity in  
 $\alpha$ -lytic protease through molecular dynamics unfolding simulations

by

Neema L. Salimi

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

*Copyright 2009*

*by*

*Neema L. Salimi*

## Preface

A doctoral dissertation, though it is authored by an individual, is the result of years of collective effort. None of the work contained herein would have been possible without the ideas, efforts, experiments, and support of a multitude of colleagues, friends and family. This dissertation is as much theirs as it is mine.

- My advisor, David Agard, is owed the utmost gratitude. Much of the science contained herein resulted from many productive meetings between the two of us. His breadth and depth of knowledge is, to me, unparalleled, and his support through both scientific and personal hard times unquestionably bettered my work and my life. I cannot hope to sum up his contributions in a paragraph; the best I can do is assert this dissertation would have been impossible without him.
- My colleagues in the Agard Lab truly made the lab a productive environment for science. I have turned to many of them with multiple questions and always found them willing to help. Luke Rice, Nobuyuki Ota, Bosco Ho, Timo Street, Justin Kollman, Albion Baucom, Chris Waddling, Brian Kelch, Cynthia Fuhrmann, Stephanie Truhlar, and Pinar Erciyas have all contributed to my scientific growth.
- Working in the same lab for over six years will also build strong friendships, and while I will miss most everyone in the lab, several people have really made the time go by quicker. Brian Kelch and I sat next to each other for four years before he moved on to Berkeley, and I have missed his shenanigans ever since. Kristin Krukenberg and I sat across from each other ever since she joined the lab, and now she is leaving at the same time as I am. I enjoy talking with her about just about anything. Mariano Tabios is one of the few people in the lab with whom I

can talk about sports, which is a pleasant distraction from what can sometimes be the tedium of science. Finally, Laura Lavery, one of the few who works the late shift, like me, will be sorely missed. We can play the music only the two of us like when we work late, and we talk about life, grad school, sports, and science. She joined the lab several years after I did, and I would like to think I helped in mentoring her during her first few years of grad school.

- UCSF sports a great environment for collaboration, and many folks in several labs helped me throughout my graduate career. Special recognition is deserved for John Chodera and Vince Voelz, as they were both great sounding boards for ideas.
- The two other members of my thesis committee, Ken Dill and Robert Fletterick, have been more than helpful. I have learned much from both of them.
- My classmates in the Biophysics entering class of 2002 were pretty much awesome, and we shared many memories. Sam Pfaff, Caleb Bashor, Chris Farady, Jeremy Wilbur, and Quincey Justman are all fine scientists, fine people, and I am lucky to call them my friends.
- My family was always supportive, even through the many “When are you going to finish?” questions. My father, Mahmoud Salimi, pushed me from a young age to excel, yet always let me choose my own path. My mother, Margaret Salimi, has always been my biggest fan. My sister, Lailah Rice, has helped me keep things in perspective. And numerous family members (I have many) have been there with words of encouragement the last seven years.

# **Understanding the unfolding mechanism and origins of extreme cooperativity in $\alpha$ -lytic protease through molecular dynamics unfolding simulations**

*Neema L. Salimi*

*Laboratory of Dr. David A. Agard*

$\alpha$ -lytic protease ( $\alpha$ LP), a bacterial serine protease of the chymotrypsin family, has evolved both kinetic stability and an extreme unfolding cooperativity in order to limit proteolysis. Trypsin, a well-studied metazoan homolog, is degraded at rates up to 100x faster than  $\alpha$ LP even though it has approximately the same global unfolding rate. Previous experimental studies have implicated the interface between  $\alpha$ LP's two domains as critical to the unfolding pathway and its cooperativity. To investigate this, I performed multiple high temperature molecular dynamics unfolding simulations on both  $\alpha$ LP and trypsin. The simulations revealed a robust unfolding pathway that featured preferential disruption of the domain interface, primarily at three regions: the Domain Bridge, the C-terminal domain  $\beta$ -hairpin and *cis*-proline turn, and the N-terminus. I developed a metric for measuring global unfolding cooperativity, and it showed correctly that  $\alpha$ LP unfolded cooperatively, while trypsin did not. I then applied an information-theory-derived measure of cooperativity developed by Voelz, based on pairs of contacts in the two proteins, to the simulations, allowing me to look at cooperativity at the residue level. By graphing the contact cooperativity as a network, I showed that the  $\alpha$ LP network is significantly larger and more connected than that of trypsin, again showing a much higher

global unfolding cooperativity. Using only the early parts of the simulations, the cooperativity network highlights contacts broken cooperatively around the transition state ensemble. These network graphs also identify residues that are key centers of cooperativity and if mutated, may disrupt  $\alpha$ LP's unfolding cooperativity. Experimental studies are currently underway in the lab to test the hypotheses created from these simulations.

## Table of Contents

<b>Preface .....</b>	<b>iii</b>
<b>Understanding the unfolding mechanism and origins of extreme cooperativity in <math>\alpha</math>-lytic protease through molecular dynamics unfolding simulations .....</b>	<b>v</b>
<b>Table of Contents .....</b>	<b>vii</b>
<b>List of Figures.....</b>	<b>ix</b>
<b>List of Tables .....</b>	<b>xi</b>
<b>Introduction.....</b>	<b>1</b>
Background.....	1
Specific Aims.....	5
<b>Chapter 1: Unfolding Simulations Reveal the Mechanism of Extreme Unfolding Cooperativity in the Kinetically Stable <math>\alpha</math>-Lytic Protease.....</b>	<b>6</b>
Preface .....	6
Synopsis.....	7
Background.....	8
Results.....	12
Discussion.....	33
Materials and Methods .....	40
Acknowledgements .....	44
Postscript.....	46
<b>Chapter 2: Understanding unfolding cooperativity through an information theory measure of contact pair cooperativity in molecular dynamics simulations.....</b>	<b>49</b>
Preface .....	49
Synopsis.....	51
Background.....	52
Results.....	57
Discussion.....	72
Materials and Methods .....	76



Acknowledgments .....	79
Postscript.....	80
<b>Chapter 3: Accurately measuring the distortion of aromatic rings in crystal structures and molecular dynamics trajectories: <math>\alpha</math>LP F228 as a case study.....</b>	<b>81</b>
Preface .....	81
Synopsis.....	82
Background.....	83
Results.....	85
Discussion.....	92
Methods .....	93
Acknowledgments.....	94
<b>Chapter 4: Conclusions and future directions.....</b>	<b>95</b>
How does $\alpha$ LP unfold so cooperatively?.....	95
Future directions – computation.....	97
Future directions – biochemistry.....	98
<b>References.....</b>	<b>100</b>
<b>Appendix 1: Supplemental Material for Chapter 1 .....</b>	<b>109</b>
<b>Appendix 2: Supplemental Material for Chapter 2 .....</b>	<b>112</b>
<b>Appendix 3: A new kinetically stable protease alignment .....</b>	<b>115</b>
Description.....	115
Alignment .....	115

## List of Figures

Figure 1.1: The structure of $\alpha$ LP.....	10
Figure 1.2: $\alpha$ LP unfolds significantly and reproducibly at high temperature but is stable at 298K. ....	14
Figure 1.3: Selected structures from the 500K1 simulation illustrate the $\alpha$ LP unfolding pathway.....	18
Figure 1.4: Conformational clustering effectively defines the exit from the native state.	21
Figure 1.5: Property-based landscapes clearly separate native from non-native conformations.....	25
Figure 1.6: The structure of the $\alpha$ LP TSE. ....	27
Figure 1.7: Contacts at the domain interface are preferentially broken at the unfolding transition. ....	29
Figure 1.8: $\alpha$ LP unfolds significantly more cooperatively than trypsin.....	32
Figure 1.9: Solvation of the domain interface during unfolding differs significantly between $\alpha$ LP and trypsin. ....	37
Figure 2.1: The structures of $\alpha$ LP and trypsin.....	56
Figure 2.2: MCOOP explained .....	58
Figure 2.3: The distribution of MCOOP values in $\alpha$ LP and trypsin.....	60
Figure 2.4: $\alpha$ LP and trypsin cooperativity networks.....	62
Figure 2.5: Selected clusters from the cooperativity networks.....	65
Figure 2.6: The $\alpha$ LP <sub>early</sub> cooperativity network.....	69
Figure 2.7: Highly cooperative residue distributions differ between trypsin, $\alpha$ LP, and $\alpha$ LP <sub>early</sub> .....	71

Figure 3.1: The distribution of Phe and Tyr bend angles in ultra-high resolution structures.....	86
Figure 3.2: The distortion of Phe rings in $\alpha$ LP at room temperature.....	89
Figure 3.3: F228 remains distorted until after $\alpha$ LP has unfolded.....	92
Figure A1.1: Representative conformations of the $\alpha$ LP TSE from each simulation show both the similarity and diversity of the TSE.....	109
Figure A1.2: C $\alpha$ RMSD for trypsin control and unfolding simulations.....	110
Figure A1.3: X-ray structure of trypsin and members of its unfolding TSE from each simulation.....	110
Figure A2.1: The large cooperativity difference between $\alpha$ LP and trypsin is not solely due to simulation number.....	112
Figure A2.2: Additional clusters identified in Figure 2.4.....	113
Figure A2.3: MCOOP distributions for abbreviated simulations.....	114

## List of Tables

Table 1.1: Time (ns) at the native cluster exit for the five $\alpha$ LP unfolding simulations. ...	21
Table 2.1: Parameters of the three MCOOP distributions and networks. ....	60
Table 2.2: Residues making up each cluster. ....	63
Table 3.1: Statistics for the distribution of bend angles .....	87
Table 3.2: Statistics for Phe ring distortions in the 298K $\alpha$ LP simulation. ....	90
Table A1.1: Parameter loadings for the $\alpha$ LP Principal Components Analysis landscape. .....	109
Table A1.2: Selected properties of the $\alpha$ LP crystal structure and TSE. ....	109
Table A1.3: Properties of the trypsin TSE.....	111

## **Introduction**

### ***Background***

$\alpha$ -lytic protease ( $\alpha$ LP), a bacterial serine protease of the chymotrypsin family, has long been studied for its unusual energy landscape. Its folded and enzymatically active state is less energetically stable than both a molten-globule-like intermediate and the fully unfolded state (Sohl, Jaswal et al. 1998). The native state is maintained by its kinetic stability;  $\alpha$ LP unfolds on the timescale of one year (Sohl, Jaswal et al. 1998). The native state's metastability is a consequence of an even slower unfolding rate on the order of millennia (Sohl, Jaswal et al. 1998). To get to the native state, an N-terminal pro region acts as a folding catalyst and is degraded once the active enzyme is formed. Why would  $\alpha$ LP evolve such an odd energy landscape? An assay comparing the degradation rates of  $\alpha$ LP and its metazoan homologs trypsin and chymotrypsin produced a remarkable result;  $\alpha$ LP is degraded up to a hundred-fold more slowly than trypsin and chymotrypsin (Jaswal, Sohl et al. 2002). Even more surprising, trypsin unfolds at the same rate as  $\alpha$ LP, and yet is degraded much faster (Truhlar, Cunningham et al. 2004). To avoid proteolysis,  $\alpha$ LP must suppress partial unfolding to a much higher degree than its metazoan homologs, which it does, as its autolysis rate is only slightly faster than its unfolding (Jaswal, Sohl et al. 2002). It does this by having a very rigid native structure; hydrogen exchange protection factors for 31 core backbone amides exceed  $10^9$  (Jaswal, Sohl et al. 2002).  $\alpha$ LP's suppression of partial unfolding, its extreme unfolding cooperativity, provide it a key functional benefit in harsh environments.

Kinetic stability and extreme unfolding cooperativity are conserved amongst other bacterial proteases homologous to  $\alpha$ LP. SGPB has a slightly reduced kinetic stability relative to  $\alpha$ LP, but folds much faster, in between  $\alpha$ LP and trypsin (Truhlar, Cunningham et al. 2004). A thermophile, TFPA, has a much increased kinetic stability relative to  $\alpha$ LP, especially at high temperatures. Another homolog, NAPase, secreted from a bacterium found in bathroom tile grout, is also kinetically stable with a particular resistance to unfolding at acidic pH. Studies in the Agard Lab on  $\alpha$ LP and these homologs have led to greater insight into the mechanistic underpinnings of kinetic stability and extreme unfolding cooperativity.

Interestingly, many of the previous studies have implicated the  $\alpha$ LP domain interface as critical to the unfolding pathway. Both the temperature and denaturant dependence of unfolding led to a “cracked egg” model of unfolding, where  $\alpha$ LP’s two domains separate but remain relatively intact (Jaswal 2000). On the folding side, the two individual  $\alpha$ LP domains do not fold independently, nor can they reconstitute the active enzyme, unlike both trypsin and chymotrypsin (Duda and Light 1982; Higaki and Light 1986; Cunningham and Agard 2003). A study on the Domain Bridge, the covalent linkage between the two domains in bacterial proteases, showed that the amount of surface area it buries inversely correlates with the unfolding rate of the protein, i.e., the larger the Domain Bridge, the slower the unfolding (Kelch and Agard 2007). Finally, elimination of an inter-domain salt bridge, found in  $\alpha$ LP and not in NAPase, drastically reduced  $\alpha$ LP’s susceptibility to acidic pH unfolding, like that of NAPase (Kelch, Eagen et al. 2007).

However, we still lack a comprehensive picture of  $\alpha$ LP unfolding. Most proteins whose energy landscapes are extensively studied are subjected to  $\phi$ -value analysis, a procedure where many mutants, often hydrophobic deletions (e.g. Leu to Ala), are generated and have their folding/unfolding kinetics measured (Matouschek, Kellis et al. 1989; Fersht, Matouschek et al. 1992). From the ratio of the change in the Transition State Ensemble's (TSE) stability to the change in overall stability, structure in the TSE can be inferred. Many of these mutants would make measuring  $\alpha$ LP folding kinetics impossible, due to significantly reduced expression and/or folding kinetics below the detection limit of our extremely sensitive enzyme assay.

Several labs, primarily the Daggett Lab, have applied high temperature molecular dynamics simulations to the problem of describing unfolding pathways and the structures of TSEs. Many model systems for protein folding have been characterized in this manner, and the computational results generally compare favorably to experimentally determined  $\phi$ -values (Li and Daggett 1994; Li and Daggett 1996; Fulton, Main et al. 1999; Day and Daggett 2005; Scott, Randles et al. 2006). The simulations can also be used in a predictive manner, as the Daggett Lab showed with a faster folding variant of CI2 designed from the simulated TSE structure (Ladurner, Itzhaki et al. 1998).

Here, I apply unfolding simulations to the problem of  $\alpha$ LP's kinetic stability and extreme cooperativity. Like others, I hope to provide both explanatory and predictive power to biochemical experiments from the atomic resolution accessible in these simulations. As a significant challenge, unfolding cooperativity has not been studied previously by molecular dynamics, and it was unclear when starting the project if it could be measured computationally. This dissertation contains my attempt to characterize the

unfolding pathway of  $\alpha$ LP, applying novel methods to ensure robustness, and understand the mechanism of its unfolding cooperativity, with new techniques aimed at discovering the critical residues involved.



### *Specific Aims*

The goal of this dissertation is to understand the unfolding pathway and cooperativity of  $\alpha$ -lytic protease at high resolution by addressing the following questions:

- How does the simulated  $\alpha$ LP unfolding pathway compare with experimental studies?
- What is the role of the domain interface in  $\alpha$ LP unfolding?
- Can unfolding cooperativity be measured in a simulation and is the experimental difference in cooperativity between  $\alpha$ LP and trypsin maintained?
- What insights can be gained from investigating cooperativity at the level of contact pairs?

# **Chapter 1: Unfolding Simulations Reveal the Mechanism of Extreme Unfolding Cooperativity in the Kinetically Stable $\alpha$ -Lytic Protease**

## *Preface*

Despite many recent biochemical experiments on the  $\alpha$ LP energy landscape, we still lacked a comprehensive description of its unfolding pathway. There was also little known about how  $\alpha$ LP unfolded so cooperatively while metazoan serine proteases did not. I undertook molecular dynamics unfolding simulations of both  $\alpha$ LP and trypsin in order to answer these questions.

As of this writing, this work in this chapter had been submitted to *PLoS Computational Biology* for publication and was under review. David Agard appears as the second author, having contributed intellectually to the science and assisted in the writing of the paper. I performed all of the simulations and analysis and wrote the paper, appearing as the first author.

## *Synopsis*

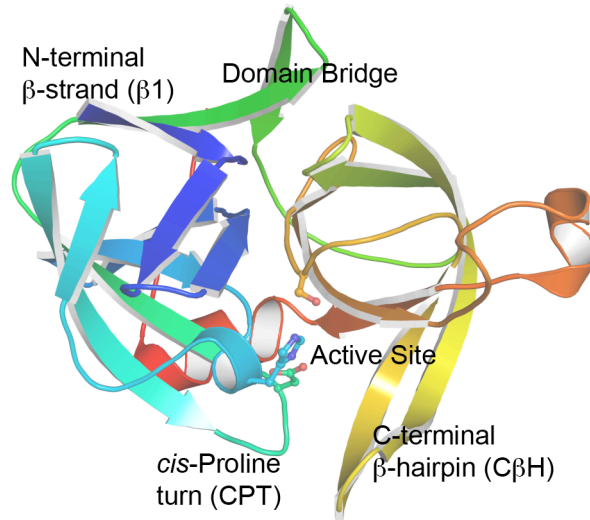
Kinetically stable proteins, those whose stability is derived from their slow unfolding kinetics and not thermodynamics, are examples of evolution's best attempts at suppressing unfolding. Especially in highly proteolytic environments, both partially and fully unfolded proteins face potential inactivation through degradation and/or aggregation, hence, slowing unfolding can greatly extend a protein's functional lifetime. The prokaryotic serine protease  $\alpha$ -lytic protease ( $\alpha$ LP) has done just that, as its unfolding is both very slow ( $t_{1/2} \approx 1$  year) and so cooperative that partial unfolding is negligible, providing a functional advantage over its thermodynamically stable homologs, such as trypsin. Previous studies have identified regions of the domain interface as critical to  $\alpha$ LP unfolding, though a complete description of the unfolding pathway is missing. In order to identify the  $\alpha$ LP unfolding pathway and the mechanism for its extreme cooperativity, we performed high temperature molecular dynamics unfolding simulations of both  $\alpha$ LP and trypsin. The simulated  $\alpha$ LP unfolding pathway produces a robust transition state ensemble consistent with prior biochemical experiments and clearly shows that unfolding proceeds through a preferential disruption of the domain interface. Through a novel method of calculating unfolding cooperativity, we show that  $\alpha$ LP unfolds extremely cooperatively while trypsin unfolds gradually. Finally, by examining the behavior of both domain interfaces, we propose a model for the differential unfolding cooperativity of  $\alpha$ LP and trypsin involving three key regions that differ between the kinetically stable and thermodynamically stable classes of serine proteases.

## ***Background***

$\alpha$ -lytic protease ( $\alpha$ LP), a prokaryotic serine protease of the chymotrypsin family, has evolved an unusual energetic landscape, providing it a functional advantage over its metazoan homologs. Unlike most proteins,  $\alpha$ LP's active state is not stabilized by thermodynamics, but by a large kinetic barrier to unfolding, with an unfolding  $t_{1/2}$  of  $\sim 1$  year (Sohl, Jaswal et al. 1998). While thermodynamically stable homologs like trypsin have similar unfolding rates, they are degraded at rates up to 100x faster than  $\alpha$ LP under highly proteolytic conditions (Jaswal, Sohl et al. 2002; Truhlar, Cunningham et al. 2004). In addition, the rates of  $\alpha$ LP unfolding and degradation are nearly identical, indicating that partial unfolding leading to proteolysis is negligible. Therefore,  $\alpha$ LP's functional advantage is derived from not only its very slow unfolding, which it shares with trypsin, but also its suppression of local unfolding events that would render it protease-accessible. Thus, it appears that the evolution of  $\alpha$ LP has generated such extreme cooperativity in unfolding in order to maximize its functional lifetime under harsh conditions. The cost of maximizing resistance to unfolding comes in the form of extremely slow folding ( $t_{1/2} \sim 1800$  years) and the consequent loss of thermodynamic stability of the active state relative to the unfolded state (Sohl, Jaswal et al. 1998; Truhlar, Cunningham et al. 2004). However,  $\alpha$ LP also evolved a large Pro-region folding catalyst, which speeds folding by nine orders of magnitude and is then degraded by the mature protease, decoupling the folding and unfolding landscapes so that unfolding resistance can be maximized (Sohl, Jaswal et al. 1998; Jaswal, Sohl et al. 2002; Cunningham and Agard 2004).

Given  $\alpha$ LP's unusual energetic landscape and its reliance on kinetic stability, much effort has focused on elucidating its unfolding mechanism in detail. Native-state

hydrogen-deuterium exchange showed over half of its 194 backbone amides are well-protected from exchange, and 31 have protection factors greater than  $10^9$  (Jaswal, Sohl et al. 2002). This extreme rigidity is spread throughout both domains and is indicative of  $\alpha$ LP's high unfolding cooperativity. Thermodynamic decomposition of the unfolding energetics into entropic and enthalpic contributions suggested a prominent role for the extensive domain interface in unfolding, with the critical step involving solvation of the domain interface while the individual domains remain relatively intact (Jaswal, Truhlar et al. 2005). Mutational studies on  $\alpha$ LP inspired by the acid-resistant homolog NAPase were consistent with this hypothesis. The distribution of salt-bridges in NAPase and  $\alpha$ LP differ markedly; replacement of a salt-bridge at  $\alpha$ LP's domain interface with an intra-domain salt-bridge (as in NAPase) resulted in significant increases in  $\alpha$ LP's resistance to low pH unfolding (Kelch, Eagen et al. 2007). A major component of the domain interface, the Domain Bridge (Figure 1), is the only covalent linkage between the two domains. This structure exists only in prokaryotic proteases and varies considerably among  $\alpha$ LP and its homologs. The area buried by the domain bridge is inversely correlated with the high-temperature unfolding rate for four kinetically stable proteases, indicating both its relevance and that it is weakened early in unfolding (Kelch and Agard 2007). Another domain interface component is a  $\beta$ -hairpin in the C-terminal domain ( $C\beta H$ ), unique to kinetically stable proteases, that forms part of the active site (Figure 1). Substitution of a more stable  $\beta$ -turn was consistent with an unfolding pathway where  $C\beta H$  loses its domain interface contacts early in unfolding (Truhlar and Agard 2005). Despite much progress, we still lack a global picture of  $\alpha$ LP unfolding, especially at high resolution.



**Figure 1.1: The structure of  $\alpha$ LP.**

The molecule is colored dark blue at the N-terminus progressing to red at the C-terminus. Important structural regions for this work are labeled, including the active site (the catalytic triad of H36, D63, and S143 are represented in ball-and-stick), the N-terminal  $\beta$ -strand ( $\beta$ 1, blue), the *cis*-proline turn (CPT, teal), the Domain Bridge (green), and the C-terminal  $\beta$ -hairpin ( $C\beta$ H, yellow).

For higher-resolution views of protein folding/unfolding, researchers have often turned to  $\phi$ -value analysis (Matouschek, Kellis et al. 1989; Fersht, Matouschek et al. 1992; Itzhaki, Otzen et al. 1995; Fersht 2000). These studies involve large-scale protein engineering experiments which investigate the molecule's folding and unfolding kinetics after making perturbing mutations, normally hydrophobic deletions. By analyzing sufficiently large numbers of perturbations, structure in the transition state ensemble (TSE) can be inferred and a folding/unfolding mechanism can be proposed. Unfortunately, the extremely slow folding and unfolding rates for  $\alpha$ LP make large-scale  $\phi$ -value analysis on  $\alpha$ LP impractical. As an alternative, we decided to investigate the  $\alpha$ LP unfolding pathway computationally in order to explain previous experiments and guide new ones.

High-temperature molecular dynamics (MD) unfolding simulations offer the highest structural and temporal resolution for studying protein unfolding, but their results must be validated experimentally. Daggett and co-workers have been pioneers in this field, using Chymotrypsin Inhibitor 2 (CI2) as a model system to show how well simulated unfolding calculations agreed with experimental  $\phi$ -values and were even able to predict faster folding mutants (Li and Daggett 1994; Li and Daggett 1996; Ladurner, Itzhaki et al. 1998; Day and Daggett 2005). Further work on other proteins by multiple groups has established MD unfolding simulations as a useful tool in examining protein unfolding at atomic resolution while correlating well with experiments (Lazaridis and Karplus 1998; Fulton, Main et al. 1999; Scott, Randles et al. 2006; Oroguchi, Ikeguchi et al. 2007).

A critical step in analyzing unfolding simulations is accurately pinpointing the TSE from the multitude of conformations generated. Because the TSE is experimentally accessible through a molecule's folding and unfolding kinetics, its identification computationally can be used for both explanatory and predictive purposes. Various methods for identifying the TSE have been used in the past, breaking down into conformational clustering and landscape methods (Li and Daggett 1994; Lazaridis and Karplus 1998; Kazmirski, Li et al. 1999; Day and Daggett 2005; Scott, Randles et al. 2006). Conformational clustering relies on all-versus-all comparisons of conformations, often by  $C\alpha$  RMSD, while landscapes separating native from unfolded structures can be generated using properties of the conformations, such as the fraction of native contacts or secondary structure.

Here, we report the results of multiple MD simulations carried out at high temperature in order to probe the mechanism of  $\alpha$ LP's extremely cooperative unfolding. Due to the robustness and cooperativity of  $\alpha$ LP unfolding, the same TSE is obtained using either conformational clustering or landscape methods. The simulated unfolding pathway for  $\alpha$ LP matches well with previously described experiments and provides atomic resolution to previous models for  $\alpha$ LP unfolding which highlight the role of the domain interface. In addition, we have performed similar simulations on trypsin with the goal of understanding the observed experimental differences in unfolding cooperativity. Through a novel method for calculating cooperativity in MD simulations, we show  $\alpha$ LP unfolds significantly more cooperatively than trypsin, mirroring the experimental results. Finally, by analyzing the domain interfaces of both proteins during unfolding, we propose a mechanism for how this differential cooperativity is achieved.

## ***Results***

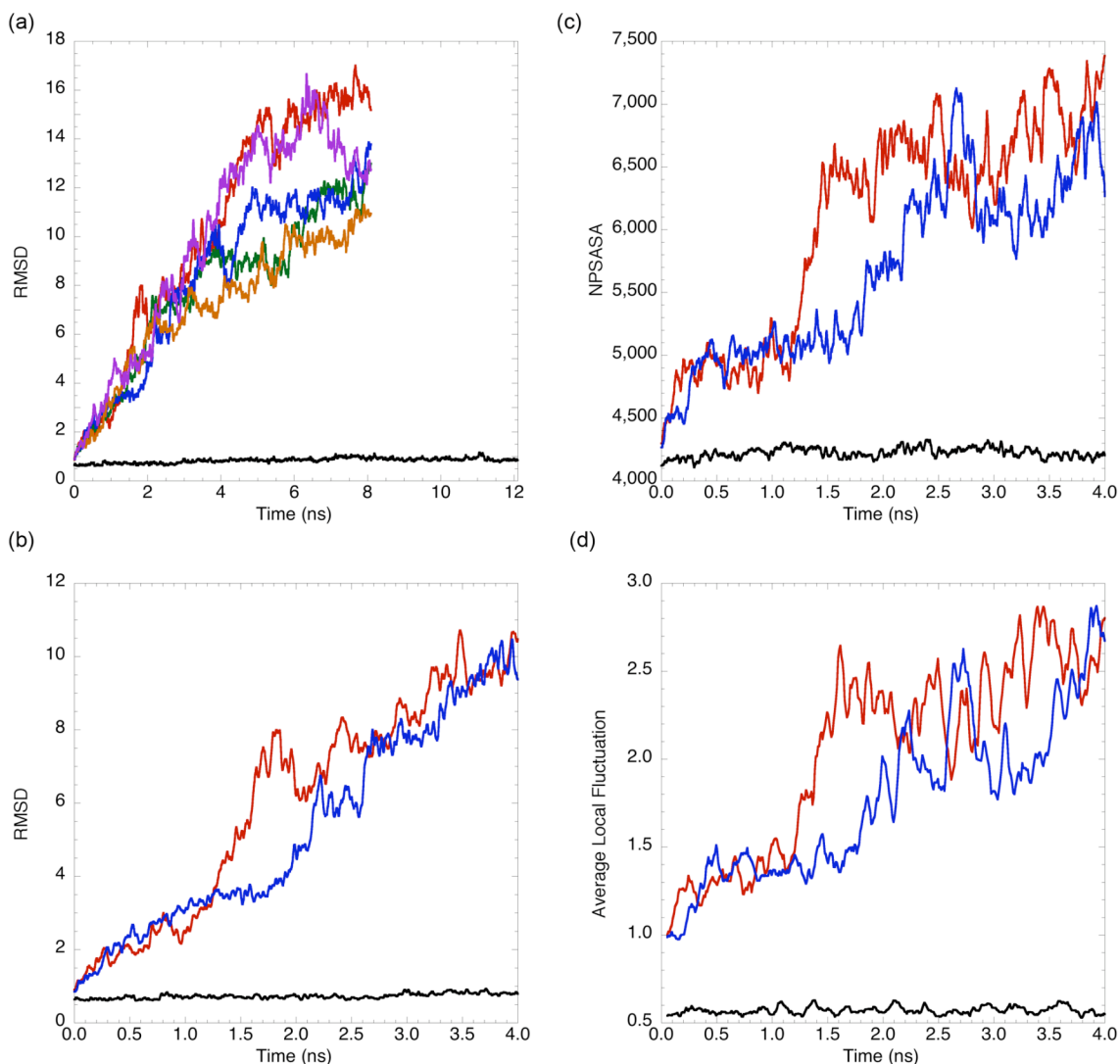
### *Unfolding Simulations*

Simulations were performed with NAMD (Phillips, Braun et al. 2005) using the CHARMM22 (MacKerell, Bashford et al. 1998) forcefield and TIP3P explicit water (full details in Methods). To test for proper behavior in our simulations, a 298K MD simulation of  $\alpha$ LP was performed for 12.1 ns.  $\alpha$ LP was quite stable, averaging 0.84 Å C $\alpha$  RMSD to the crystal structure (Fuhrmann, Kelch et al. 2004) over the course of the simulation and 0.87 Å C $\alpha$  RMSD over the last 1 ns, with a maximum of 1.32 Å (Figure 2A). A previous 1 ns MD simulation of  $\alpha$ LP at 300K using a different force field and simulation conditions also found little deviation from the crystal structure (average 0.83



Å C $\alpha$  RMSD) (Ota and Agard 2001). A long loop comprising residues 163-178 (Figure 1, middle right, orange) and several residues at turns contribute most of the differences and have higher than average B-factors in the crystal structure (Fuhrmann, Kelch et al. 2004). At 298K, there is little additional exposure of non-polar solvent accessible surface area (NPSASA), with an average increase of 5.5 % in exposure (Figure 2C). It should be noted that the rigidity of  $\alpha$ LP as seen by 298K simulation is considerably greater than what is observed for other proteins (Li and Daggett 1994; Fulton, Main et al. 1999; Scott, Randles et al. 2006), consistent with the very low crystallographic B-factors (Fuhrmann, Kelch et al. 2004) and high hydrogen exchange protection factors (Jaswal, Sohl et al. 2002) seen previously.

Five independent 8.1 ns MD simulations at 500K were conducted to determine the unfolding pathway of  $\alpha$ LP, with the C $\alpha$  RMSD of each plotted in Figure 2A. Visual inspection of the trajectories and the high C $\alpha$  RMSDs attained indicated that  $\alpha$ LP had unfolded in each simulation. By contrast, simulations at 450K showed little unfolding at similar timescales making them impractical for analysis (data not shown). Each trajectory shows a generally increasing C $\alpha$  RMSD throughout the simulation, though there is significant variation in the rates of increase, periods of no change or decrease in C $\alpha$  RMSD, and final C $\alpha$  RMSD, as expected for independent simulations. Because relatively high RMSDs were reached in the first 4 ns of the simulations, we hypothesized that the major unfolding transition occurred in that timeframe (Figure 2B).



**Figure 1.2:  $\alpha$ LP unfolds significantly and reproducibly at high temperature but is stable at 298K.**

(a) At 500K,  $\alpha$ LP unfolds quickly and fully in the five 8.1 ns unfolding simulations while it remains native-like at 298K as measured by C $\alpha$  RMSD (black, 298K; red, 500K1; green, 500K2; blue, 500K3; orange, 500K4; purple, 500K5). (b,c,d) Colors used are the same as in (a). 500K1 and 500K3 were chosen due to the relatively large difference in their unfolding times. (b) C $\alpha$  RMSD for the first 4 ns of 298K, 500K1, and 500K3 indicates unfolding occurs early at high temperature. (c) The NPSASA for the first 4 ns of 500K1, 500K3, and 298K is shown. After a short thermal equilibration, both 500K1 and 500K3 reach values  $\sim 5000 \text{ \AA}^2$  and level off until exposing much more non-polar surface at 1.3 and 1.8 ns, respectively. At 298K, very little increase is seen in NPSASA. (d) ALF measures short-term fluctuations in structure and is an indicator of conformational flexibility of the molecule's current state. For both 500K1 and 500K3, conformational flexibility is low and then suddenly rises concurrently with NPSASA. For all but (d), the data is smoothed with a 0.019 ns running average.

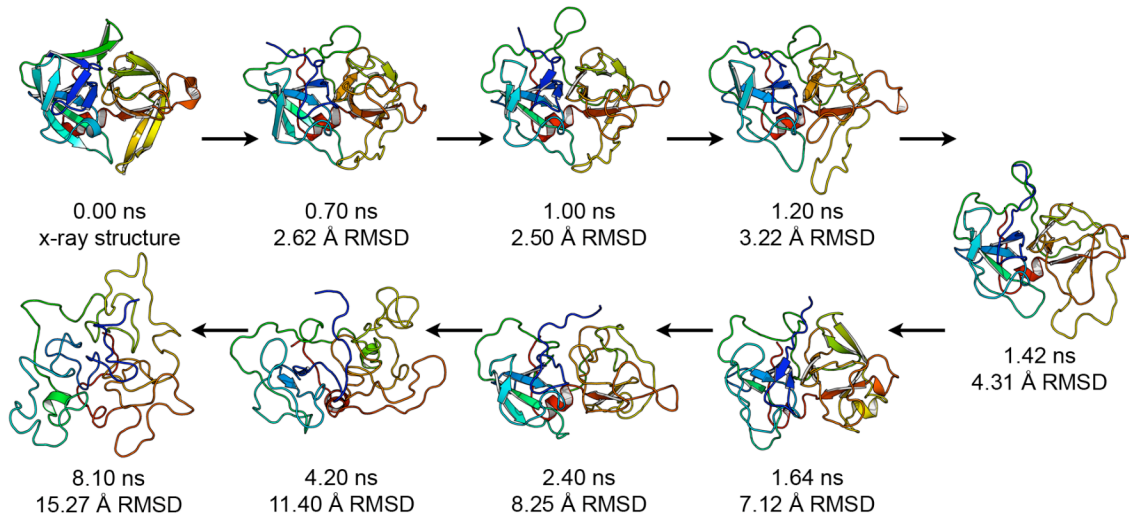
To confirm that unfolding had occurred, we examined molecular properties orthogonal to C $\alpha$  RMSD early in the simulations. These properties, non-polar solvent accessible surface area (NPSASA) and a new metric termed Average Local Fluctuation (ALF), can distinguish native from non-native conformations without directly comparing them to the crystal structure. First, non-polar amino acid side-chains, normally buried in a protein's interior, become exposed upon unfolding, increasing NPSASA. The NPSASA for the first 4 ns of 298K1 (for comparison), 500K1, and 500K3 is plotted in Figure 1C. 500K1 and 500K3 were chosen for clarity due to a large difference in unfolding time. Both exhibit relatively small increases to  $\sim 5000 \text{ \AA}^2$  within the first 0.3 ns, consistent with thermal equilibration. NPSASA then increases very slowly, unlike C $\alpha$  RMSD, until it rapidly increases at 1.3 and 1.8 ns for 500K1 and 500K3, respectively. These sharp rises are followed by another slowly increasing phase that is highly variable for the rest of the simulations.

The second property, ALF, relies on the notion, derived from funnel energy landscape models of protein folding/unfolding, that molecules in the unfolded ensemble can explore many more conformations than those in the native ensemble (Dill and Chan 1997). For  $\alpha$ LP, where the unfolding barrier has been shown experimentally to be extremely high, cooperative, and entropic in nature, it is certain that conformational space on the folded side of the TSE is quite restricted relative to the unfolded side (Jaswal, Sohl et al. 2002; Jaswal, Truhlar et al. 2005). If unfolding simulations capture this ensemble behavior, there would be bottlenecks or barriers in the unfolding landscape. ALF was created to assay for these barriers, as it measures the rate of conformational change throughout a simulation (details in Methods). ALF for the first 4 ns of 298K1 (for

comparison), 500K1, and 500K3 is plotted in Figure 2D. In the first 0.3 ns of both simulations, ALF increases slightly from 1.0 to 1.3 Å due to thermal equilibration. It remains relatively flat until rapid increases beginning at 1.3 and 1.8 ns for 500K1 and 500K3, respectively, resulting in a permanently higher ALF. In 500K3, ALF increases less sharply relative to 500K1, rapidly decreasing and then recovering in the middle of its rise  $\sim 2.0$  ns, which has implications for identifying its TSE (see below). The large and permanent increases in conformational flexibility measured by ALF and their coincidence with similar increases in NPSASA are indicative of seeing true unfolding transitions.

Structurally, the early stages of  $\alpha$ LP's unfolding pathway are quite consistent among the five unfolding simulations, though the simulations tend to diverge once the molecule becomes much less native-like. As we will show below, these early events constitute the major unfolding transition and are the primary focus of this work. First, we will describe the pathway in detail for 500K1, with several important conformations shown in Figure 3, and then note any important differences in other simulations. A movie of the full 500K1 unfolding pathway is provided in the Supporting Information (Video S1). For the first several hundred picoseconds,  $\alpha$ LP thermally equilibrates and reaches  $\sim 2$  Å C $\alpha$  RMSD to the crystal structure, with small surface loops the major source of this small deviation. At 0.7 ns, a large loop comprising residues 163-178 unique to  $\alpha$ LP becomes more mobile, though its flexibility is somewhat limited by a disulfide bond between residues C137 and C170. Because this loop is not conserved in kinetically stable proteases and is relatively mobile at 298K, we feel its overall impact on the unfolding pathway is small. At 1.0 ns, the Domain Bridge, a  $\beta$ -hairpin connecting the two domains of  $\alpha$ LP, becomes more mobile but remains intact (Figures 1 and 3). Between 1.2 and 1.4

ns,  $\alpha$ LP begins to unfold much more significantly, though the distortions are confined to four main structural areas: the N-terminal strand  $\beta$ 1, the Domain Bridge, a region near the active site comprising the C $\beta$ H and a *cis*-proline-containing turn (residues 58-63, CPT), and the 163-178 loop (Figures 1 and 3).  $\beta$ 1 pulls away from the body of the protein and becomes highly flexible. The Domain Bridge breaks tertiary contacts with nearby residues and its two strands separate. Contacts between the CPT and the C $\beta$ H break as the two pull away from each other, and the C $\beta$ H strands separate. The 163-178 loop remains highly flexible, causing residues 160-162, which form part of the substrate binding groove, to separate from the  $\beta$ -barrel and push the C $\beta$ H away from the body of the protein. These regions continue to unfold, accelerating the unfolding of nearby structure, though several regions remain relatively well-structured at 1.64 ns, including the  $\beta$ -sheets  $\beta$ 4- $\beta$ 7- $\beta$ 6 and  $\beta$ 14- $\beta$ 15- $\beta$ 16, and the C-terminal  $\alpha$ -helix (Figure 3). The C-terminal  $\beta$ -barrel unfolds and further weakens the domain interface, with very few native-like interactions bridging the two domains at 2.4 ns (Figure 3). By 4.2 ns, little residual structure remains, as the C $\alpha$  RMSD is 11.4 Å, though the molecule does continue to unfold, reaching a C $\alpha$  RMSD over 16 Å within 8 ns (Figure 3). The presence of three disulfide bonds most likely prevents more extreme unfolding.



**Figure 1.3: Selected structures from the 500K1 simulation illustrate the  $\alpha$ LP unfolding pathway.**

Time in the simulation and C $\alpha$  RMSD to the crystal structure are indicated. See the text for a full description.

Early on, each of the unfolding simulations follows a similar trajectory to that of 500K1 although with variability in the timing (Figure 2), beyond this, some other differences do exist. In 500K4,  $\beta$ 5 unfolds much earlier relative to the other simulations, separating from  $\beta$ 2 and  $\beta$ 6 and partially exposing the interior of the N-terminal domain to solvent. The turn connecting  $\beta$ 5 to the more stable  $\beta$ 6 (Figure 1, upper left, light blue) is quite flexible in all five unfolding simulations and has some of the highest B-factors in the crystal structure, which may explain part of this behavior (Fuhrmann, Kelch et al. 2004; Kelch and Agard 2007). In 500K3, the Domain Bridge does break some tertiary contacts with surrounding regions early in unfolding, but its two strands separate relatively late. The N-terminal  $\beta$ 1 does not completely separate from the body of the protein in 500K2 and 500K3 early on, as it does in the other three simulations, but its contacts are somewhat disrupted in both. Other differences at early time points appear to

be relatively minor and are to be expected given five independent high temperature unfolding simulations.

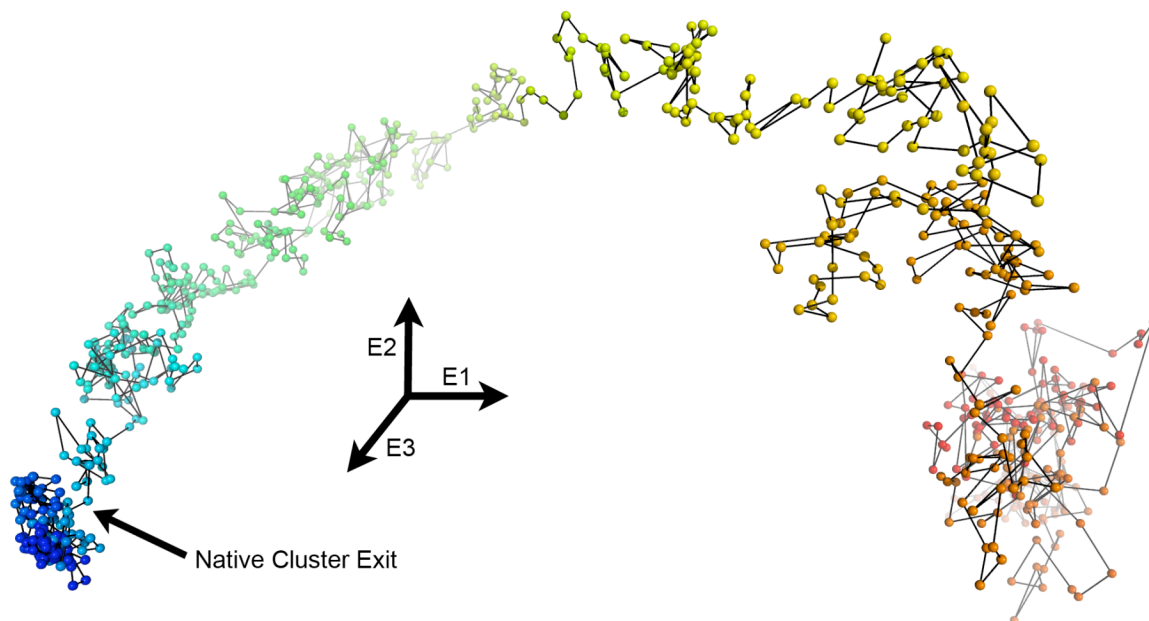
### *Determining the Location of the TSE*

Because computational studies of protein unfolding are severely restricted in the number of molecules that can be simulated, they must use the vast amount of information present in each simulation in order to identify the TSE. As in other types of single-molecule experiments, there will be significant variation within the properties of the ensembles, such as time to unfold. Unlike experimental studies, where there is often a single reporter of the molecule's conformation, such as tryptophan fluorescence, MD simulations provide every conformation sampled, an enormous amount of data. However, there is no *a priori* way to say whether a particular three-dimensional structure is "folded" or "unfolded." The challenge then is to derive properties from the conformations, either those directly computable from each structure or those that rely on comparing structures to each other, that can be used to clearly separate the folded from the unfolded conformations.

Previous studies investigating the nature of a protein's TSE by unfolding simulations have often determined TSEs from individual simulations and combined them into an overall TSE (Fulton, Main et al. 1999; Day and Daggett 2005; Jemth, Day et al. 2005). These approaches depend on the assumption that the TSE is a small region of conformational space at the edge of the native basin, hence identifying them requires methods that clearly separate native from non-native conformations. One method that has had considerable success is a conformational clustering procedure pioneered by Li and

Daggett (Li and Daggett 1994; Li and Daggett 1996). A pairwise  $C\alpha$  RMSD matrix is generated for all trajectory conformations and then projected down into two or three dimensions using multi-dimensional scaling. Visual clustering then separates the native conformations from the non-native, placing the TSE at the exit of the native cluster. While the method does require a significant level of subjective judgment, the Daggett group has had good success correlating results of their unfolding simulations to protein engineering studies of the same proteins. Conformational clustering was performed for each of the unfolding simulations here, with the three-dimensional projection of the 500K1 trajectory shown in Figure 4. Individual conformations extracted every 10 ps are shown as spheres and are connected chronologically by sticks; the color goes from blue to red as the simulation progresses. The first 1.41 ns of 500K1 is tightly clustered around the native state (lower left) and then rapidly moves away from the native state, forming much less dense clusters as it progresses through the simulation. Similar behavior is seen for the other unfolding simulations, allowing them to be effectively clustered (Table 1). However, it is much more difficult to identify a common TSE by conformationally clustering all five unfolding simulations simultaneously; hence we sought a method that would allow a common TSE to be generated, testing the conformationally clustered TSE.





**Figure 1.4: Conformational clustering effectively defines the exit from the native state.**

3-D representation of conformational clustering of 500K1 generated by multi-dimensional scaling of the all-versus-all two-fit  $C\alpha$  RMSD. Each sphere is a conformation from every 10 ps of 500K1 and is connected by sticks to the preceding and following conformation. The earliest conformations are colored blue and the latest red. E1, E2, and E3 represent the first through third eigenvectors from the multi-dimensional scaling. The exit from the native cluster is identified by the arrow and is at 1.41 ns.

**Table 1.1: Time (ns) at the native cluster exit for the five  $\alpha$ LP unfolding simulations.**

	Conformational Clustering	NPSASA-Native Contacts	PCA Landscape
500K1	1.41	1.41	1.41
500K2	1.83	1.80	1.79
500K3	1.92	2.18	2.17
500K4	1.40	1.46	1.48
500K5	1.98	1.94	1.97

The only significant difference between the conformational clustering and the landscape methods is for 500K3.

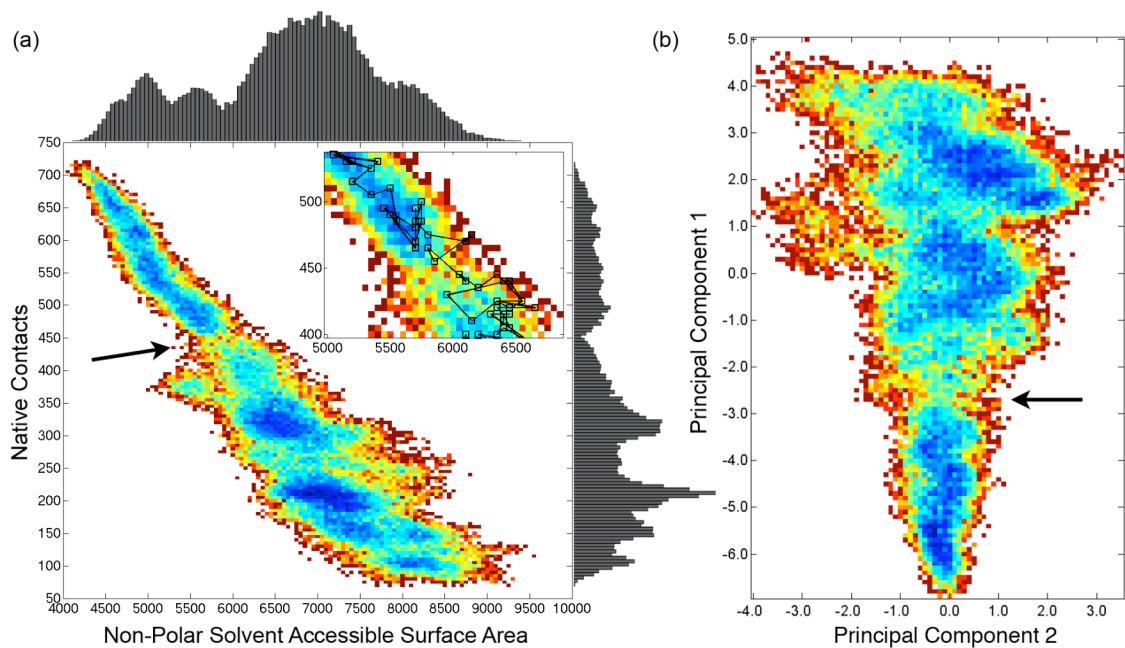
Although the ALF metric captures some of the significant changes during unfolding, it should be possible to gain a better picture of the unfolding process across all of the unfolding simulations, by not looking as a function of time, but rather through changing properties. Using common protein folding/unfolding metrics (here, the number of native contacts and NPSASA) as order parameters, we have computed a single two-dimensional unfolding landscape that integrates data from all the simulations despite their individual differences in timing (Figure 5A). Histograms of the individual metrics are shown at the top and right of the landscape. The landscape shows three well-populated basins (dark blue), one native-like (upper left) and two progressively less native (middle and lower right). There is a bottleneck in the landscape, shown enlarged in the inset and centered around 450 native contacts and 5900 Å<sup>2</sup> NPSASA, that separates the native from non-native basins. Also shown in the inset is a trace of the 500K1 simulation, at 10 ps intervals, for clarity (the landscape was constructed using conformations at 1 ps intervals, a total of 40500 conformations). Significantly, all simulations cross this bottleneck only once, implying a shared barrier to unfolding with these order parameters. The actual crossing transition occurs at different times in the different simulations, for example occurring between 1.41 and 1.42 ns for 500K1 (Table 1). We propose that this barrier is the location of the  $\alpha$ LP TSE in these simulations and have generated a TSE from the structures making up the barrier (Table 1).

In reality, the  $\alpha$ LP unfolding landscape is highly multi-dimensional and is only approximated by NPSASA and native contacts, which are clearly highly correlated. In order to utilize more of those dimensions, ten parameters were measured for each conformation (details and full listing in Methods). Principal components analysis (PCA)

was used to eliminate the inherent correlations in the parameters and allow visualization in less than ten dimensions. The first two principal components explain 90% of the variance in the parameters and were used to generate a landscape as above (Figure 5B). Again, the region comprising native-like conformations is well-separated from the non-native region by a sparsely populated barrier centered around -2.7 on PC1 and 0.0 on PC2. Crossing times for all of the simulations are within 30 ps of the crossing times in the NPSASA/native contacts landscape, and, as above, we have generated a TSE from the PCA landscapes (Table 1). The first principal component, which contains relatively equal weightings from all ten parameters, is mostly a function of each conformation's nativeness (Table S1). There is little variation in the second principal component in the native-like region, and the simulation trajectories begin to diverge more significantly upon reaching the unfolding barrier. The second principal component is dominated by the size of the molecule and backbone exposure to solvent, as the three largest components are non-native mainchain hydrogen bonds, polar SASA, and radius of gyration (Table S1).

With the exception of 500K3, there is remarkable agreement on the TSE location between the landscape methods and the conformational clustering method, despite the vast differences between them. For the other four simulations, the TSEs generated by all three methods are qualitatively identical and quantitatively differ only slightly. For several reasons, we believe the TSE generated from the landscape is the more accurate one. First, the visual clustering is inherently more subjective than the landscape methods, as, at least with  $\alpha$ LP, there is no ambiguity in both the locations of the barriers in the unfolding landscapes or that each simulation only crosses them once. Second, visually

clustering the 500K3 trajectory is more difficult than the other four trajectories, making determination of the termination of the native cluster somewhat ambiguous. In addition, the conformational changes in 500K3 between 1.92 and 2.17 ns are more similar to the other unfolding simulations' changes prior to the TSE, arguing that the landscape TSE is the correct one. Finally, the coincidence of four out of five conformational clustering TSEs with the barriers in the landscapes created by all five simulations argues strongly that these barriers are the true TSE location.



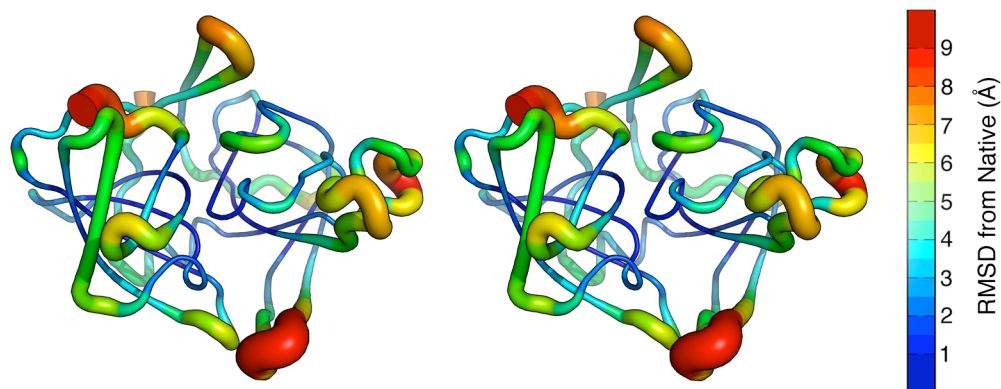
**Figure 1.5: Property-based landscapes clearly separate native from non-native conformations.**

(a) The unfolding landscape is generated from all five unfolding simulations using native contacts and NPSASA as order parameters. 1-D histograms of native contacts and NPSASA are found to the right and above the landscape, respectively. The landscape was generated by taking the negative natural logarithm of the 2-D histogram with white being unobserved in the simulations, dark red the least populated, and progressing to dark blue as the most populated. The native state is in the upper left corner. A less populated region (indicated by the arrow) centered around 450 native contacts and  $5900 \text{ \AA}^2$  separates native-like conformations from non-native conformations and represents the TSE. (inset) Zoomed-in view of TSE region, with trace of 500K1 overlaid. 500K1 crosses the TSE barrier only once and in less than 10 ps, between 1.41 and 1.42 ns; other simulations exhibit similar behavior. (b) Principal components analysis was used to reduced ten conformational properties to two dimensions (see Methods for list of properties). Coloring is the same as in (a). The native state is the well-populated region at the bottom of the figure and is separated from the non-native state by a barrier near -2.7 in PC1 (indicated by the arrow). Note that significant spread in PC2 is only seen after the TSE, as many more conformations are accessible in the unfolded state.

### *Unfolding Pathway and the TSE*

For the remainder of this work, the  $\alpha$ LP TSE is derived from the PCA landscape, generated by taking the conformations spanning the barrier crossing for each of the individual simulations (10 ps, conformations saved at 1 ps intervals) and combining them, yielding a TSE with 50 conformations. Some general properties of the TSE are listed in Table S2. Due to heterogeneity in large portions of the molecule, it is difficult to visualize the entire set of conformations (representative members are shown in Figure S1). As one way of visualizing the TSE, all TSE conformations and the crystal structure were superimposed using the structural superposition program THESEUS and the average deviation from the crystal structure at each  $C\alpha$  over all conformations was computed (Theobald and Wuttke 2006; Theobald and Wuttke 2008). These deviations were then mapped onto the crystal structure by color and thickness of the tube used to represent the backbone, as seen in Figure 6. Several observations can be made from this representation. First, significant deviations from the crystal structure are confined to several regions, notably those mentioned above. Much of the molecule is quite native-like, including the sheet  $\beta 2$ - $\beta 3$ - $\beta 4$ - $\beta 7$ - $\beta 6$  in the N-terminal domain and most of the  $\beta$ -barrel in the C-terminal domain. Second, as evident in stereo, the “front” face of  $\alpha$ LP as depicted deviates far more from native than the “back” face. The “front” face contains the active site and these deviations would severely disrupt enzymatic activity. In addition, preliminary native state hydrogen exchange experiments found that denaturing agents had a more significant effect on the “front” face of  $\alpha$ LP (Davis 1996). Third, with the exception of the 163-178 loop, which is not conserved, unfolding of the regions identified

in each of the unfolding simulations,  $\beta 1$ , the Domain Bridge, and the active site hairpins, would disrupt the domain interface and expose much of it to solvent.



**Figure 1.6: The structure of the  $\alpha$ LP TSE.**

Deviations from native in the  $\alpha$ LP TSE are restricted to several regions, mostly in the domain interface. Stereo view of the average  $C\alpha$  RMSD at each residue in the PCA landscape TSE from the crystal structure is mapped onto the crystal structure. Both the thickness of the cartoon and the color indicate the deviation from native, with thicker representations meaning larger deviations.

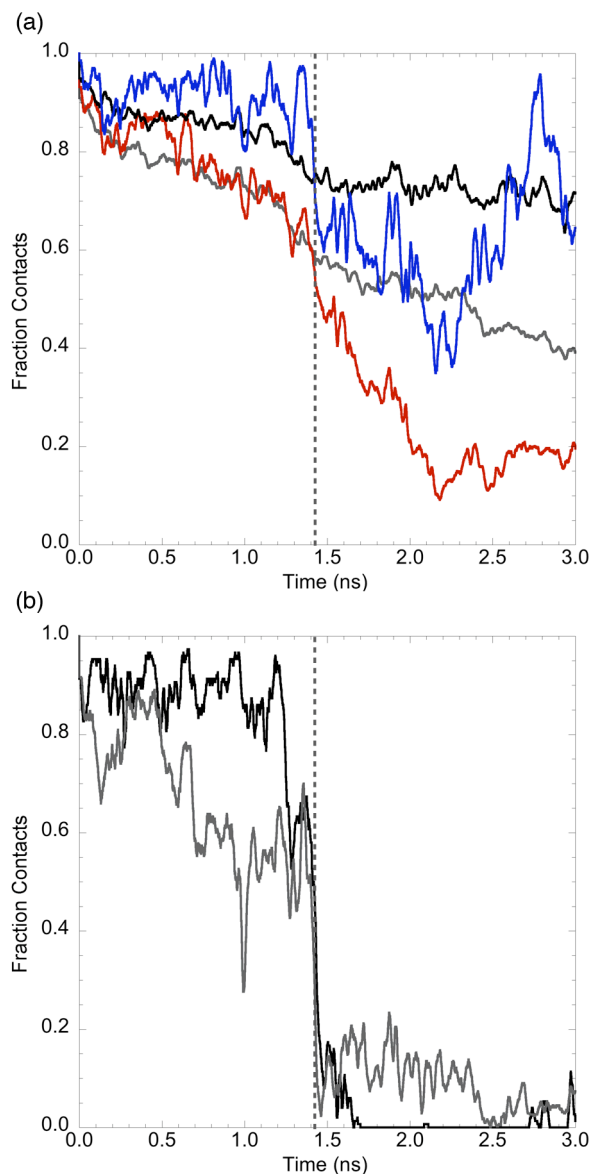
### *The Domain Interface's Role in Unfolding*

The entropic nature of the  $\alpha$ LP unfolding barrier previously led us to hypothesize a solvated domain interface at the TSE, and investigations of the pH-dependence of unfolding has lent credence to that model (Jaswal, Truhlar et al. 2005; Kelch, Eagen et al. 2007). The TSE model presented here (Figure 6) is consistent with the individual domains remaining well-folded throughout the unfolding transition, but is seemingly at odds with the hypothesis that the domains open up in the TSE. To better investigate the domain interface's response to unfolding, we calculated the number of residue-residue intra-domain and inter-domain contacts present in each simulated conformation and normalized them by the corresponding number present in the crystal structure (shown for

500K1 in Figure 7A). Note that at the TSE (1.4 ns), the drop in inter-domain contacts is much more steep and continues much longer than the relatively shallow drop in intra-domain contacts. This effect is exaggerated if only contacts present in the crystal structure are considered. Gray and red curves represent these native intra-domain and inter-domain contacts, respectively. As before, native inter-domain contacts are being lost much more quickly at the TSE. At 2.0 ns, just 0.6 ns after the TSE, only 15% of native inter-domain contacts remain while 50% of native intra-domain contacts are present. This general pattern holds for the other unfolding simulations, providing additional evidence that a key step in  $\alpha$ LP unfolding is the opening of the domain interface.

The Domain Bridge is an integral part of the  $\alpha$ LP domain interface and has been experimentally implicated as a determinant of the unfolding rate (Kelch and Agard 2007). To quantify its role in unfolding, we have calculated the normalized number of native contacts it makes, as above, though using atom-atom contacts due to the relatively small number of residues. Plotted in Figure 7B are the fraction of native contacts between two residues both in the Domain Bridge (DB-DB, black) and between one residue in the Domain Bridge and any other residue (DB-O, green) for the first 3 ns of 500K1. DB-DB contacts are quite stable until the molecules begins to unfold significantly at 1.2 ns, reaching about 60% of native, and then losing all native contacts right at the TSE at 1.41 ns. DB-O contacts are lost more gradually prior to the TSE than DB-DB contacts, but they experience the same steep loss at the TSE. With the exception of 500K3 as noted above, the other unfolding simulations exhibit similar behavior. The high unfolding cooperativity of the Domain Bridge and its coincidence with the TSE observed here is consistent with the previous experimental studies.





**Figure 1.7: Contacts at the domain interface are preferentially broken at the unfolding transition.**

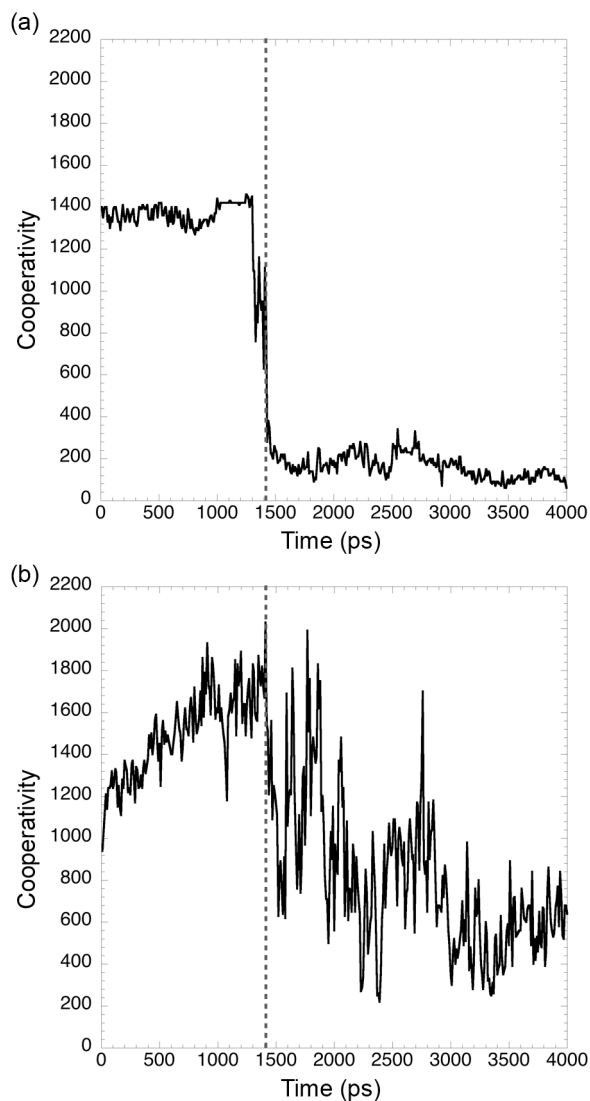
(a) The fraction of intra-domain (black), inter-domain (blue), native intra-domain (gray), and native inter-domain (red) are shown for the first 3 ns of 500K1. Inter-domain contacts experience a sharp drop at the native cluster exit (dashed vertical line, 1.41 ns) and continue to decline. Intra-domain contacts are lost more gradually. Shortly after unfolding, ~90% of native inter-domain contacts are lost permanently. (b) The fraction of native domain bridge-domain bridge (black) and native domain bridge-other (gray) contacts for the first 3 ns of 500K1. Both decline sharply at the native cluster exit (dashed vertical line) and do not return to native-like values. For both (a) and (b), the data is smoothed with a 0.019 ns running average.

## *Unfolding Cooperativity*

A critical feature for  $\alpha$ LP's kinetic stability is its extremely high unfolding cooperativity. Previous work has shown that while  $\alpha$ LP and trypsin, a thermodynamically stable homolog, have similar unfolding rates,  $\alpha$ LP unfolding is much more cooperative as measured by proteolysis, providing it a functional advantage in highly proteolytic environments (Jaswal, Sohl et al. 2002; Truhlar, Cunningham et al. 2004). Because determining the origins of this remarkable difference is crucial for understanding the molecular basis for kinetic stability, we sought to compare the behaviors of  $\alpha$ LP and trypsin as revealed by unfolding simulations. Four 10.1 ns unfolding simulations at 500K were performed for trypsin. Although a thorough discussion of the details of the trypsin TSE and unfolding pathway will be presented elsewhere, the general behavior of these simulations is reported in the Supporting Information (Figures S2 and S3, Table S3).

To quantitatively compare unfolding cooperativity, we developed a new metric defined by how many conformations were similar (based on a C $\alpha$  RMSD threshold) to the  $i$ th conformation within the  $n$  total simulation conformations. The cooperativity graph for a perfectly cooperative unfolding transition would be high and flat for the beginning of the simulation, drop steeply at the TSE, and then be much lower for the duration of the simulation. Specifically, it would have a value of  $j$  from 1 to  $j$ , where  $j$  is the TSE conformation, and drop to a value  $k \ll j$  after the TSE. Less cooperative transitions would feature gradually increasing and/or decreasing values prior to the TSE and less steep drops after the TSE. Cooperativity for  $\alpha$ LP (500K1) and trypsin (500K2T) are shown in Figure 8A and 8B, respectively. The cooperativity profile for  $\alpha$ LP is very similar to that of the hypothetical perfectly cooperative unfolding transition. Before the

TSE at 1410 ps, the value is near 1400 and relatively flat. It drops sharply right at the TSE, and then is much lower for the duration of the simulation. Trypsin, on the other hand, unfolds much less cooperatively. Its increasing profile from 0 to 900 ps represents gradual unfolding, because structures that have partially unfolded are similar to both the native structure and to more unfolded conformations. It has no clear steep drop from the native state as  $\alpha$ LP does, only a gradual and very noisy decline. Its values post-TSE are much higher than those observed for  $\alpha$ LP, which suggests a more rugged, gradual unfolding process. Cooperativity plots for the other simulations show the same general trends and the behavior is qualitatively similar with different choices of  $C\alpha$  RMSD thresholds. We believe this work is the first example of both measuring cooperativity in simulated unfolding and comparing it across two proteins where that difference has functional relevance.



**Figure 1.8:  $\alpha$ LP unfolds significantly more cooperatively than trypsin.**

Cooperativity is measured by counting the number of sampled conformations  $< 3 \text{ \AA}$   $C\alpha$  RMSD (two-fit  $C\alpha$  RMSD, see Methods) from the conformation at each time point. (a) Cooperativity for the first 4 ns of 500K1. Starting flat and steeply dropping indicates a very cooperative unfolding transition for  $\alpha$ LP. (b) Cooperativity for the first 4 ns of 500K2T (trypsin). Trypsin unfolds much less cooperatively than  $\alpha$ LP, as seen by the gradual rise early in the simulation and the gradual and noisy decline starting at 1.4 ns. (a) and (b) Vertical dashed line indicates position of the native cluster exit in each simulation.

## ***Discussion***

A major motivation for this study was providing atomic resolution to previous biochemical experiments on  $\alpha$ LP unfolding, but first those lower resolution results must be reproduced. A comprehensive analysis of experimental data on protein unfolding barriers revealed a stark difference between those of  $\alpha$ LP and thermodynamically stable proteins: the  $\alpha$ LP unfolding barrier is significantly more entropic, suggesting the  $\alpha$ LP TSE is considerably more native-like than those for thermodynamically stable proteins (Jaswal, Truhlar et al. 2005). In addition, m-value analysis of unfolding found the fraction of SASA buried at the  $\alpha$ LP TSE was computed to be 80%, also highly-native like (Jaswal 2000). The simulation-derived TSE reported here is quite similar to the native structure, with an average C $\alpha$  RMSD of  $4.4 \pm 0.4$  Å and with 38% of C $\alpha$  atoms being less than 2.0 Å from native. The average fractional SASA at the TSE is  $82 \pm 2$  %, slightly higher but quite consistent with the value derived from experiment. One possibility for the slight deviation is that the elevated temperature in the simulations shifts the TSE somewhat towards the native, a modest Hammond effect that was seen for CI2 (Day, Bennion et al. 2002; Day and Daggett 2005).

Previous experiments have shown a large role for the domain interface in  $\alpha$ LP unfolding. The thermodynamic analysis referenced above suggested a possible model for the TSE: a “cracked egg” where the two  $\beta$ -barrel domains are largely intact but the extensive domain interface between them is disrupted (Jaswal, Truhlar et al. 2005; Kelch, Eagen et al. 2007). Relocation of salt bridges spanning the domain interface significantly decreased  $\alpha$ LP’s sensitivity to low pH unfolding, consistent with the “cracked egg” model (Kelch, Eagen et al. 2007) (P. Erciyas, private communication). The simulations

presented here confirm the disruption of the domain interface at the TSE, provide atomic detail as to how it happens, and extend these insights to two other critical structural regions:  $\beta 1$  and the CPT and C $\beta$ H.

The Domain Bridge, the covalent linkage between  $\alpha$ LP's two domains, has been shown to modulate the unfolding rate (Kelch and Agard 2007). The simulations support this; they reveal that many of its native contacts are lost at the TSE, including separation of its strands, allowing it to make non-native contacts. The domain bridge makes several contacts with the N-terminal  $\beta$ -strand  $\beta 1$ , which is also significantly disrupted at the TSE. Our results indicate a probable coupling of the unfolding of the Domain Bridge and  $\beta 1$ , though the coupling is less evident in 500K2 and 500K3. When full-length Pro- $\alpha$ LP is synthesized, the C-terminus of the Pro region is covalently connected to the protease's N-terminus. As the protease domain folds, it gains proteolytic activity, cleaving the Pro- $\alpha$ LP junction that is positioned across the active site (Silen, Frank et al. 1989; Sauter, Mau et al. 1998). The active site is 20 Å away from the location of the N-terminus in the native state and hence folding requires a significant rearrangement of the N-terminal strand. The flexibility of the N-terminus at the TSE in our simulations is consistent with its requirements during Pro-assisted folding. The last region at the domain interface disrupted at the TSE forms part of the active site.

Previous studies on  $\alpha$ LP have also implicated the C $\beta$ H as important to the folding/unfolding landscape. Mutations in the hairpin affected both the unfolding rate and the Pro-catalyzed folding rate (Peters, Shiau et al. 1998; Truhlar, Cunningham et al. 2004; Ho and Agard 2008). The Pro- $\alpha$ LP complex structure revealed that this hairpin forms a larger five-stranded  $\beta$ -sheet with Pro; mutants disrupting the interface there

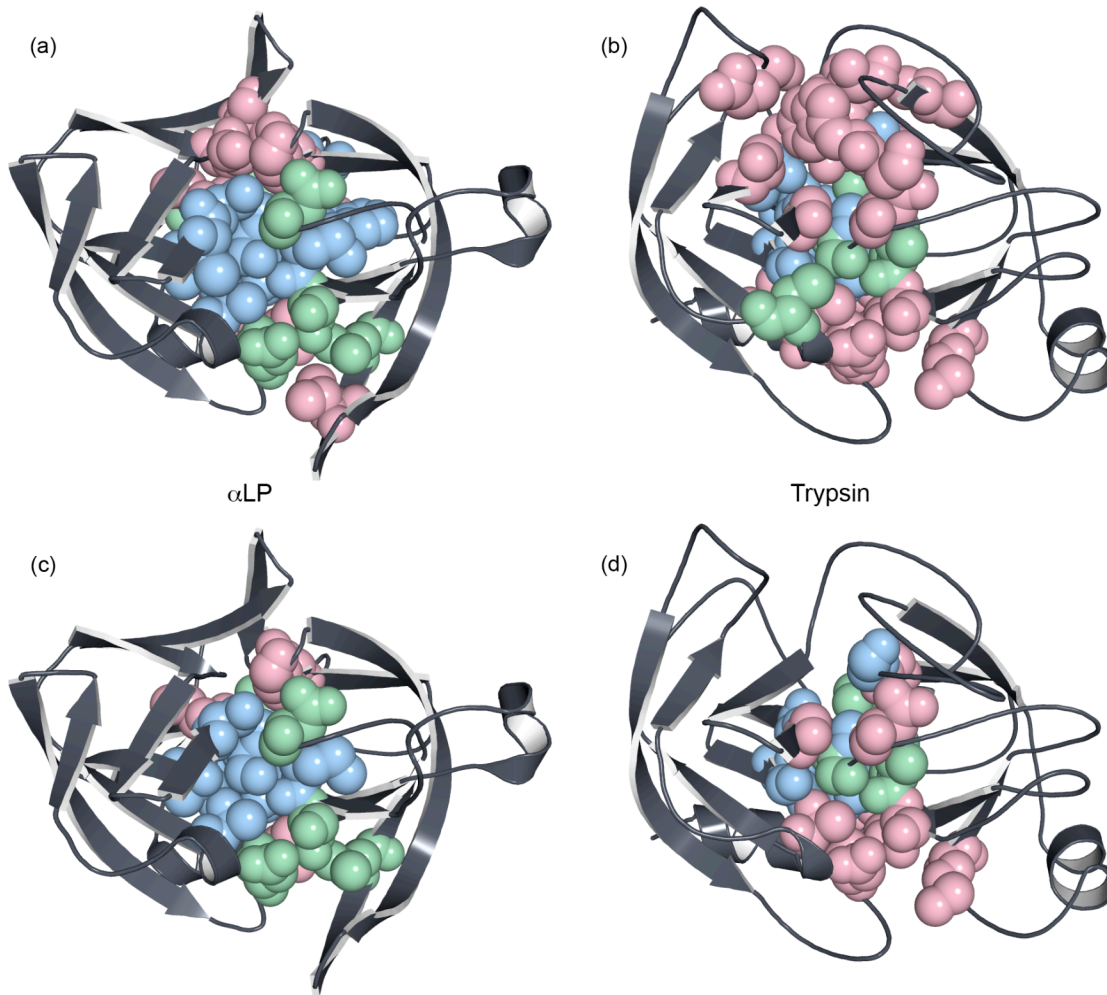
significantly weaken Pro's foldase activity (Sauter, Mau et al. 1998). The hairpin forms several side-chain contacts and two main-chain hydrogen bonds with the CPT in the native state; CPT residues F59, which forms the bulk of the contacts with the C $\beta$ H, and *cis*-P60 are both completely conserved in kinetically stable proteases. The amides in these hydrogen bonds have relatively weak protection factors compared to the rest of the protein, consistent with them being broken at the TSE (Jaswal, Sohl et al. 2002). These contacts are also relatively long-range in sequence space, requiring that the molecule must give up significant conformational entropy in bringing them together, again arguing that they are broken early in unfolding. In our simulations, once the contacts between the two structures are broken, C $\beta$ H pulls away from the body of the protein and its strands separate; the presence of the Pro region would keep the hairpin in a position ready to make contacts with the *cis*-proline turn, stabilizing the TSE. By understanding the  $\alpha$ LP unfolding pathway and TSE in atomic detail, we can begin to explore how the Pro region stabilizes the TSE and accelerates folding  $10^9$ -fold.

The three regions of the domain interface disrupted at the TSE have something else in common: they are only found in the kinetically stable proteases and not the thermodynamically stable family members, such as trypsin.  $\alpha$ LP and trypsin are good structural homologs; 120 (of 198)  $\alpha$ LP's C $\alpha$ 's have an equivalent position in trypsin, with 99 of them within 2.0 Å of their trypsin equivalent (Lesk and Fordham 1996). It seems an unlikely coincidence that the regions of  $\alpha$ LP that unfold at the TSE happen to be in the 1/3 of the protein that is not homologous to trypsin. In fact, for both the  $\alpha$ LP and trypsin simulations, the structurally conserved regions are much more native-like at the TSE and beyond than are the non-conserved regions. Significantly, large parts of the  $\alpha$ LP domain

interface are made up of the non-conserved regions, likely resulting in the dramatic differences observed between folding of individual  $\alpha$ LP and trypsin domains. For both chymotrypsin and trypsin, the two domains fold independently and upon mixing will form the active enzyme (Duda and Light 1982; Higaki and Light 1986). By contrast, active  $\alpha$ LP cannot be reconstituted from unfolded individual domains even in the presence of the Pro region (Cunningham and Agard 2003).  $\alpha$ LP's cooperativity in folding echoes that of the unfolding reaction and likely involves the Domain Bridge and other regions of the domain interface which are distinct from its metazoan homologs.

By closely examining the differences between the domain interfaces of  $\alpha$ LP and trypsin, we can begin to discover the mechanism of  $\alpha$ LP's unfolding cooperativity. The buried residues (less than 5% exposed in the crystal structure) of the  $\alpha$ LP and trypsin domain interfaces are shown in space-filling spheres in Figure 9A and 9B, respectively. Residues colored light red are exposed to solvent at the TSE, while residues still buried at the TSE are further subdivided into light green and light blue residues, which at 600 ps post-TSE are either exposed to solvent or still buried, respectively. For both  $\alpha$ LP and trypsin, residues near the "top" and "bottom" of the molecules are more likely to be exposed at the TSE, while the "middle," which contains the conserved active site, has fewer red residues. Clearly, more of trypsin's buried domain interface residues are exposed to solvent at the TSE than for  $\alpha$ LP. In addition, the  $\alpha$ LP core is much more blue than trypsin, as this core is much more resistant to solvation even post-unfolding than its metazoan counterpart.





**Figure 1.9: Solvation of the domain interface during unfolding differs significantly between  $\alpha$ LP and trypsin.**

(a) - (d) Residues colored light red are solvent-exposed at the TSE, light green residues become exposed within 600 ps of the TSE, and light blue residues are still buried 600 ps past the TSE. (a) and (c)  $\alpha$ LP, (b) and (d) trypsin. (a) and (b) All buried domain interface residues, (c) and (d) the subset of (a) and (b) where the position is conserved and found at the domain interfaces of both prokaryotic and metazoan proteases. Notably, many fewer buried residues of the  $\alpha$ LP domain interface are solvated at the TSE compared to trypsin, even after eliminating the non-conserved positions.

As alluded to previously, much of the domain interface is not conserved between the two families of proteases. Figures 9C and 9D focus on the conserved interface, and correspond to 9A and 9B, respectively, after removing all residues that are not common to both domain interfaces (this implies position and sequence conservation). Here, the

difference between  $\alpha$ LP and trypsin is striking; over half of the positions in trypsin are red (solvent exposed), while residues in  $\alpha$ LP are generally exposed to solvent much later, over half of them blue. An area near the active site (see Figure 1) comprising  $\alpha$ LP residues D63, L131, and S159 (all green) and their trypsin equivalents D84, M160, and S192 (all light red) is particularly interesting. Again, this region is composed of the C $\beta$ H unique to the kinetically stable proteases, and though it unfolds at the TSE, only once it completely unfolds does it begin to expose the conserved core to solvent. This is not the case in trypsin, where the different architecture, composed of loops, allows relatively small unfolding events to expose the buried interface.

Examining the differences between the full-domain interface and the conserved domain interface figures then highlights the non-conserved regions. At the  $\alpha$ LP Domain Bridge, its unfolding exposes relatively little of the domain interface at the TSE, while the much larger equivalent area in trypsin is quite solvated. An important difference between the two proteases is that the  $\alpha$ LP Domain Bridge is a compact, cooperative substructure, a simple  $\beta$ -hairpin. In trypsin, the domain interface is formed by two long and relatively floppy loops, which are inherently less cooperative than the domain bridge. Many of the non-conserved domain interface residues in  $\alpha$ LP are also in secondary structure or tightly constrained turns near the Domain Bridge or active site (i.e. G5 and G6 connect  $\beta$ 1 to  $\beta$ 2 and interact with the Domain Bridge, V79 and V88 form the base of the Domain Bridge, V128 is in the C $\beta$ H), while in trypsin these areas are formed with much less constrained loops. The differences seen here in the Domain Bridge and active site regions provide evidence that extreme unfolding cooperativity is generated for these

two-domain proteins by using highly cooperative substructures to protect the rest of the domain interface from solvent.

Intriguingly, increased protease resistance mediated through high inter-domain cooperativity has been observed in an unrelated system (Young, Skordalakes et al. 2007). A screen of the *Escherichia coli* proteome for protease resistance found 40 proteins, one of which was the glycolytic enzyme phosphoglycerate kinase (PGK) (Park, Zhou et al. 2007). Young et al. found that while the *E. coli* and *Saccharomyces cerevisiae* enzymes had similar stabilities, the yeast PGK unfolded and was degraded much faster than the *E. coli* PGK (Park, Zhou et al. 2007). The difference was attributed to the domain interface; the separated domains of yeast PGK fold independently and are quite stable, unlike the *E. coli* PGK domains, analogous to the difference between prokaryotic  $\alpha$ LP and eukaryotic trypsin.

The costs of evolving extreme unfolding cooperativity are high; for  $\alpha$ LP, the bacterium must synthesize a 166 residue protein to catalyze  $\alpha$ LP folding after which it is immediately degraded.  $\alpha$ LP's extremely slow folding is a consequence of the large energy gap between its unfolded/molten globule states and the TSE (Sohl, Jaswal et al. 1998). One likely contributor that has been previously noted is its high glycine content, as glycines in formed structures lose much conformational entropy relative to unstructured glycines (D'Aquino, Gómez et al. 1996). These glycines, which make up 18% of kinetically stable proteases, are used to form tight turns and tight packing in areas where even an alanine would be sterically hindered (Sohl, Jaswal et al. 1998; Fuhrmann, Kelch et al. 2004). Like most proteins, the metazoan proteases have much lower glycine content (about 9%) and have many correspondingly longer loops than the prokaryotic proteases.

These loops, like those in the domain interface of trypsin, are likely the reason for trypsin's lack of cooperative unfolding.

Finally, the idea that a protein's folding transition state is determined by its native structure, as shown through studies of Contact Order and folding rates, poses interesting questions for this class of proteases (Plaxco, Simons et al. 1998; Ivankov, Garbuzynskiy et al. 2003). While trypsin fits well in the Contact Order plot,  $\alpha$ LP is an extreme outlier, perhaps not surprising given its remarkably slow folding (Jaswal, Truhlar et al. 2005). The two proteins have the same fold and would be expected to have similar TSEs. Here, we have identified the TSEs for both, and remarkably, those TSEs both contain much of the conserved core of the fold. However, the regions where the two proteases differ are critical parts of the TSE structures. While the general structure of the TSE may be mostly determined by the native structure, the details, such as highly cooperative units making up the domain interface in  $\alpha$ LP and not trypsin, can provide large functional advantages depending on the environment of the particular protein.

## ***Materials and Methods***

### *Simulations*

1SSX and 5PTP PDBs were used for  $\alpha$ LP and trypsin, respectively. All non-protein and hydrogen atoms were removed and hydrogens were added back with XPLORE (Brunger 1992). For residues with multiple conformations, the "A" conformation was used. Protein molecules were placed in cubic boxes with a minimum of 12 Å distance to the edge and solvated with TIP3P explicit water and chloride counter-ions using Packmol (Martínez and Martínez 2003), where the approximate density was determined by the

density of liquid water at the corresponding temperatures.[48] The number of atoms for 298K  $\alpha$ LP, 500K  $\alpha$ LP, 298K trypsin, and 500K trypsin were 32760, 28005, 33223, 28468, respectively. All simulations were performed using NAMD 2.5 with the CHARMM22 forcefield (MacKerell, Bashford et al. 1998; Phillips, Braun et al. 2005). Simulations were carried out with periodic boundary conditions, a 12 Å cutoff for non-bonded interactions, and Particle Mesh Ewald for long-range electrostatics. A timestep of 1 fs was used and snapshots were saved every 1 ps. Each system was equilibrated using the following protocol. The protein was fully constrained and the solvent was minimized for 500 steps using a conjugate gradient algorithm. The solvent was equilibrated for 100 ps under NPT conditions (298K and 1.01325 bar or 500K and 27 bar) using Berendsen coupling for both pressure (100 fs relaxation time) and temperature (2.0 ps coupling constant) (Berendsen, Postma et al. 1984). The solvent was then fully constrained and the protein was minimized for 50 steps. The entire system was then minimized for 50 steps. Finally, the system was equilibrated for 100 ps under the same NPT conditions. Multiple independent simulations were generated by starting the whole-system equilibration using different random number seeds for each. After equilibration, production simulations were carried out in the NVE ensemble, with the box size fixed at its final size from the equilibration. One 298K  $\alpha$ LP (12.1 ns), five 500K  $\alpha$ LP (8.1 ns each), one 298K trypsin (3.6 ns), and four 500K trypsin (10.1 ns each) simulations were performed for 96.6 ns total simulation time.

### *Two-fit C $\alpha$ RMSDs*

In several analyses presented here (Conformational Clustering, ALF, and Cooperativity), C $\alpha$  RMSDs were calculated with two fits to the target structure in order to lessen the impact of a small number of poorly aligning residues. Structures were aligned using all C $\alpha$  atoms and the mean and standard deviation of the deviations were calculated. C $\alpha$  atoms whose deviations were greater than two standard deviations above the mean were discarded for the second fit and calculation of C $\alpha$  RMSD. This fitting procedure eliminated an average of 5% of the C $\alpha$  atoms.

### *Average Local Fluctuation (ALF)*

For all overlapping 90 ps windows in a simulation, all pairwise two-fit C $\alpha$  RMSDs were calculated for the 10 snapshots (10 ps intervals), resulting in 45 two-fit RMSDs at 801 windows for an 8.1 ns simulation. These RMSDs were averaged to give the ALF at the midpoint of each window. ALF therefore measures the extent of short-timescale (90 ps) fluctuations throughout the simulation, as it is the mean RMSD between any two snapshots within a short time window.

### *Conformational Clustering*

For each simulation, pairwise two-fit RMSDs were calculated at 10 ps intervals, forming a symmetric N x N matrix, with N = 810 for  $\alpha$ LP and N = 1010 for trypsin unfolding simulations. Multi-dimensional scaling, as implemented in the MATLAB Statistical Toolbox, was used to calculate the first three eigenvectors of the RMSD

matrix. The resulting three-dimensional graph, where each point represents a single conformation, was visually clustered to identify the native state ensemble and its exit.

### *Contacts*

Atoms less than 4.6 Å apart or 5.4 Å apart if one of the atoms was C or S and more than two residues separated in the primary sequence were judged to be in contact. A contact was defined as native if the two residues had a contact in the crystal structure. For the purposes of defining inter-domain, intra-domain, and domain bridge contacts in αLP, the N-terminal domain is residues 1-82 and 184-198, the Domain bridge is residues 78-88, and the C-terminal domain is residues 83-183.

### *Native contacts-NPSASA landscape*

For each simulation snapshot, the number of native residue-residue contacts and the NPSASA were calculated. The values were binned into a two-dimensional histogram using bin sizes of 5 native contacts and 50 Å<sup>2</sup>. The landscape was generated by taking the negative natural logarithm of the bin counts at each position.

### *Principal components landscape*

Ten conformational properties were used to generate the landscape: Cα RMSD, native intra-domain atom-atom contacts, native inter-domain atom-atom contacts, non-native intra-domain atom-atom contacts, non-native inter-domain atom-atom contacts, radius of gyration, non-polar SASA, polar SASA, non-native main-chain hydrogen

bonds, native main-chain hydrogen bonds. Properties were scaled by dividing by subtracting the mean value and dividing by the standard deviation for each. Principal components analysis was performed with the MATLAB Statistics Toolbox. Loadings for each term in the PCA are shown in Supplemental Table 1. A two-dimensional histogram was computed using the first two principal components, with a bin size of 0.1 units. The landscape was generated by taking the negative natural logarithm of the bin counts at each position.

### *Cooperativity*

Two-fit  $C\alpha$  RMSDs were calculated for each pair of snapshots (10 ps intervals to reduce the number of pairwise comparisons) in a simulation. Cooperativity was defined as the number of snapshots less than 3 Å of the above  $C\alpha$  RMSD at each time point in the simulation multiplied by the snapshot interval (10 ps). Results were qualitatively similar using thresholds of 3.5 and 4.0 Å.

### *Molecular Graphics*

PyMOL (DeLano 2002) was used to generate Figures 1, 3, 4, 6, and 9.

### *Acknowledgements*

We thank P. Erciyas, Drs. J. Chodera, B. Kelch, and L. Rice for helpful discussions and P. Erciyas, Drs. J. Chodera, B. Ho, Q. Justman, and T. Street for critical



reading of the manuscript. N. L. S. was supported by an NSF Graduate Research Fellowship and an NIH NIGMS Training Grant (GM-008284).

## *Postscript*

### *How else can the simulated TSE be verified?*

While it would still be unfeasible to perform extensive  $\phi$ -value analysis on  $\alpha$ LP to examine its TSE experimentally, there are other approaches that could be and were undertaken. The first was what is called  $p_{\text{fold}}$  analysis. A conformation that is truly a member of the TSE should refold exactly half the time and unfold exactly half of the time. Thus, one could start many simulations using TSE conformations and simply count how many refold or unfold.  $p_{\text{fold}}$  has been applied to small proteins, often in implicit solvent, with impressive results (Gspomer and Caflisch 2002). I attempted to do the same with conformations from the 500K1  $\alpha$ LP simulation.

I took conformations spanning a range of 500 ps on both sides of the TSE and simulated them at 325K, a slightly elevated temperature in order to ensure movement along the folding/unfolding trajectory. Because the unfolding simulations were performed in explicit solvent, I performed the simulations with explicit solvent, which makes them much slower than implicit solvent simulations. Each simulation was at least 1 ns. Unfortunately, little changed during the simulations, as they were most likely too short and possibly at too low a temperature for  $\alpha$ LP and explicit solvent. At this time I began focusing on trypsin simulations, thinking they were a better use of limited computational resources. I had some discussions with John Chodera in Vijay Pande's lab about using his `folding@home` setup to do similar simulations, but nothing came of them.

The second possibility was to use a different mutagenesis strategy to experimentally verify the TSE. Tobin Sosnick and colleagues have developed what they call  $\psi$ -value analysis as a method of examining TSE structure (Krantz, Dothager et al. 2004; Sosnick, Dothager et al. 2004). The method relies on double His mutants, which are placed so that they will be surface exposed and able to coordinate a divalent metal cation, such as  $Zn^{2+}$  or  $Ni^{2+}$ , in the native state. Because the cation will stabilize the His-His contact up to several kcal/mol, the level of formation of the His-His contact in the TSE can be determined by measuring the folding/unfolding kinetics as a function of [cation]. The benefit of  $\psi$ -value analysis over  $\phi$ -value analysis is that the mutations are at worst only weakly stabilizing, which would presumably allow better expression and folding kinetics compared to destabilizing hydrophobic deletion mutants. Pinar Erciyas had planned on doing some of these experiments, but she became more involved in mutagenizing  $\alpha$ LP's inter-domain salt bridges, following up on work started by Brian Kelch and based on insights from the NAPase structure (Kelch, Eagen et al. 2007).

*Are other experiments also supporting the critical role of the domain interface?*

As mentioned above, Pinar Erciyas has been following up on the inter-domain salt bridges and their role in unfolding. She has managed to generate all combinations of one and two salt bridge deletions, with intriguing results. Her mutations confirm that deletion of the salt bridges reduces  $\alpha$ LP's sensitivity to low pH. One particular mutation, Del3 (R64A/E182Q) increases the unfolding rate at pH 5 by about two orders of magnitude. Del1 (E8A/R105S) only slightly increases the  $\alpha$ LP unfolding rate, but acts synergistically

with Del3 to produce a significantly faster unfolding mutant. The mechanism of this synergy will need to be explored.

Bosco Ho has developed a somewhat novel methodology for performing “pulling” experiments using molecular dynamics, and has recently begun applying them to  $\alpha$ LP and trypsin. In what are very preliminary results, it appears that when he pulls the two  $\beta$ -barrels apart, there is a significant barrier to the separation of  $\alpha$ LP domains but a much smaller barrier for trypsin. These studies will need to be replicated and made more rigorous before putting significant stock in them, but appear to be interesting enough to move forward on.

## **Chapter 2: Understanding unfolding cooperativity through an information theory measure of contact pair cooperativity in molecular dynamics simulations**

### *Preface*

This work began after I remembered a talk by Vince Voelz, a former graduate student in Ken Dill's lab. He talked about a cooperativity metric that he used to progressively fold proteins through an algorithm he and others in the lab developed. When I began this project, Vince's methods had only been published in his dissertation, but have since been published in a paper studying the simulations of peptide fragments from larger proteins and their use in predicting protein structure. MCOOP, the term for the cooperativity metric he devised (and that is used here), plays a very small role in his paper.

After finishing the unfolding simulations, David Agard and I began thinking of ways to test their results. Because the functional difference between  $\alpha$ LP and trypsin is the large difference in unfolding cooperativity, the obvious experiment was finding mutations that would eliminate  $\alpha$ LP's advantage over trypsin. Even though we had identified the unfolding pathway and hypotheses about which residues were critical, the list of mutations would still be large. Once I remembered Vince's method and applied it to my simulations, we had found a way to much more rigorously identify target residues for mutation.

As of this writing, this chapter is a draft that will soon be submitted for publication, most likely to *Protein Science*. David Agard will appear as the second author, having contributed intellectually to the science and assisted in the writing of the paper. I performed all of the analysis and wrote the paper, and will be appearing as the first author.

## *Synopsis*

Protein unfolding kinetics are often modeled as a two-state process, with a rate-limiting Transition State Ensemble (TSE) separating the native from unfolded states. However, partially unfolded states between the native state and TSE can be transiently populated and are subject to aggregation and/or proteolysis under certain conditions. Some proteins, such as  $\alpha$ -lytic protease ( $\alpha$ LP), have evolved mechanisms for extreme cooperativity in unfolding, i.e. not detectably populating partially unfolded states. This cooperativity allows  $\alpha$ LP a longer functional lifetime in highly proteolytic environments, an advantage not found in its metazoan homolog trypsin. Previous biochemical studies pointed to the domain interface as key to  $\alpha$ LP's unfolding cooperativity. In order to determine the mechanism of cooperativity in  $\alpha$ LP, I previously carried out high temperature molecular dynamics unfolding simulations of  $\alpha$ LP and trypsin. These simulations confirmed the role of the domain interface in unfolding, showed correctly that  $\alpha$ LP unfolds cooperatively while trypsin does not, and revealed that domain interface structural elements found in  $\alpha$ LP and not in trypsin were responsible for the cooperativity. Here, I apply new methodology from Dill and co-workers to investigate cooperativity at the contact and residue level from the previous simulations. Networks of cooperative contact pairs reveal how cooperativity is distributed throughout each protease, with  $\alpha$ LP clearly more cooperative than trypsin. Mapping network clusters onto the protein structure identifies cooperatively unfolding units. Residues involved in many cooperative contacts are also identified; these residues present excellent targets for mutagenesis aimed at disrupting  $\alpha$ LP's unfolding cooperativity.

## ***Background***

Many biochemical systems utilize cooperative mechanisms in order to carry out their functions. These processes, including O<sub>2</sub> binding by hemoglobin (Mills, Johnson et al. 1976), ATP binding and hydrolysis by GroEL/GroES (Bochkareva, Lissin et al. 1992), and the folding of small, single-domain proteins (Dill, Bromberg et al. 1995), increase their efficiency through cooperativity by significantly reducing the vast number of potential intermediates between end states, many of which may be non-functional or harmful. For protein unfolding, cooperativity can also provide large functional benefits.

Cooperative unfolding describes a process by which the protein unfolds in an all-or-nothing manner, i.e., not detectably populating partially unfolded states. What is the advantage of cooperative unfolding? Unfolded polypeptides, whether they are whole proteins or short stretches of a larger molecule, are susceptible to inactivation through aggregation with other polypeptides and proteolysis by proteases. Because partial unfolding occurs faster than the global unfolding rate, partial unfolding increases the rate of protein inactivation under conditions conducive to aggregation and/or proteolysis. Hence, high unfolding cooperativity can extend a protein's functional lifetime under harsh conditions.

For  $\alpha$ -lytic protease ( $\alpha$ LP), a prokaryotic serine protease, extreme unfolding cooperativity slows degradation by a factor of 20 relative to its metazoan homolog trypsin, even though their global unfolding rates are nearly identical (Jaswal, Sohl et al. 2002; Truhlar, Cunningham et al. 2004). The cooperativity is seen in significantly reduced native state fluctuations, experimentally through  $\alpha$ LP's low B-factors and extremely large hydrogen exchange protection factors (Jaswal, Sohl et al. 2002;



Fuhrmann, Kelch et al. 2004) and computationally through low C $\alpha$  RMSDs in molecular dynamics (MD) simulations (Ota and Agard 2001) and a novel method of discovering flexible regions in proteins (Ho and Agard 2009). The evolutionary consequence of  $\alpha$ LP's unfolding cooperativity is the instability of its native state relative to the unfolded state;  $\alpha$ LP is kinetically stable, meaning its native state is maintained only by its very slow unfolding rate ( $t_{1/2} = 1$  year) (Sohl, Jaswal et al. 1998; Truhlar, Cunningham et al. 2004). Structurally, both  $\alpha$ LP and trypsin are composed of two domains, each made up of a six-stranded  $\beta$ -barrel, rotated approximately perpendicular to each other (Figure 1). The active site lies between the two domains and is well-conserved, but other parts of the domain interface differ significantly between the kinetically stable bacterial proteases and the thermodynamically stable metazoan proteases. Recent biochemical evidence suggested that the  $\alpha$ LP domain interface is broken early in unfolding, and that the Domain Bridge, a hairpin connecting the two domains in  $\alpha$ LP, is a key modulator of the unfolding rate (Kelch and Agard 2007; Kelch, Eagen et al. 2007).

In order to gain a comprehensive view of the  $\alpha$ LP unfolding pathway and insight into its remarkable unfolding cooperativity, we previously performed multiple high temperature MD unfolding simulations of both  $\alpha$ LP and trypsin (Salimi 2009). For  $\alpha$ LP, we found good agreement between previous biochemical experiments and the computational pathway, showing that the domain interface, including the Domain Bridge, is preferentially broken at the transition state ensemble (TSE). Importantly, inter-domain contacts in regions not homologous with trypsin were broken before the conserved region. For trypsin, the non-homologous regions unfolded much earlier and allowed the core conserved region to be exposed to solvent at the TSE, unlike in  $\alpha$ LP. Using a novel

method of measuring global cooperativity based on conformational similarity, we showed  $\alpha$ LP unfolds much more cooperatively than trypsin, with the divergent regions of the domain interface playing a key role.

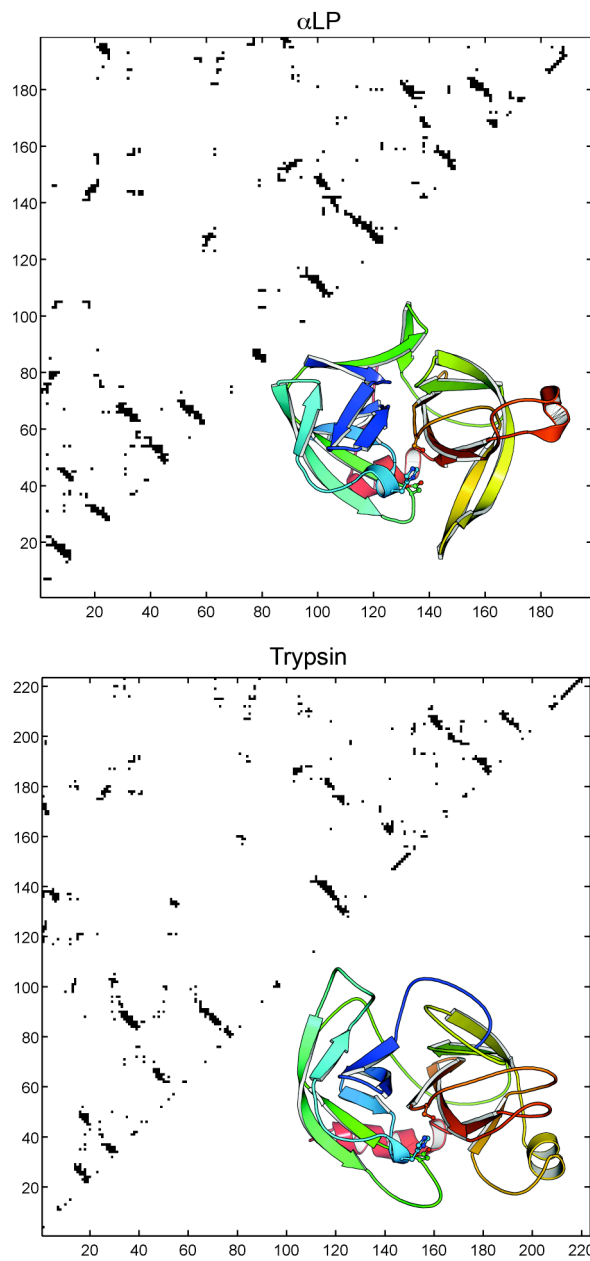
Here, we take advantage of recent work from Dill and co-workers in order to describe unfolding cooperativity at the residue level and examine its distribution throughout a protein. In that publication, Voelz et al. built on previous models of cooperativity from simple model systems (Chan and Dill 1991; Dill, Fiebig et al. 1993) and devised a metric for calculating the cooperativity of pairs of residue-residue contacts from MD simulations based on information theory (Voelz, Shell et al. 2009). Defined simply, contact pair cooperativity (MCOOP), is the extra information gained from knowing the joint probabilities for the formation of contacts  $c_x$  and  $c_y$  in addition to their marginal probabilities. Equation 1 quantifies that information in bits

$$MCOOP = \sum_{c_x} \sum_{c_y} p(c_x, c_y) \log_2 \frac{p(c_x, c_y)}{p(c_x)p(c_y)} \quad (1)$$

where the formation of contacts  $c_x$  and  $c_y$  is binary and  $p(c_x)$ ,  $p(c_y)$ , and  $p(c_x, c_y)$  are the probabilities of each contact state. MCOOP ranges from 0 bits (no cooperativity) to 1 bit (perfect cooperativity). Figure 2 illustrates MCOOP for two hypothetical distributions of a contact pair, where in both cases the individual contacts are formed at  $p = 0.35$  (state 1), but the distribution on the left is much more cooperative, as the  $p(1,1)$  and  $p(0,0)$  states are observed far more than would be expected from the individual contact probabilities.

Voelz et al. applied MCOOP and other metrics to equilibrium simulations of small peptides derived from larger proteins as a first step in predicting native contacts from a protein's primary sequence (Voelz, Shell et al. 2009). Here, we find MCOOP

produces remarkable insight into our non-equilibrium unfolding simulations of large proteins, once potential biases are removed. From the simulations, we construct cooperativity networks of native contacts for  $\alpha$ LP and trypsin that differ profoundly in both their size and connectivity. Mapping the networks back on to the structure reveals both cooperative elements conserved between the two proteases and the elements that determine  $\alpha$ LP's extreme unfolding cooperativity. Importantly, our results are being used to design mutations that undermine  $\alpha$ LP's unfolding cooperativity, a goal previous mutations have failed to achieve.



**Figure 2.1: The structures of  $\alpha$ LP and trypsin.**

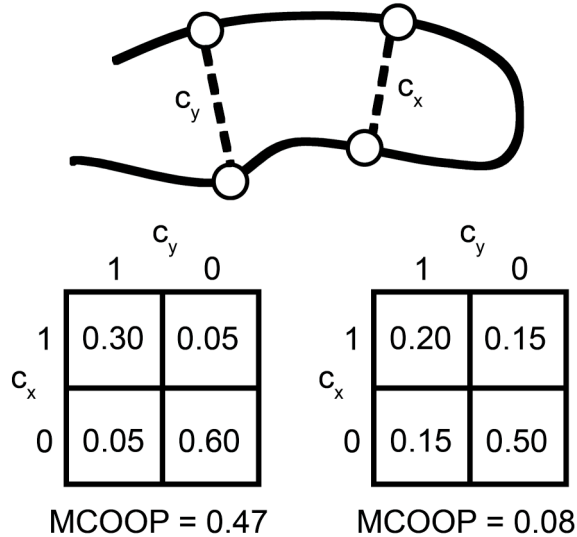
Contact maps for each protein show the contacts present in the crystal structures and formed with  $p \geq 0.8$  in 298K simulations. Cartoon renderings of molecules are colored blue at the N-terminus progressing to red at the C-terminus.

## **Results**

### *Calculating MCOOP Cooperativity*

The simulations analyzed here were performed with NAMD (Phillips, Braun et al. 2005) using the CHARMM22 force field (MacKerell, Bashford et al. 1998) with TIP3P explicit water (Jorgensen, Chandrasekhar et al. 1983) and at a temperature of 500K to achieve unfolding (Salimi 2009). There were five 8.1 ns  $\alpha$ LP simulations and four 10.1 ns trypsin simulations and conformations were saved each picosecond, resulting in a total of 40500 and 40400  $\alpha$ LP and trypsin conformations, respectively. Approximately 22% and 19% of the  $\alpha$ LP and trypsin conformations, respectively, occur before the TSE (Salimi 2009).

Though MCOOP can be calculated for any pair of contacts, here, we restrict ourselves to only native contacts for two reasons: 1) the number of possible contacts for a protein the size of  $\alpha$ LP or trypsin is on the order of  $10^4$ , resulting in  $10^8$  potential contact pairs, an increase of about three orders of magnitude over native contact pairs, and 2) the simulations are non-equilibrium unfolding simulations started from the native state, making native contacts much more relevant. Native contacts were defined on the residue level using atom-atom distance criteria (see Methods). Contacts not reliably formed in 298K simulations (Salimi 2009) were filtered from the native contacts (see Methods), as were contacts between disulfide bonded residues, which cannot be broken in the simulations. This resulted in 678 native contacts for both  $\alpha$ LP and trypsin, and contact maps for both are shown in Figure 1.



**Figure 2.2: MCOOP explained**

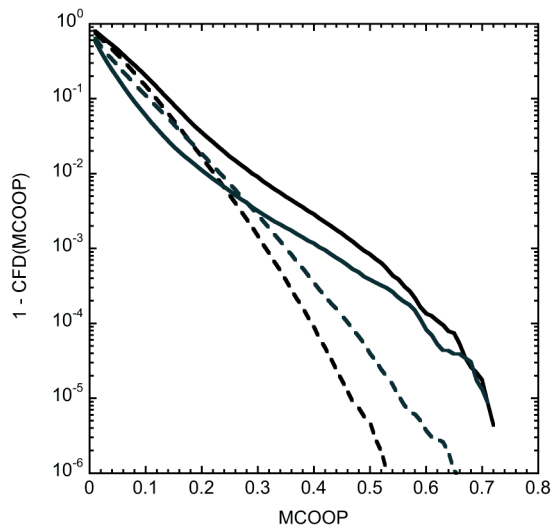
For hypothetical contacts  $c_x$  and  $c_y$ , there are four possible contact states: both formed (1,1),  $c_x$  formed and  $c_y$  broken (1,0),  $c_x$  broken and  $c_y$  formed (0,1), and both broken (0,0). Each box contains an example where  $c_x$  and  $c_y$  are each formed with  $p = 0.35$ , however, the left box has a much more cooperative distribution of the four contact states compared to the right box. MCOOP values for each distribution are calculated using Equation 1.

Using all simulation conformations, MCOOP was calculated for each pair of native contacts for both  $\alpha$ LP and trypsin. The resulting distributions were approximately exponential, with the majority of MCOOP values below 0.05 and 0.02 for  $\alpha$ LP and trypsin, respectively. Figure 3 shows 1 - the cumulative frequency distributions (CFD) for  $\alpha$ LP and trypsin in solid black and gray lines, respectively. Strikingly, the  $\alpha$ LP curve remains significantly above trypsin curve until very high values of MCOOP, indicating a marked increase in the cooperativity of  $\alpha$ LP unfolding at the contact level relative to trypsin.

While Voelz et al. empirically chose 0.3 bits as a threshold for defining a cooperative contact pair in their equilibrium simulations (Voelz, Shell et al. 2009), it may

not be appropriate for this work due to the nature of our simulations. Here, we define MCOOP thresholds based upon null MCOOP distributions in order to perform rigorous statistical testing and correct for differences in both the number and length of simulations used in the analysis. The null distributions are computed using a non-parametric permutation methodology (see Methods) and can then be compared to the observed distributions. The dashed black and gray lines in Figure 3 are these null distributions for  $\alpha$ LP and trypsin, respectively. At high MCOOP, the null CFDs are shifted heavily to the left, indicating that the cooperativities observed from combining multiple simulations are not just due to unfolding from the native state. Importantly, the difference between the observed and the null CFDs is much greater for  $\alpha$ LP than for trypsin, a further indication of its more cooperative unfolding.

Because the null CFD is heavily dependent on the number of simulations used in its generation, directly comparing the null  $\alpha$ LP and trypsin CFDs is difficult. To better compare the proteins, we also calculated observed and null CFDs for each of the five combinations of four  $\alpha$ LP simulations and observed a rightward shift of the null CFDs to approximately overlay with the trypsin null CFD (Figure S1A). However, the observed CFDs also moved generally rightward, though not to the same extent as the null CFDs. As can be seen in the ratio of observed to null CFDs (Figure S1B), trypsin unfolds far less cooperatively than all the combinations of four or five  $\alpha$ LP simulations.



**Figure 2.3: The distribution of MCOOP values in  $\alpha$ LP and trypsin.**

The y-axis is 1 - Cumulative Frequency Distribution (CFD) of the MCOOP values, ranging from 1 at MCOOP = 0 to 0 at MCOOP = 1. The observed distributions for  $\alpha$ LP (solid black line) and trypsin (solid gray line) are approximately exponential, with only a small fraction of contact pairs having large MCOOP values. Null distributions, created through a randomization method (see Methods) for  $\alpha$ LP (dashed black line) and trypsin (dashed gray line) are shifted considerably to the left of the observed distributions, showing that the combination of non-equilibrium simulations provides real information above a background level.

**Table 2.1: Parameters of the three MCOOP distributions and networks.**

	$\alpha$ LP	Trypsin	$\alpha$ LP <sub>early</sub>
# Contacts	678	678	678
# Pairs	229503	229503	229503
MCOOP threshold from null CFD	0.32	0.36	0.29
Observed CFD at threshold	0.00697	0.00173	0.00285
# Observed pairs at threshold	1600	398	654
# Observed contacts at threshold	362	191	253
Mean cooperative pairs per contact	8.8	4.2	5.2
Network Edges	1519	328	561
Network Nodes	266	102	148
Network Degree	11.4	6.4	7.6

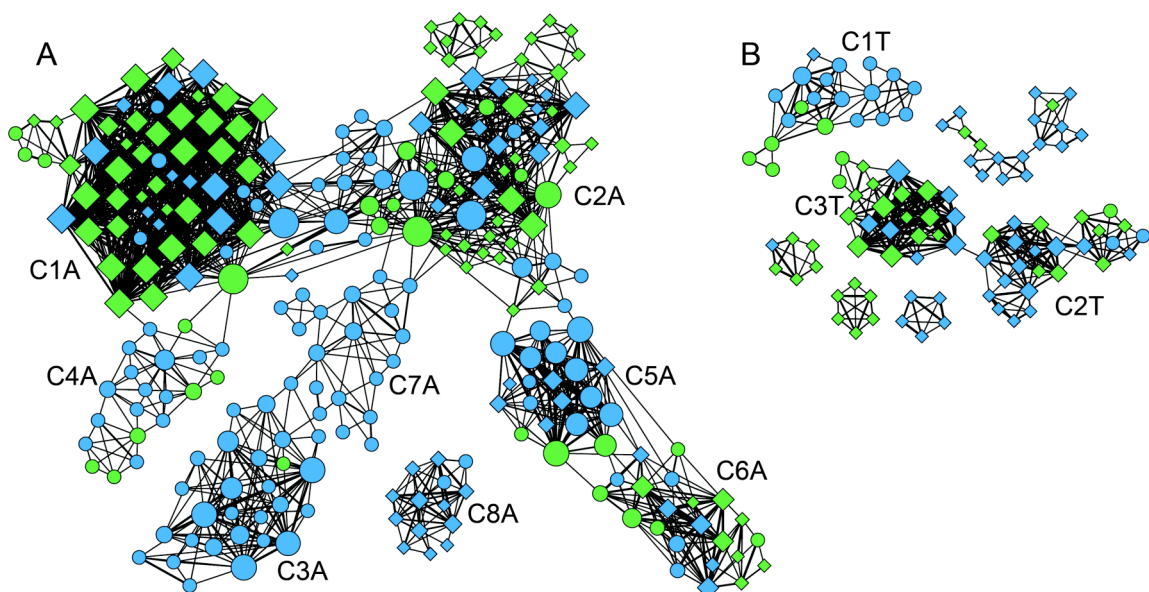
Numbers for networks represent final filtered networks as described in the Methods.



From the null CFDs, we defined MCOOP thresholds at the  $p=0.001$  level, rounded to the nearest hundredth, resulting in thresholds of 0.32 for  $\alpha$ LP and 0.36 for trypsin (Table 1). Using these thresholds, cooperativity networks were created for both  $\alpha$ LP and trypsin, where each node is a contact and the edges indicate an MCOOP value greater than or equal to the threshold value. While the vast majority of edges, 95% for  $\alpha$ LP and 82%, are found in large, high degree networks, the rest are found in tiny, sparsely connected networks often with only two or three nodes or singletons connected to the larger network. Because these weakly connected contacts contribute very little to the cooperativity distribution for the two proteins, they were filtered out (see Methods). Table 1 summarizes the network parameters after filtering.

To visualize the cooperativity networks, we used Cytoscape (Shannon, Markiel et al. 2003) to generate the 2-D network graphs in Figure 4. When viewed this way, the unfolding cooperativity difference between  $\alpha$ LP and trypsin is unmistakable. Furthermore, extra dimensions of the network can be easily visualized by manipulating node and edge colors, shapes, and sizes. First, the edge thickness increases with increasing MCOOP, with the highest MCOOP values found in densely connected subnetworks within the larger networks. The node size increases with the degree of the node, or how many other nodes to which it is connected. The maximum degree is 47 in  $\alpha$ LP and only 18 in trypsin;  $\alpha$ LP has 57 contacts with a degree of at least 18. Using node shapes, the networks separate early unfolding (circles,  $\leq 44\%$  fraction formed) from late unfolding contacts (diamonds,  $> 44\%$  fraction formed) and reveal the progression of unfolding. Not surprisingly, circles and diamonds tend to cluster together for both proteins. Intriguingly, the two largest clusters in  $\alpha$ LP, C1A and C2A, which consist

mostly of late unfolding contacts, are connected by early unfolding contacts, suggesting a clear progression of unfolding cooperativity. Most importantly, the node color denotes the conservation of the contact, green nodes conserved contacts between  $\alpha$ LP and trypsin and blue contacts non-conserved. For  $\alpha$ LP, there is a significant correlation between early unfolding and non-conserved contacts ( $p < 0.0001$ ), while unfolding timing and conservation are uncorrelated in trypsin. By constructing network graphs of  $\alpha$ LP and trypsin contact cooperativity, one can easily discern that the two homologs have evolved markedly different unfolding mechanisms.



**Figure 2.4:  $\alpha$ LP and trypsin cooperativity networks.**

Network nodes are contacts, linked to each other by edges when the MCOOP value for the pair exceeds the threshold defined by the null distributions. Edge thickness increases with MCOOP value, and node size increases with increasing degree (number of connected edges). Contacts formed  $\leq 44\%$  of the simulation are circles,  $> 44\%$  are diamonds. Green nodes are conserved contacts between  $\alpha$ LP and trypsin, blue nodes are not conserved. The  $\alpha$ LP network (A) is significantly larger and more densely connected than the trypsin network (B). Important node clusters, identified with MCODE, are identified with labels. Note that breakdown between early and late unfolding contacts as well as conserved and non-conserved contacts varies strongly amongst the clusters.

The cooperativity networks are built up of several well-defined clusters of various sizes that make relatively few connections between each other. One exception, mentioned above, is the large number of connections between clusters C1A and C2A in the  $\alpha$ LP network. We used an automated approach, MCODE (Bader and Hogue 2003), to identify clusters; the results confirm the clustering visually evident in the network graphs (Figure 4). Even though only pairwise cooperativities are computed here, because the simulations are unfolding from the native state and only native contacts are analyzed (anti-cooperative contact pairs will be rare), we observe multi-contact cooperativities as clusters in the network.

**Table 2.2: Residues making up each cluster.**

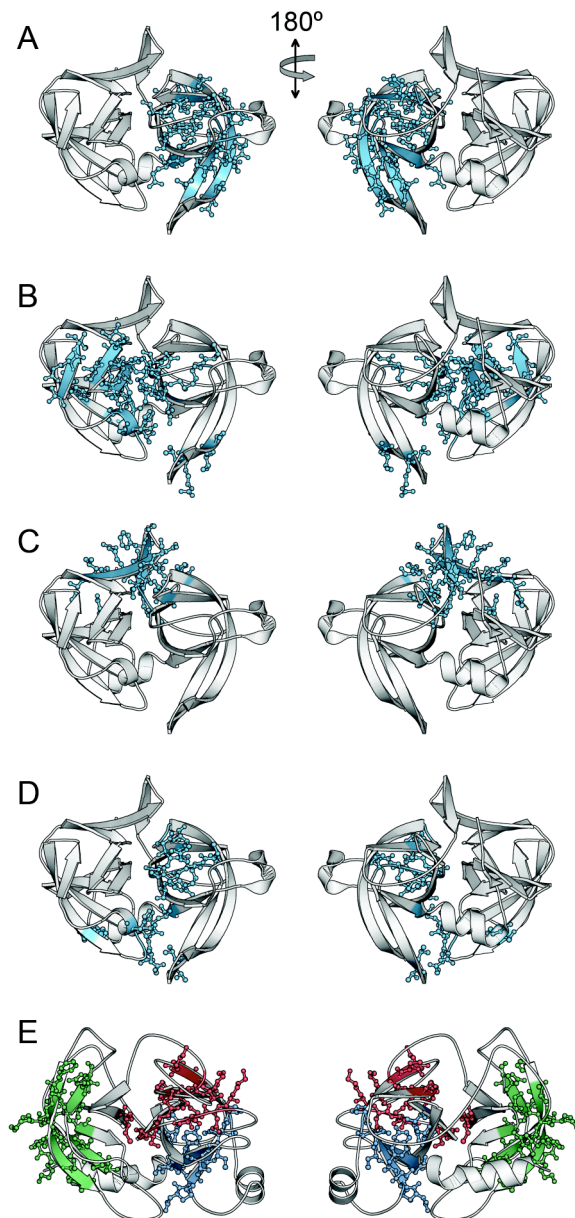
Cluster	Residues
C1A	135-138,158,160-167,179-190,198-201A,210,213,225-229
C2A	31-34,40-45,53-55,57-58,64,106,190,194-196,198,212-213
C3A	46,114-120D,120H-123,139,157,200,207,209
C4A	57,89,100-102,138,142-143,176-177,185-194,213,229,238
C5A	32-33,60-66,82-88
C6A	54,58-59,84-90,103-108
C7A	16-19,43-44A,52,120B-120C,139,198
C8A	45-48,231,235,240-243
C1T	16-18,42-44,138-140,142,156-158,188A-191,194-196
C2T	51,64-66,82-89,104-109
C3T	161-162,179-183,199,210-211,226-230
C1E	15A-31,43-44A,47,51-52,54,84,108,112-119,120A-120C,139,157,198
C2E	120-120D,120H-123,200,207-209
C3E	56-57,94-103,138,140-156,175-178,180,185-194,199,213,226,229,233
C4E	42-44A,54,193-197,212-214

Residue numbering is based on homology to chymotrypsin.

As one method of relating the cooperativity networks to protein structure, we have simply identified all residues present in a cluster and mapped them onto the corresponding structure (Figure 5 and Figure S2). Not surprisingly, contacts making up

these clusters form contiguous parts of the proteins' tertiary structures. This does not imply that only nearby contacts can form cooperative pairs; of the 1519 edges in the  $\alpha$ LP network, 42 of them are between pairs of contacts at least 12 Å apart in the crystal structure, often far apart in the primary sequence. There is only one cooperative contact pair at that distance in trypsin, the lone connection between C2T (I83-L108) and C3T (F181-Q210).

$\alpha$ LP clusters C1A, C2A, C3A, and C4A are shown in Figure 5A-D; these have a particular importance and will be discussed here, while the other clusters are shown in Figure S2. C1A (Figure 5A) encompasses 33 residues, much of the C-terminal  $\beta$ -barrel, with very high MCOOP values and late unfolding contacts (Table 2). Notably, two strands involved in enzymatic activity and substrate binding, both of which are found at the domain interface, are not represented in C1A. C2A (Figure 5B) comprises residues from both domains, especially those forming the active site at the intersection of the two  $\beta$ -barrels (Table 2). The catalytic triad (S195, H57, and D102) are all found in C2A, as are C-terminal domain residues that form the binding groove for peptide substrates. Intriguingly, several contacts between the two strands of a C-terminal domain hairpin unique to kinetically stable proteases ( $C\beta H$ ), are found in C1A, C2A, and in the area between them. Those in C2A (bottom right of Figure 5B) are found closer to the  $\beta$ -turn than those in C1A; because the strands in  $C\beta H$  separate beginning at the turn, contacts closer to the turn are more cooperative with the relatively earlier breaking C2A contacts. Both C1A and C2A consist of many conserved contacts (Figure 4A); these make up the conserved structural elements of the serine protease family and tend to unfold later, as noted above.



**Figure 2.5: Selected clusters from the cooperativity networks.**

Residues in each cluster are shown in blue ball-and-stick. For each cluster, the right panel represents a  $180^\circ$  rotation of the left panel about the vertical axis. (A) C1A is comprised of many residues making up the C-terminal domain  $\beta$ -barrel. (B) C2A spans the domain interface, along with several residues in both domains away from the interface. (C) C3A contains the Domain Bridge and several nearby residues. (D) C4A makes up part of the Domain interface and active site; it unfolds rather early. (E) Trypsin clusters C1T (red), C2T (green), and C3T (blue) are shown all on one structure because no residue overlaps between the clusters. Some similarities exist between the  $\alpha$ LP and trypsin clusters, particularly C1A and C3T.

Clusters C3A and C4A, on the other hand, contain many fewer conserved contacts and unfold relatively early (Figure 4A). C3A (Figure 5C) encompasses the Domain Bridge and several residues interacting with it (Table 2). The Domain Bridge is unique to kinetically stable proteases, hence only one contact from C3A, S46-L114, which lies just outside of it, is conserved. Unlike the C $\beta$ H, which unfolded at the turn, the Domain Bridge is most stable at the turn, with contacts furthest from the turn broken earliest. The contacts of C4A form a snaking line up the front face of the protein, clustered in two areas (Table 2). The first lies at the base of the active site, containing contacts between a *cis*-proline containing hairpin turn in the N-terminal domain (CPT) and C-terminal domain residues in the C $\beta$ H and flanking the C-terminal  $\alpha$ -helix. The breaking of these inter-domain contacts not found in trypsin is a key step in the unfolding pathway, as described previously (Salimi 2009). The other area, largely made up of contacts between T142 and T143 and residues 189-194, forms the top of the active site. Kinetically stable proteases feature Thr almost exclusively at positions 142 and 143, while nearly all metazoan proteases have Gly at 142 and a large insertion beginning at position 142, so position 143 has no equivalent (Lesk and Fordham 1996). Several of the contacts T142 makes are conserved in trypsin, but not cooperative. The side chain of T143 occupies the same site as the trypsin N-terminus and makes the same four contacts with residues 189-191 and 194; in trypsin, three of these contacts are cooperative.

The trypsin network contains three large clusters and four smaller clusters (Figure 4B), and with the exception of C2T and C3T, none of them are connected, unlike  $\alpha$ LP. Figure 5E shows the three large clusters, C1T (red), C2T (green), and C3T (blue), all mapped on to the trypsin crystal structure. Unlike many of the  $\alpha$ LP clusters, they contain

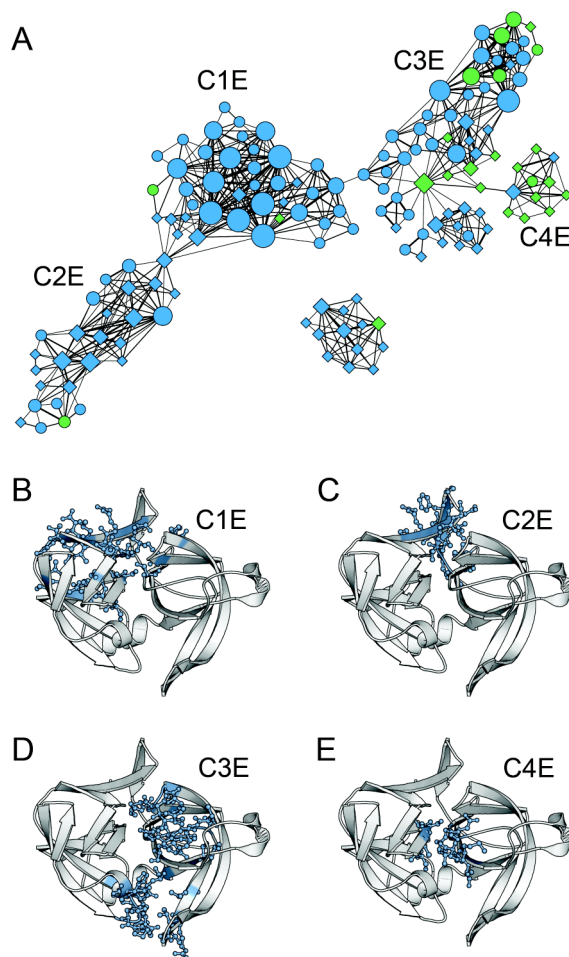
no overlapping residues and form very distinct structural clusters (Table 2). C1T differs from the other trypsin clusters in two major ways: most of its contacts are broken early, and it contains the only inter-domain contacts in the cooperativity network. The inter-domain contacts are some of the same conserved contacts found in  $\alpha$ LP cluster C2A, which form the protease active site. The majority of the contacts in C1T involve the many interactions made by the trypsin N-terminus (I16), which is critical for stabilizing the enzyme's specificity pocket (Kossiakoff, Chambers et al. 1977). C2T consists of multiple contacts between three  $\beta$ -strands and a single contact involving a fourth strand in the N-terminal  $\beta$ -barrel. Similarly, many cooperative contacts are found between two  $\beta$ -strands (residues 179-183 and 226-230) in the C-terminal  $\beta$ -barrel in C3T, with several contacts involving two more  $\beta$ -strands also included.

Due to the non-equilibrium nature of unfolding simulations, MCOOP for native contacts will be sensitive to the length of the simulation portion analyzed. If we analyzed significantly longer simulations, the fraction of time most of the native contacts would be formed would decrease to probabilities too small to have significant cooperativity (i.e., for  $p_x = 0.05$ ,  $\max \text{MCOOP} \approx 0.29$ ). In the same way, if we shorten the simulations, contacts that remain formed cannot be cooperative. In order to investigate cooperativity earlier in unfolding, around the TSE, we analyzed only conformations from the beginning of the simulation to 1 ns past the TSE. For  $\alpha$ LP, there was significant cooperativity observed, but not for trypsin (Figure S3). Table 1 summarizes the results for the early  $\alpha$ LP unfolding ( $\alpha$ LP<sub>early</sub>).

As before, a cooperativity network was generated for  $\alpha$ LP<sub>early</sub> (Figure 6A). Because many of  $\alpha$ LP's native contacts have not unfolded, the  $\alpha$ LP<sub>early</sub> network is

considerably smaller than the one in Figure 4A. Color coding remains the same as Figure 4, green contacts are conserved in trypsin, with circles representing contacts that have a formation probability of  $< 0.63$  (the fraction of simulation time prior to the TSE) and diamonds for  $\geq 0.63$ . Four main clusters are of note here, C1E-C4E (Figure 6), all of which have similarities to clusters in the main  $\alpha$ LP network (Table 2). C1E (Figure 6B) consists of many of the interactions of the N-terminal  $\beta$ -strand and those of the N-terminal part of the Domain Bridge, of which the vast majority are not found in trypsin and unfold early. C2E (Figure 6C), like C3A, contains the intra-Domain Bridge contacts and several C-terminal domain contacts, which are again not conserved in trypsin. A key contact bridging C1E and C2E is V120B/V120J, a highly conserved contact in bacterial proteases in the middle of the Domain Bridge. C3E (Figure 6D) is quite similar to C4A, with most of the contacts unfolding earlier and some of them conserved in trypsin. Finally, C4E (Figure 6E), a smaller cluster weakly connected to C3E, is made up of almost exclusively conserved and later unfolding contacts. These contacts, part of C2A in the full simulation network, mediate inter-domain contacts in the core of  $\alpha$ LP and break after the TSE and many of the other inter-domain contacts in C1E, C2E, and C3E. The two non-conserved contacts in C4E are a consequence of the divergence between bacterial and metazoan proteases. G44A/G196 connects C4E to the main network, and G44A is an inserted residue found only in kinetically stable proteases. S43/G193 only exists in bacterial proteases, where S18 is absolutely conserved, while most metazoan proteases have Gly at that position. Preliminary results indicate the  $\alpha$ LP mutant S43G does not express and most likely does not fold correctly (P. Erciyas, private communication).





**Figure 2.6: The  $\alpha\text{LP}_{\text{early}}$  cooperativity network.**

(A) The network is generated in the same manner as Figure 4, but contacts formed  $\leq 63\%$  (the fraction of simulation time prior to the TSE) of the simulation are circles,  $> 63\%$  are diamonds. (B-E) Clusters from the network, represented as in the left panels of Figure 5. (B) C1E centers around the N-terminal  $\beta$ -strand. (C) C2E contains the Domain Bridge, as in C3A. (D) C3E is similar to C4A, but contains more residues. (E) C4E comprises contacts spanning the core of the domain interface. These conserved contacts only unfold after the other domain interface contacts in the other clusters surrounding them unfold.

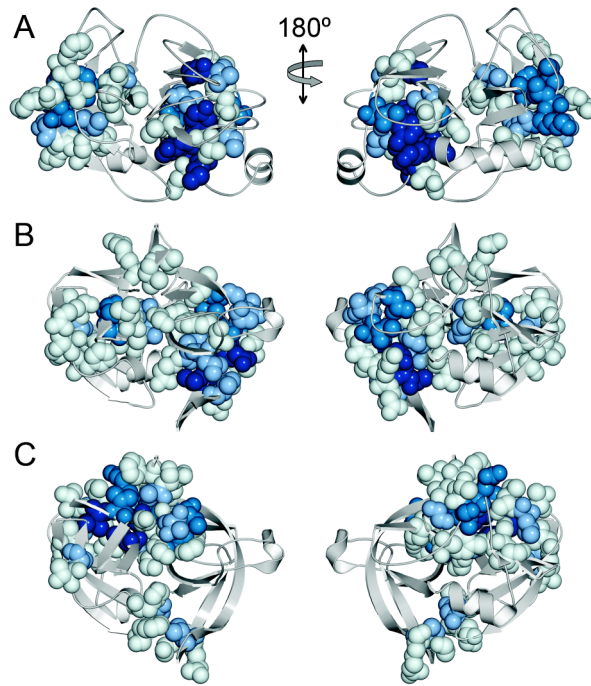
Showing every residue in a particular network cluster allows one to see the extent of the cooperativity, but also puts each residue on a level playing field. As both the number of cooperative pairs formed by a contact and the number of contacts formed by a residue vary throughout the proteins, so will the number of cooperative pairs involving any individual residue will vary. 13.6%, 14.1%, and 8.5% of the residues in  $\alpha\text{LP}$ ,

$\alpha\text{LP}_{\text{early}}$ , and trypsin, respectively, make up over 50% of the residues in cooperative pairs; hence, cooperativity is concentrated in relatively few key residues. This is especially true in trypsin, as the bottom 50% of residues contribute 0.3% of the total cooperativity, versus 6.1% in  $\alpha\text{LP}$  and 6.0% in  $\alpha\text{LP}_{\text{early}}$ . We have therefore mapped these “hot spots” for cooperativity onto the structures to see how they cluster.

Figure 7 highlights the cooperativity “hot spots” for the three networks, with residues contributing at least 0.0085% (approximately 20% of the residues) of the total cooperativity shown in space-filling and the color progressing from light blue to dark blue as the residue contributes more cooperativity. Trypsin is shown in Figure 7A. The most cooperative residues (dark blue) are all found in the C-terminal domain: I16, M180, F181, Y228, T229, K230. F181 is by far the most cooperative, making up 7.2% of the total cooperativity. For  $\alpha\text{LP}$  (Figure 7B), the most cooperative residues span the domain interface, though, as in trypsin, the C-terminal domain has more. Of the top nine most cooperative residues in  $\alpha\text{LP}$ , eight are in the C-terminal domain, and four of them, I114, L180, T181, and F228 are also in trypsin’s top nine, suggesting that the high unfolding cooperativity of the C-terminal  $\beta$ -barrel is conserved.

The distribution of cooperative residues in  $\alpha\text{LP}_{\text{early}}$  (Figure 7C) stands in marked contrast to that of either trypsin or full simulation  $\alpha\text{LP}$ . Highly cooperative residues cluster together around the Domain Bridge, N-terminus, T142 and T143, and the C $\beta$ H-CPT interactions. Intriguingly, of the eight most cooperative residues in  $\alpha\text{LP}_{\text{early}}$ , seven have no close structural equivalent in trypsin, and the other residue co-varies between bacterial proteases (aliphatic sidechains, V44 in  $\alpha\text{LP}$ ) and metazoan proteases (predominantly Gly, some Ala, G44 in trypsin). In our earlier work (Salimi 2009), we

showed that unfolding proceeded through a preferential disruption of inter-domain contacts, especially those not conserved between  $\alpha$ LP and trypsin. Figure 7C echoes that finding, and also shows that these structural elements unfold cooperatively, unlike in trypsin.



**Figure 2.7: Highly cooperative residue distributions differ between trypsin,  $\alpha$ LP, and  $\alpha$ LP<sub>early</sub>.**

Residues making up  $\geq 0.85\%$  of the total number of cooperative contacts in each network are shown in space-filling. The color goes from light blue to dark blue as the fraction of cooperative contacts increases. (A) Highly cooperative residues in trypsin comprise little of the domain interface. (B) Unlike in trypsin, many residues in  $\alpha$ LP are amongst the highly cooperative residues. Several of the dark blue residues are at equivalent positions of dark blue residues in trypsin. (C) The residue distribution of  $\alpha$ LP<sub>early</sub> shifts significantly from (B), as the most cooperative residues are cluster around the N-terminus, Domain Bridge, T142 and T144, and the C $\beta$ H-CPT region. These make up much of the domain interface that is not conserved with trypsin.

## ***Discussion***

Previously, we defined global unfolding cooperativity through a conformational similarity metric; the idea being that for a perfectly cooperative unfolding transition, conformations will be very similar and native-like until passing the TSE, when they will rapidly diverge. We found, as was previously shown experimentally (Jaswal, Sohl et al. 2002), that  $\alpha$ LP unfolds cooperatively, while trypsin does not. We attributed the difference to the domain interfaces of the two proteases; the non-conserved domain interface regions in  $\alpha$ LP protect the core from solvation better than those in trypsin because those elements are structured to only unfold cooperatively. Here, we assess cooperativity at the level of individual contacts, which allows for a direct test of that hypothesis and provides a greater insight into the mechanism of unfolding and cooperativity.

The amount and distribution of contact cooperativity also differs significantly between  $\alpha$ LP and trypsin. There are more than four times as many cooperative contact pairs in  $\alpha$ LP compared to trypsin and twice as many contacts involved in cooperative pairs. Cooperativity is also spread over more far more residues in  $\alpha$ LP compared to trypsin. The  $\alpha$ LP cooperativity distribution is consistent with its hydrogen exchange profile; over half of its amide protons have protection factors  $> 10^4$ , with 31 of those  $> 10^9$  (Jaswal, Sohl et al. 2002). For trypsin, not only is the cooperativity network much smaller, but unlike  $\alpha$ LP, there are very few connections between clusters of strong cooperativity. Each cluster is often made up of a single structural unit, such as the contacts between two adjacent  $\beta$ -strands, that unfolds cooperatively. In the  $\alpha$ LP network, these clusters are connected to others through highly connected bridging contacts, while

trypsin only has one such connection. In addition, these clusters, which represent cooperative unfolding units, are large and contain many more residues in  $\alpha$ LP compared to trypsin. In every way,  $\alpha$ LP unfolds more cooperatively than trypsin; more residues are involved in cooperative contacts, the cooperative units these residues comprise are larger, and these cooperative units are more heavily linked to other units through bridging contacts.

Though the  $\alpha$ LP and trypsin networks differ considerably at first glance, perhaps unsurprisingly, some similarities exist for the two homologs. Many of the trypsin clusters have equivalent clusters in  $\alpha$ LP, such as C3T and C1A, both of which comprise a significant portion of the C-terminal  $\beta$ -barrel. The most highly cooperative residues in each network, those making up > 50% of the cooperative contacts (27 in  $\alpha$ LP, 19 in trypsin), are much more likely to be structurally conserved between the two proteins ( $\alpha$ LP  $p < 0.0001$ , trypsin  $p = 0.026$ ). Seven of these residues are highly cooperative for both proteases and are critical residues in establishing their structures. One of these residues in  $\alpha$ LP, F228, through a distortion of its side chain, plays an important role in suppressing unfolding (Fuhrmann, Kelch et al. 2004) (B.A. Kelch submitted). The F228A mutant unfolds 16x faster than wild-type and is critical for folding, as it folds too slowly to be detected, at least two orders of magnitude slower than wild-type (B.A. Kelch submitted).

By restricting the cooperativity analysis to the early parts of unfolding, we identify the cooperative unfolding units on the way to and just past the TSE. The  $\alpha$ LP<sub>early</sub> network contains only 148 contacts out of 678 native contacts, though of the 530 not included, only 157 are due to not breaking significantly before the end of the analyzed

period. Another 105 have very weak cooperativity and were filtered from the network, leaving 268, more than half of the contacts not included, as just not significantly cooperative. By applying stringent thresholds, we can identify the relatively few contacts that not only unfold early, but also unfold cooperatively with many other contacts, not independently. This is only possible because we have carried out the analysis on multiple independent simulations, which reduces the inherent bias of the non-equilibrium simulations.

As mentioned previously, the highly cooperative residues of  $\alpha\text{LP}_{\text{early}}$  tend not to be conserved in trypsin, highlighting the role of divergent structural regions in unfolding to the TSE. Only seven of the 28 residues making up  $> 50\%$  of the cooperative contacts are structurally conserved in trypsin, with just three, D102 of the catalytic triad and D194 and G196 flanking the catalytic Ser, having sequence conservation across all serine proteases of the family. Within the family of kinetically stable proteases, 12 of the 28 residues have near complete sequence conservation. However, nine highly cooperative residues have little sequence conservation, two of them near the N-terminus and the rest in the Domain Bridge. If these residues are all critical to unfolding cooperativity, shouldn't they be very well conserved? For some residues, the polypeptide backbone, and not the side chain, makes the critical contacts, often hydrogen bonds between adjacent  $\beta$ -strands. This is also clearly true for the four Gly residues that are highly cooperative and conserved in bacterial proteases. For highly cooperative residues where the amino acid identity is strongly conserved (and not Gly), we argue that the side chain is playing a critical role in unfolding, which is experimentally accessible through mutagenesis. For example, we know that the Domain Bridge plays a role in determining unfolding rate

(Kelch and Agard 2007). While many of the cooperative residues in the Domain Bridge are not conserved, three (V120B, V120J, and V121) are, forming a hydrophobic cluster at its base, where its unfolding begins. Experiments eliminating these contacts as well as others, with the goal of weakening  $\alpha$ LP's unfolding cooperativity, are currently underway.

Molecular dynamics unfolding simulations produce a wealth of data that is most certainly underutilized. Many model systems for protein folding and human disease alleles have been extensively simulated, generating novel insight and hypotheses (Rutherford and Daggett ; Li and Daggett 1994; Lazaridis and Karplus 1998; Fulton, Main et al. 1999; Day and Daggett 2005; Scott, Randles et al. 2006; Oroguchi, Ikeguchi et al. 2007; Anderson and Daggett 2008; Rutherford, Alphantery et al. 2008; Rutherford and Daggett 2008; Steward, Armen et al. 2008). Recently, the Daggett Group, a pioneer in the use of these simulations, has begun an immense project they termed "Dynameomics," an attempt to characterize the native state and unfolding dynamics of common protein folds through molecular dynamics (Beck, Jonsson et al. 2008; Benson and Daggett 2008). Frankly put, there is an enormous amount of simulation data that already exists and more generated each day. We believe that analyzing contact cooperativity in these simulations, particularly in cases where protein stability or unfolding cooperativity are paramount, is an excellent approach for identifying the residues critical for maintaining the protein's structure. For protein folding model systems, this approach has the potential to rigorously define cooperative unfolding residues and structural units by simulation, for comparison to experimental data. By comparing cooperativity networks of simulated mutants, investigators may gain insight

into relative plasticities of proteins. Many alleles of proteins implicated in human genetic diseases have defects in protein stability, resulting in inactivation through premature degradation, aggregation, or insufficiency. Analyzing the cooperativity of early unfolding in wild-type and mutant forms should offer new insight into the mechanism for these defects and possibly point to new directions for treatment. Most importantly, we feel that every experiment, including molecular dynamics simulations, should be predictive, generating new hypotheses to test. The analysis of contact cooperativity adds much predictive power to unfolding simulations by revealing the critical contacts underlying a protein's stability.

## ***Materials and Methods***

### *Simulations*

Full simulation details were described previously (Salimi 2009).

### *Contacts*

Atoms pairs  $< 4.6 \text{ \AA}$  apart or atom pairs including at least one C or S atom  $< 5.4 \text{ \AA}$  apart were judged to be in contact, making their parent residues in contact. Only contacts formed in the crystal structures (1SSX for  $\alpha$ LP, 5PTP for trypsin) and at least three residues distant in the primary sequence were considered. These resulting native contacts were then filtered, removing contacts with a  $p_{\text{form}}$  from 298K simulations (Salimi 2009) of  $< 0.8$ , as these are quite weak, and contacts between two residues that are



disulfide-bonded, which cannot be broken in the simulations and hence cannot be cooperative. This resulted in 678 native contacts for analysis in both  $\alpha$ LP and trypsin.

### *MCOOP*

Contact states for every native contact pair were tabulated for every unfolding simulation conformation, 40500 conformations for  $\alpha$ LP and 40400 for trypsin. For the  $\alpha$ LP and trypsin full simulation analyses, the contact states for all conformations were summed. For the  $\alpha$ LP<sub>early</sub> and trypsin<sub>early</sub> analyses, conformations from the beginning of each simulation to one ns past the TSE (Salimi 2009) were used, 13820 conformations for  $\alpha$ LP, 11660 conformations for trypsin. MCOOP for each contact pair was calculated using Equation 1.

### *Null MCOOP distributions*

Because the unfolding simulations here are non-equilibrium in nature, there will be an inherent sampling bias, given that each simulation begins from the same starting structure. In addition, the length and number of simulations will also influence the MCOOP distribution. To correct for these factors, a non-parametric permutation method was developed to generate null distributions of MCOOP values. For each simulation except one, the list of contacts is randomly shuffled. Contact states for each pair of contacts are then tabulated based on the order of the contact list, not the identity of the contact pairs, i.e., in the vast majority of cases, adding contact states for different contact

pairs together. MCOOP is then calculated for these scrambled contact pairs. The procedure is then repeated 20 times to get an average null MCOOP distribution.

For example, a hypothetical protein has three contacts A, B, and C, and two unfolding simulations (S1 and S2) were performed. For S1, the list remains (A,B,C), and for S2 it is randomly shuffled, in this case (B,C,A). The list of pairs is thus (AB,AC,BC) for S1 and (BC,BA,CA) for S2. The contact states are tabulated for the combination of  $AB_{S1}$  and  $BC_{S2}$ ,  $AC_{S1}$  and  $BA_{S2}$ , and  $BC_{S1}$  and  $CA_{S2}$  and MCOOP is calculated.

### *Cooperativity Networks*

All contact pairs with an MCOOP value greater than the  $p=0.001$  value from the null MCOOP distributions were initially included in the cooperativity networks. The networks were then visualized with Cytoscape (Shannon, Markiel et al. 2003). A force-directed layout was used to better visualize the network, which was then adjusted manually. Singleton nodes were removed iteratively. Small networks with either fewer than six nodes or an average degree of less than 3.0 were also removed. These filters remove a relatively small fraction of network edges and preserve the highly connected main network. Clustering of the network was performed with MCODE (Bader and Hogue 2003).

### *Highly Cooperative Residues*

Each contact pair is made up of either three (with one appearing twice) or four residues. For a given network, there are four times the number of residues as there are

edges, and any given residue can participate in up to 50% of the cooperative contacts, if it appears twice in every cooperative contact pair. If the network were completely random, the fraction would be  $1/N$ , where  $N$  is the number of residues in the protein, 0.0051 for  $\alpha$ LP and 0.0045 for trypsin. This fraction of cooperative residues was computed for each residue in the three networks and shown in Figure 7. Residues with a fraction  $> 0.0085$  are shown in space-filling. Residues are colored according to the cooperativity fraction:  $> 0.0085$ , light blue;  $> 0.0150$  medium blue;  $> 0.0215$ , blue;  $> 0.0280$ , dark blue.

### ***Acknowledgments***

We would like to thank Dr. V. Voelz for introducing us to MCOOP and many helpful discussions. We would also like to thank Dr. V. Voelz and P. Erciyas for critical reading of the manuscript. N.L.S was supported by NIH NIGMS Training Grant GM-008284 and an NSF Graduate Research Fellowship.

## *Postscript*

*Can we now explore these computational hypotheses experimentally?*

By identifying the most highly cooperative residues, the ones that make the most cooperative contacts, the obvious question is whether if their elimination disrupts  $\alpha$ LP's extreme unfolding cooperativity. Previous mutants Brian Kelch made that greatly increased the unfolding rate had no effect on cooperativity, including F228A. Interestingly, F228 is highly cooperative in the simulations, however, that is only for the full simulation network. In the  $\alpha$ LP<sub>early</sub> network, it has no cooperativity, as few of its contacts unfold early on.

Pinar Erciyas is characterizing some mutants I posited to disrupt cooperativity, including Y33A, S43G, V120BA, V120JA, V121A, V167A, and L180A. There are some difficulties to be worked out. S43G apparently does not express and likely has difficulty folding. Y33A has sharply reduced proteolytic activity, which may make its characterization difficult. The others remain to be characterized, but I have high confidence that at least some of them will weaken  $\alpha$ LP's unfolding cooperativity.

### **Chapter 3: Accurately measuring the distortion of aromatic rings in crystal structures and molecular dynamics trajectories: $\alpha$ LP F228 as a case study**

#### *Preface*

This work of course began with the realization that F228 is distorted from planarity in Cynthia Fuhrmann's 0.83 Å crystal structure of  $\alpha$ LP. Both Brian Kelch and I had looked at high resolution structures in the PDB to determine how common a phenomenon it was. One problem with our methods was that it was sensitive to both out-of-plane distortion (which we wanted) and in-plane distortion (which we didn't). Given that issue, I came up with the method presented here, which only measures out-of-plane distortion. Luke Rice was helpful in showing me the linear algebra I didn't know in order to write the program code.

Some of the work presented here was written up by Brian as part of a manuscript on mutations of F228 and surrounding residues, of which I appear as the 3rd author for the methodology contributions. This chapter contains only work that I have performed, as nearly all the work in Brian's manuscript is his own. Also, data relating to the distortion of F228 during simulations is not a part of that work or other published work.

## *Synopsis*

A recent ultra-high resolution structure of  $\alpha$ -lytic protease ( $\alpha$ LP) solved to 0.83 Å uncovered what was then a novel distortion in the side chain of F228. Its phenyl ring is distorted from planarity by nearly 6°, with an estimated energetic penalty of ~4 kcal/mol. This distortion was shown to be conserved in the subfamily of pro-containing serine proteases that have the slowest unfolding. Most surprisingly, mutants alleviating the distortion actually sped  $\alpha$ LP folding to a higher degree than unfolding, indicating that the distortion was present to a greater extent in the Transition State Ensemble (TSE). Thus, it appears the role of the F228 distortion is to increase the kinetic stability of  $\alpha$ LP and extend its functional lifetime. Is this distortion unique? To answer that question, I compiled a data set of sub-Ångstrom crystal structures and calculated phenyl ring distortions. To do so, I first had to develop a novel method for isolating the out-of-plane distortion or bending from other ring distortions. Applying this method to the data set yielded several interesting conclusions: 1) the distribution of phenyl ring angles is normally distributed with a mean of 180°, 2) the  $\alpha$ LP F228 distortion is quite significant, 2.7 standard deviations from the mean, and many examples of higher distortion exist in the PDB. To examine the behavior of F228 in solution, I analyzed a 298K molecular dynamics simulation of  $\alpha$ LP, and show that the distortion is perfectly maintained. Finally, the F228 distortion persists in multiple unfolding simulations of  $\alpha$ LP until after the TSE has been passed, consistent with the previous experimental results.

## ***Background***

It is axiomatic to say that phenyl rings are planar substituents in organic molecules. The delocalized electrons of the ring participate in  $\pi$ -bonding that is far more stable when the ring is completely planar. However, new evidence from high resolution protein crystal structures provides exceptions to the rule. The 0.83 Å structure of  $\alpha$ LP provides a striking example of distortion in a Phe residue and was initially hypothesized to play a role in  $\alpha$ LP's kinetic stability (Fuhrmann, Kelch et al. 2004). F228 is conserved as an aromatic residue in the chymotrypsin family, but the rotamer it adopts is determined by the neighboring residue at position 199. This position co-varies, with aromatic residues in  $\alpha$ LP and other bacterial proteases with long Pro regions and aliphatic residues in short Pro bacterial proteases and metazoan proteases. Importantly, long Pro proteases have higher unfolding barriers than the short Pro proteases (Truhlar, Cunningham et al. 2004); hence, the possible role of F228 in kinetic stability. The  $\alpha$ LP F228 rotamer forces the phenyl ring into close contact with the C $\beta$  atom of T181, which induces ring bending of 5.8°, with an energetic penalty of ~4 kcal/mol, as measured by the C $\beta$ -C $\gamma$ -C $\zeta$  angle (Fuhrmann, Kelch et al. 2004). This distortion can be measured reliably because the diffraction data is of amazingly high quality, allowing structure refinement without the nearly ubiquitously applied stereochemical restraints.

To investigate the role of F228 in  $\alpha$ LP's kinetic stability, mutants designed to alleviate the distortion were characterized (B.A. Kelch submitted). The kinetic analysis of two key mutants, T181G and Repack (T181I/W199L/Q210I) revealed unequivocally that the distortion of F228 persists in the Transition State Ensemble (TSE) and is either more distorted or has fewer compensatory interactions in the TSE compared to the native state.

Analysis of the TFPA (Kelch and Agard 2007) and NAPase (Kelch, Eagen et al. 2007) structures revealed that the distortion of F228 was conserved in these long Pro proteases, again suggesting functional relevance. As the F228 distortion destabilizes the TSE more than the native state, it increases the barrier to unfolding and  $\alpha$ LP's kinetic stability, providing a longer functional lifetime.

As methods for obtaining sub-Ångstrom diffraction data for proteins have improved, more and more ultra-high resolution structures have been deposited into the Protein Data Bank (PDB). Since these structures will have either been refined without stereochemical restraints, as  $\alpha$ LP was, or the quality of data will overwhelm the relatively weak constraints, it is important to investigate the frequency of these phenyl ring distortions. Just prior to the publication of the  $\alpha$ LP structure, another lab published the structure of a PDZ domain to 0.73 Å resolution, having observed a distorted Phe and multiple peptide bonds, also thought to be planar, distorted by nearly 20°, though without any hypothesized functional implications.

Here, I examine distortions in Phe and Tyr residues in the PDB from crystal structures with resolutions at 1.0 Å and lower. At first, I used the C $\beta$ -C $\gamma$ -C $\zeta$  angle as a metric for the phenyl ring bending, as in (Fuhrmann, Kelch et al. 2004), but this proved to be inadequate, as in-plane ring distortions can artificially inflate the “bend” angle. Therefore, I developed a method that only measures out-of-plane distortion which is directly analogous to the more intuitive C $\beta$ -C $\gamma$ -C $\zeta$  angle. This method also defines the direction of the ring bend, confirming that phenyl rings in proteins are generally planar. However, significant distortions are not rare and observed in many proteins, including those significantly more distorted than  $\alpha$ LP F228. Finally, I have applied this method to

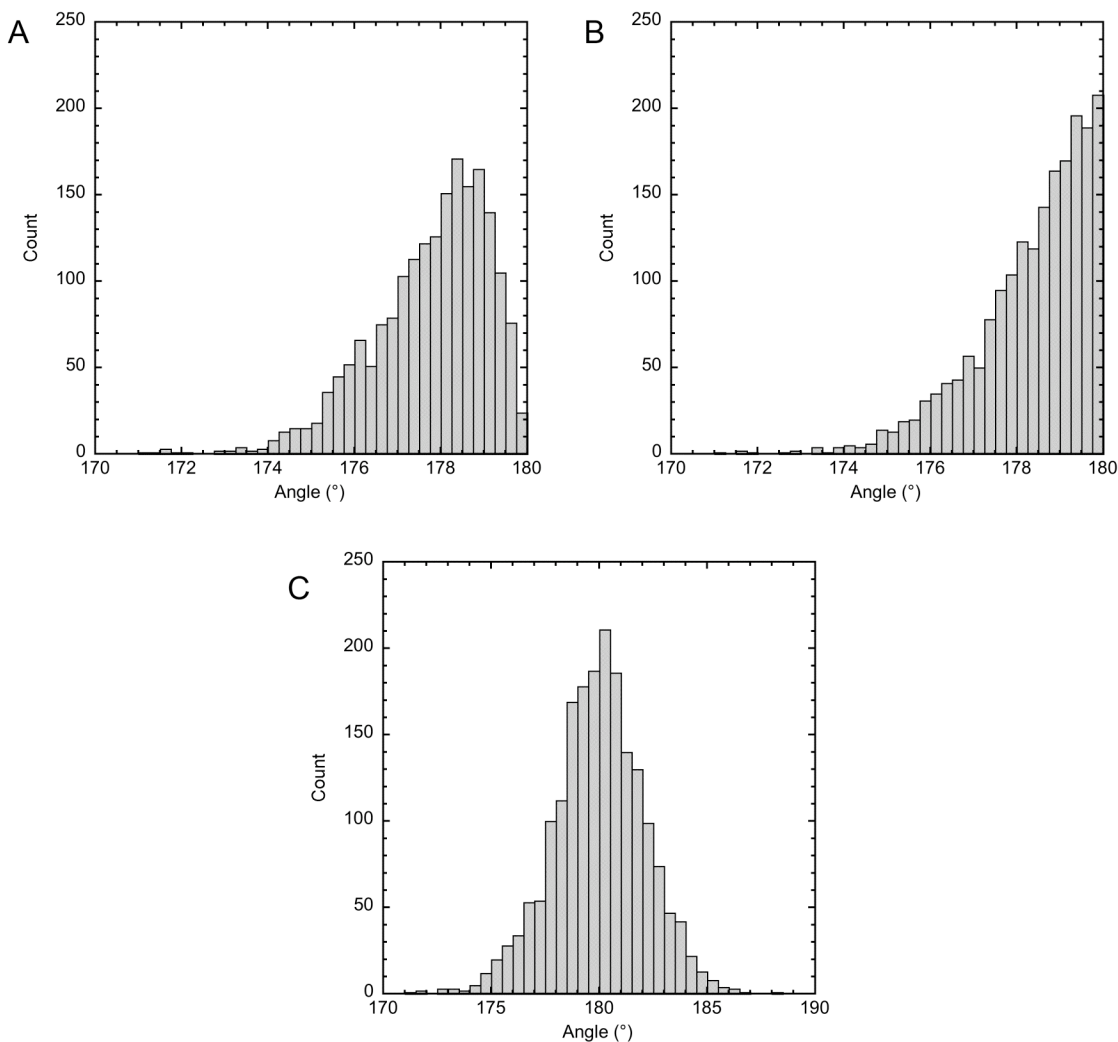


molecular dynamics trajectories of  $\alpha$ LP, both native state and high temperature unfolding, to investigate the dynamics of the distortion. Strikingly, the distortion is present in native state simulations, and is only relieved after passing the unfolding TSE, as predicted by experiment.

## ***Results***

To determine the frequency of phenyl ring distortions in proteins, I obtained all structures from the PDB solved to 1.0 Å resolution or better, which were then filtered (see Methods). The use of such high resolution structures is necessary to ensure that the atom positions are determined by the diffraction data with little influence from the stereochemical restraints. From 138 PDBs, 1944 Phe and Tyr residues were analyzed. For each residue, the ring bend can be calculated three ways. First, the  $C\beta-C\gamma-C\zeta$  angle ( $BEND_{uncor}$ ) can be calculated, as in (Fuhrmann, Kelch et al. 2004). However, this is problematic, as any in-plane distortion of that angle will be included as well. I developed a method that eliminates the in-plane distortion of the angle to provide an accurate and still intuitive measure of ring distortion. Briefly (see Methods for full details), the Phe or Tyr residue is oriented such that all out-of-plane distortion in the  $C\beta-C\gamma-C\zeta$  angle is along a single axis. Then the angle is measured with only contributions from that axis to give the corrected  $C\beta-C\gamma-C\zeta$  angle ( $BEND_{cor}$ ). Because the naming convention for Phe and Tyr residues specify different names for the  $\delta$ -carbons ( $C\delta1$  and  $C\delta2$ ), the orientation of the residue can be precisely defined, allowing the direction of the distortion to be defined. In practice, this means that the corrected and directional  $C\beta-C\gamma-C\zeta$  angle ( $BEND_{cor,dir}$ ) will still have its ideal value at  $180^\circ$ , with distortions making the angle slightly higher or

lower, depending on the orientation of C $\delta$ 1 and C $\delta$ 2. As the names of the  $\delta$ -carbons should be given randomly, the distribution of BEND<sub>cor,dir</sub> in Phe and Tyr in the PDB should center around 180°.



**Figure 3.1: The distribution of Phe and Tyr bend angles in ultra-high resolution structures.**

(A) BEND<sub>uncor</sub> (bin size 0.25°) (B) BEND<sub>cor</sub> (bin size 0.25°) and (C) BEND<sub>cor,dir</sub> (bin size 0.5°) C $\beta$ -C $\gamma$ -C $\zeta$  angles. Because the direction is not defined for (A) and (B), the maximum is 180°. Note how few angles are in the 180° bin in (A) compared to (B); this is due to in-plane distortion.

The distributions of Phe and Tyr bend angles for each of the three methodologies,  $BEND_{uncor}$ ,  $BEND_{cor}$ , and  $BEND_{cor,dir}$ , are shown in Figure 1. The difference between  $BEND_{uncor}$  and  $BEND_{cor}$  clearly shows the effect of in-plane distortion in the  $C\beta-C\gamma-C\zeta$  angle, with an increase from 18% ( $BEND_{uncor}$ ) to 39% ( $BEND_{cor}$ ) of all rings within  $1^\circ$  of  $180^\circ$ . The number of large distortions from planarity are also reduced, as the number of angles greater than  $4^\circ$  from  $180^\circ$  drops to 6.6% from 11.4%. The  $BEND_{cor,dir}$  distribution is consistent with a normal distribution centered about  $180^\circ$ , as expected. Statistics for the three distributions are shown in Table 1.

**Table 3.1: Statistics for the distribution of bend angles**

	$BEND_{uncor}$	$BEND_{cor}$	$BEND_{cor,dir}$
min	171.16	171.20	171.20
max	179.99	180.00	188.26
N	1944	1944	1944
mean	177.76	178.35	179.97
median	178.02	178.69	180.02
std. dev.	1.35	1.34	2.13

All rows, with the exception of N, are in units  $^\circ$ .

For the  $\alpha$ LP F228, its  $BEND_{cor,dir}$  is  $185.72^\circ$ , or 2.7 standard deviations away from the mean, a significant distortion from planarity. This distortion is amongst the highest in the data set, yet there are 22 other residues from 17 different proteins with larger distortions. While evaluating the model quality for these high resolution structures is outside the scope of this work, the appearance of large phenyl ring bends in this many structures argues that they are a real and relatively uncommon feature of protein structures.

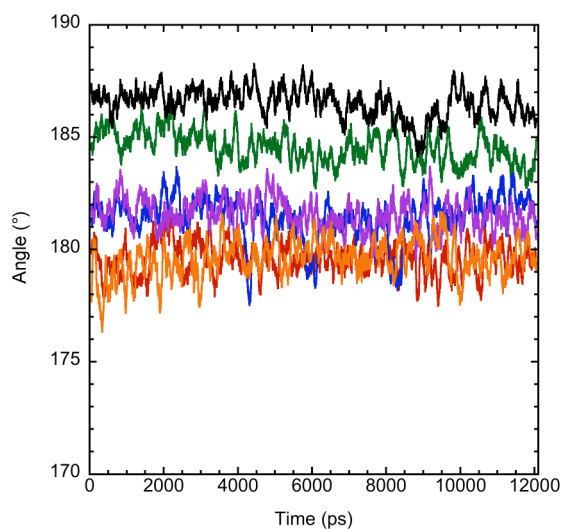
A drawback to this PDB analysis is that the state of the protein in the crystal may not resemble the protein in solution: contacts from the crystal lattice may impose

additional distortion onto phenyl rings and the cryogenic temperatures used to improve data collection may mask breathing motions which alleviate distortion at physiological temperature. One approach for investigating distortions under more physiological conditions is using molecular dynamics (MD) to simulate the protein of interest. While the physical models underlying MD have their own drawbacks, confirmation of significant phenyl ring bending (or relative planarity) by MD would add strong evidence to the results from crystallography.

Here, I have taken advantage of prior simulations of  $\alpha$ LP to investigate distortion in the six Phe residues, particularly F228, in solution. At 298K, the  $\alpha$ LP structure changes little, averaging  $< 1.0 \text{ \AA}$  C $\alpha$  RMSD to the crystal structure (Chapter 1). Somewhat surprisingly, ring bends, even modest ones, observed in the crystal structure are also seen for three Phe residues, F52, F94, and F228 (Figure 2, Table 2). Importantly, the large distortion in F228 is maintained throughout the course of the simulation. The standard deviations are quite high because the instantaneous  $\text{BEND}_{\text{cor,dir}}$  is measured from simulations with 1 fs timesteps, and snapshots saved every 1 ps. The diffraction data, on the other hand, is derived from on the order  $10^{15}$  molecules and averaged over many seconds. Future studies should attempt to save coordinates much more frequently, i.e., every 10 fs.

For several reasons, I believe that despite the high standard deviations, the simulation data is reliable. First, of the 15 pairwise t-tests between the six  $\text{BEND}_{\text{cor,dir}}$  distributions, only two produce a p value  $> 0.0001$  (F45-F120I: 0.66, F88-F94: 0.021). Second, the mean values for  $\text{BEND}_{\text{cor,dir}}$  are robust with respect to multiple parts of the simulations and unlikely due to starting conditions. Third, the behavior of F45 and F120I,

which both deviate little from  $180^\circ$ , is indicative of phenyl rings which are not subjected to forces inducing ring bending, acting as pseudo-negative controls. F85 serves this purpose particularly well, as it is considerably exposed to solvent and should rotate freely. Finally, because  $\text{BEND}_{\text{cor,dir}}$  is directional, if a distorted ring performs a ring flip, it will be observed in the trace as a sharp change in  $\text{BEND}_{\text{cor,dir}}$  to the opposite side of  $180^\circ$  with the same magnitude distortion. Several clear examples of ring flips occur in the F54 trace, and these were confirmed by examining the confirmation. Ring flips are relatively rare for the buried Phe residues in the 298K simulation, which is expected given the relatively slow rates of ring flipping from experiment (Nall and Zuniga 1990).



**Figure 3.2: The distortion of Phe rings in  $\alpha\text{LP}$  at room temperature.**

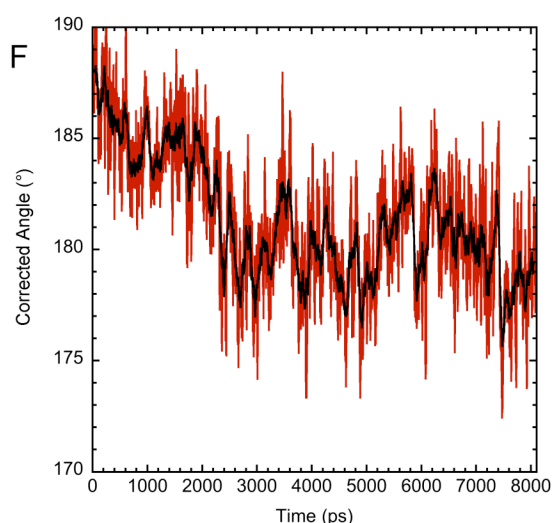
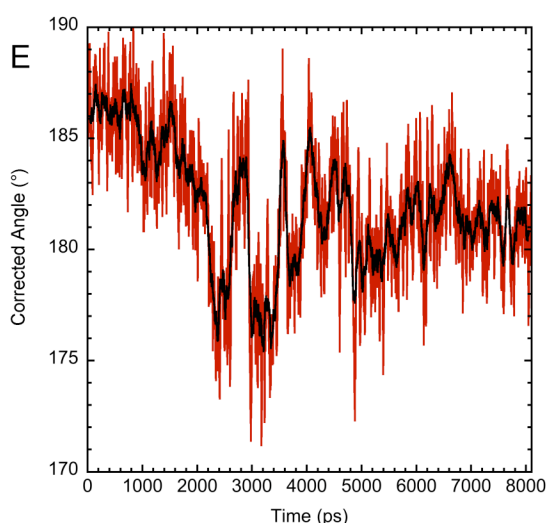
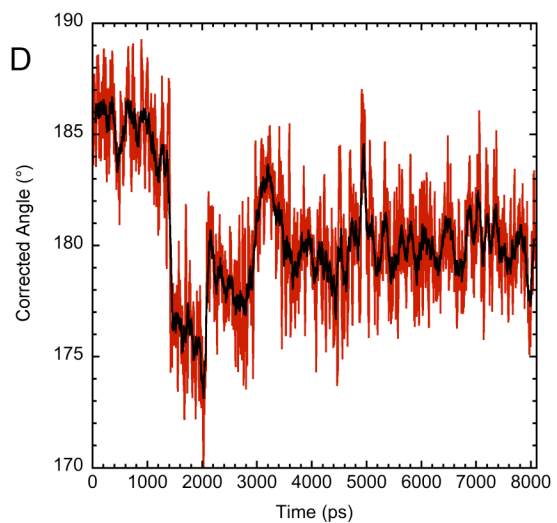
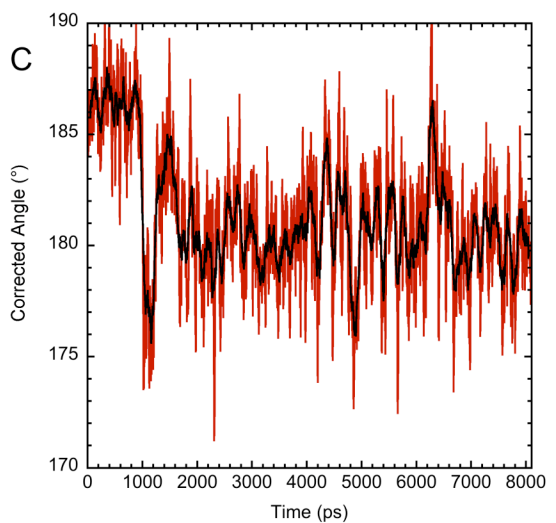
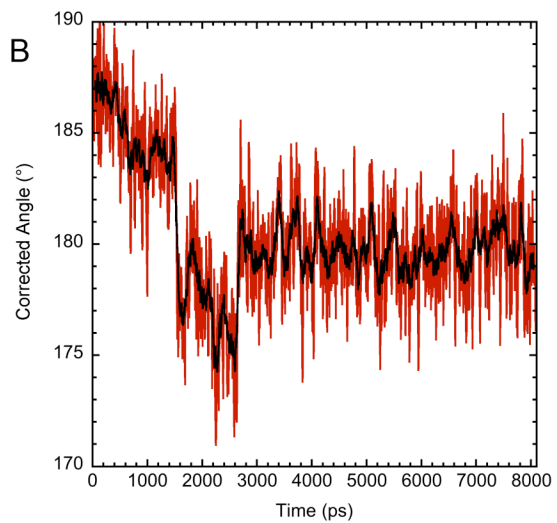
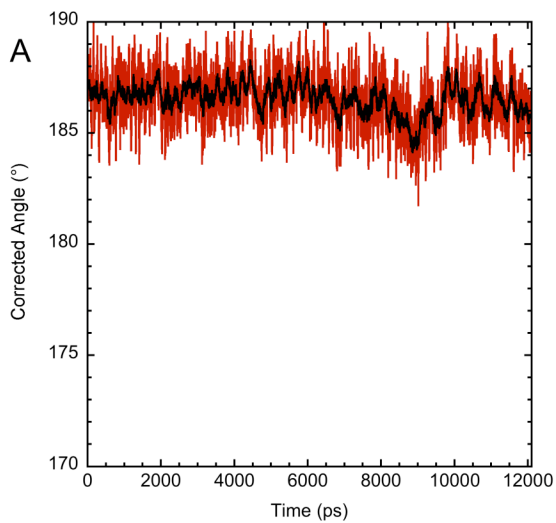
Traces are  $\text{BEND}_{\text{cor,dir}}$  for each residue with a 99 ps smoothing window applied. F45, red; F52, green; F88, blue; F94, purple; F120I, orange; F228, black.

**Table 3.2: Statistics for Phe ring distortions in the 298K  $\alpha$ LP simulation.**

Residue	Crystal	Mean	Std. Dev.
45	182.9	179.6	5.1
52	183.7	184.5	4.9
88	177.5	181.4	5.1
94	182.0	181.6	5.2
120I	182.4	179.6	5.3
228	185.7	186.5	4.9

Values are  $\text{BEND}_{\text{cor,dir}}$  in  $^{\circ}$ .

Conservation of the F228 distortion in slower unfolding long Pro proteases and characterization of mutants alleviating the distortion led to a model where F228 distortion is present through the TSE. I examined this model by measuring  $\text{BEND}_{\text{cor,dir}}$  in five 500K unfolding simulations. I hypothesized that the F228 ring would remain distorted until its local environment unfolded, which would allow it to relax back to minimal distortion. I previously determined the locations of the TSE in each simulation (Chapter 1) and examined the  $\text{BEND}_{\text{cor,dir}}$  traces for the timing of this relaxation. Figure 3 shows the traces for all five unfolding simulations and the 298K simulation for comparison. Strikingly, the F228 distortion persists until after passing through the TSE in each simulation, consistent with the experimental results. Ring flips are far more common at high temperature, likely due to both larger protein breathing motions and increased thermal energy of the ring. An examination of ring flipping of Tyr residues in cytochrome c revealed an Arrhenius-like temperature dependence of the ring flipping rate (Nall and Zuniga 1990).



*(legend on following page)*

**Figure 3.3: F228 remains distorted until after  $\alpha$ LP has unfolded.**

Traces of  $\text{BEND}_{\text{cor,dir}}$  from 298K (A) and 500K1-500K5 (B-F) unfolding simulations show the F228 ring distortion is maintained until after the protein has unfolded. Red trace is smoothed with a 19 ps window, black trace is smoothed with a 99 ps window.

***Discussion***

Only recently was it appreciated that significant distortions of planar substituents in proteins exist, due to the increase of protein structures solved to resolutions approaching those of small molecules. The F228 distortion in  $\alpha$ LP is not only fairly large, but it has been shown to be conserved and functional. Strain induced in the native state is even higher in the transition state (B.A. Kelch submitted), which slows  $\alpha$ LP unfolding. Given the many Phe and Tyr residues that have even larger ring distortions than F228, the obvious question is: “What fraction of these distortions are also functional?” Given what we have learned about  $\alpha$ LP F228, targeting highly distorted residues that are strongly conserved should provide the best chance of uncovering additional functional distortions.

The simulation results, with their strong confirmation of both the crystallography and kinetics studies, reinforce the utility of molecular simulations in studying protein structure and function. Even 500K unfolding simulations, which stress the intentions of the forcefields built to model proteins, have been shown to provide critical insight into the folding and unfolding of multiple proteins, including  $\alpha$ LP. Further insight into how the distortion manifests itself through the unfolding process may come with performing these simulations on mutants of  $\alpha$ LP that have alleviated the distortion.



## ***Methods***

### *High resolution protein structures*

I obtained all structures from the PDB solved to 1.0 Å resolution or better in December 2007. These structures were then filtered to remove structures with > 95% identity to included structures and structures containing nucleic acids. Only protein chains with > 29 residues were included in the analysis. Each protein chain in the asymmetric unit was included. Alternative conformations of Phe and Tyr residues were excluded. After these filters were applied, 138 PDB files were included containing 1944 combined Phe and Tyr residues.

### *Calculating out-of-plane distortion*

The  $C\beta-C\gamma-C\zeta$  angle ( $BEND_{uncor}$ ) in Phe is an imperfect measure of out-of-plane ring distortion because in-plane ring distortions can also reduce the angle from its idealized value of 180 degrees. To eliminate the sideways bend due to these in-plane distortions, I developed the following protocol. For each Phe, the side-chain is translated and rotated such that  $C\beta$  is at the origin,  $C\gamma$  is on the positive z-axis,  $C\delta1(x) = C\delta2(x)$ , and  $C\delta1(y) < C\delta2(y)$ . The  $C\beta-C\gamma-C\zeta$  angle is of course unaffected by this transformation and measures the uncorrected distortion. Then  $C\zeta(y)$  is set equal to 0, eliminating the sideways bend, as all out-of-plane distortion comes from the value of  $C\zeta(x)$ . The new  $C\beta-C\gamma-C\zeta$  angle ( $BEND_{cor}$ ) measures the corrected distortion. In addition, the direction of the distortion can be defined because  $C\delta1$  and  $C\delta2$  are consistently oriented. Here, the directional ( $BEND_{cor,dir}$ ) bend angle is defined by the corrected  $C\beta-C\gamma-C\zeta$  if  $C\zeta(x) < 0$ ,

and  $360 - C\beta-C\gamma-C\zeta$  if  $C\zeta(x) > 0$ . If  $C\delta 1$  and  $C\delta 2$  are assigned randomly when the crystal structure is solved, the distribution of bend angles should center at 180 degrees, which we find to be the case.

### *Molecular Dynamics Simulations*

The simulations analyzed here are described in Chapter 1. Because the instantaneous bend angle of Phe residues varies significantly in simulations, the angles are smoothed using 19 ps and 99 ps windows.

### *Acknowledgments*

I would like to thank Brian Kelch and Cynthia Fuhrmann, who were both instrumental making this project succeed.

## Chapter 4: Conclusions and future directions

### *How does $\alpha$ LP unfold so cooperatively?*

By fully suppressing partial unfolding,  $\alpha$ LP extends its functional lifetime in harsh environments (Jaswal, Sohl et al. 2002). The domain interface has long been thought to be critical in establishing this cooperativity (Cunningham and Agard 2003; Jaswal, Truhlar et al. 2005), with more recent experiments showing specific structural regions playing an important role in unfolding (Truhlar and Agard 2005; Kelch and Agard 2007; Kelch, Eagen et al. 2007). However, the mechanism of cooperativity was elusive; no perturbations had been observed to disrupt cooperativity. Hence, we sought a more global description of  $\alpha$ LP unfolding to inform future experiments.

In this work, I have applied molecular dynamics simulations to answer how  $\alpha$ LP unfolds, with the goal of discovering the mechanism of cooperativity. I have shown that simulated  $\alpha$ LP unfolding is robust, allowing the determination of the same transition state ensemble (TSE) through multiple methods. This TSE is consistent with prior experimental data: it is highly native-like, with the same fraction of surface area exposed to solvent as seen experimentally, it features early unfolding at two key interfaces already shown to play a role in unfolding, and it shows a marked disruption of the domain interface, as predicted by experiment. The TSE provides insight into how the Pro region folding catalyst accelerates  $\alpha$ LP folding so significantly, as the C-terminal  $\beta$ -hairpin (C $\beta$ H) to which it binds becomes distorted early in unfolding; the Pro region likely stabilizes the C $\beta$ H, forming the active site between the two domains.

To investigate cooperativity in unfolding, I had to develop a method to measure cooperativity, as it had not been done previously. From the cooperativity model, I showed unequivocally that  $\alpha$ LP unfolds cooperatively, a single all-or-none transition in each simulation. I then showed that its metazoan homolog trypsin, which does not unfold cooperatively *in vitro*, does not unfold cooperatively by simulation. It unfolds gradually, at key loops not found in kinetically stable proteases, one of which is a known proteolytic site. A critical difference between the two proteases is the domain interface regions flanking the highly conserved core; these regions have diverged significantly through evolution. In  $\alpha$ LP, these regions are made up of relatively small and well-structured cooperative units, the Domain Bridge and the C $\beta$ H/*cis*-proline turn region. These units only expose the core to solvent when they fully unfold. However, in trypsin, the analogous regions consist of much weaker interactions between less structured loops, and they allow much more solvation of the core domain interface early in unfolding.

I then used an information theoretical method first developed by Vince Voelz (Voelz, Shell et al. 2009) to quantify the degree of cooperativity between pairs of contacts. I first developed statistical tests to ensure the rigor of the method, which had not yet been done. This method again showed that  $\alpha$ LP unfolding significantly more cooperatively than trypsin, both in the magnitude of cooperativity and the fraction of the protein that participated in the cooperativity. Calculation of the total pairwise cooperativities led to the creation of a cooperativity network, as contacts that unfolded together clustered together. These clusters corresponded to contiguous regions of structure in the proteins, and the timing of their unfolding provided key insight into the unfolding pathways. The method was also applied to early stages of unfolding, so as to

explain the observed cooperativity of  $\alpha$ LP unfolding. Clustering identified regions at the N-terminus, Domain Bridge, and C $\beta$ H/*cis*-proline turn region which broke cooperatively near the TSE, which then allowed core domain interface contacts to break post-TSE. Finally, the method was used to identify per-residue contributions to the cooperativity, which were dominated by a small number of residues. The highly conserved residues that are also highly cooperative are prime targets for mutagenesis aimed at disrupting  $\alpha$ LP's unfolding cooperativity.

### ***Future directions – computation***

Though my work answers many questions about the  $\alpha$ LP unfolding pathway, it has by no means answered them all. And like any set of good experiments, it asks new questions and suggests new experiments. For some of the questions, computation will prove to be an asset.

Some questions are simpler to answer, such as, “How robust is the  $\alpha$ LP unfolding pathway?” I have shown that it is quite robust using five simulations, a number that the Daggett Lab has shown produces nearly the same quality of results as 100 simulations for CI2 (Day and Daggett 2005). However,  $\alpha$ LP is a much larger protein, and a more extensive set of simulations may prove useful. This is especially true with respect to the cooperativity analysis, which clearly benefits from additional simulations. Computer time is much more available and the computers are much faster now compared to the early stages of these studies, which should allow relatively large amounts of new simulation data to be collected. As of this writing, additional 500K  $\alpha$ LP simulations are underway.

MD simulations of mutants, both previously characterized and predicted, may now prove to be more useful. Though it can be difficult to characterize small changes in behavior from the mutations, advances presented here such as the cooperativity analysis provide highly quantitative results that can be tested statistically. Some mutants will be inaccessible to biochemical characterization, and hence computational study may be the only option, such as for S18G, which should have serious consequences for cooperativity and appears to not express well in culture.

There are several other possible simulation experiments that come to mind. Though I made several attempts to simulate  $\alpha$ LP unfolding while bound to the Pro region, the results have not been satisfactory, due to the restraints used to maintain the Pro region structure. Future attempts should further weaken these restraints, or apply them in such a way that allows more Pro flexibility. These simulations have the potential to better characterize how Pro accelerates  $\alpha$ LP unfolding by nine orders of magnitude. Low pH strongly accelerates the rate of  $\alpha$ LP unfolding, while weakly affecting the unfolding rate for NAPase. MD simulations at low pH may be used to test the experimental hypothesis that salt bridges across  $\alpha$ LP's domain interface are to blame.

### ***Future directions – biochemistry***

If the computational work I have presented here is sound, then it should do more than explain the results of others. Without predictive power, the value of this data is much lower. Fortunately, my work leads to many new questions that can only be answered with the right biochemical experiments. Some of these experiments are being

performed by Pinar Erciyas, though others will have to be taken up by future  $\alpha$ LP biochemists.

My work has shown that the Domain Bridge and the C $\beta$ H-CPT interactions are critical early steps in  $\alpha$ LP unfolding, consistent with experiment. However, further characterization of these regions is still warranted, either through  $\phi$ -analysis or  $\psi$ -analysis, the latter being preferable. I expect to see a decrease in the  $\alpha$ LP unfolding rate with increasing  $[\text{Ni}^{2+}]$  with properly placed double His mutants in these areas. These are also the areas which contribute to cooperativity as seen in the  $\alpha$ LP<sub>early</sub> network. Mutations of several residues, namely I16, V120B, V120J, V121, V167, and L180, that remove critical interactions in these regions have the best chance at disrupting  $\alpha$ LP's extreme unfolding cooperativity. Pinar Erciyas is currently characterizing several of these mutants, so results should be soon in coming. Several other residues, which are highly cooperative in the full simulation  $\alpha$ LP network, may also prove interesting from a mutational standpoint. F228 has already been mutated to Ala by Brian Kelch, who showed it folds too slowly to be measured and unfolds about 15 times faster than WT. He also mutated T181 to Ala and Gly. Interpretation of these mutants with regard to cooperativity is difficult given the distortion, so mutations of other residues, such as Y33, V136, I162, L227, and E229 should provide insight. At this point, Pinar Erciyas has begun characterizing Y33A, which appears to have significantly weakened proteolytic activity. E229 is also part of an inter-domain salt bridge, and Pinar Erciyas has shown that its elimination (R103A/E229Q) makes  $\alpha$ LP unfold much faster. Single mutants of this salt bridge may help sort out the salt bridge's role.

## References

- Anderson, P. C. and V. Daggett (2008). "Molecular basis for the structural instability of human DJ-1 induced by the L166P mutation associated with Parkinson's disease." Biochemistry **47**(36): 9380-9393.
- Bader, G. D. and C. W. Hogue (2003). "An automated method for finding molecular complexes in large protein interaction networks." Bmc Bioinformatics **4**: 27.
- Beck, D. A., A. L. Jonsson, et al. (2008). "Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations." Protein Eng Des Sel **21**(6): 353-68.
- Benson, N. C. and V. Daggett (2008). "Dynameomics: large-scale assessment of native protein flexibility." Protein Sci. **17**(12): 2038-50.
- Berendsen, H. J. C., J. P. M. Postma, et al. (1984). "Molecular-dynamics with coupling to an external bath." Journal of Chemical Physics **81**(8): 3684-3690.
- Bochkareva, E. S., N. M. Lissin, et al. (1992). "POSITIVE COOPERATIVITY IN THE FUNCTIONING OF MOLECULAR CHAPERONE GROEL." Journal of Biological Chemistry **267**(10): 6796-6800.
- Brunger, A. T. (1992). X-PLOR Version 3.1, A System for X-ray Crystallography and NMR, The Howard Hughes Medical Institute and Department of Biochemistry and Biophysics, Yale University.
- Chan, H. S. and K. A. Dill (1991). "Polymer Principles in Protein Structure and Stability." Annual Review of Biophysics and Biophysical Chemistry **20**(1): 447-490.



- Cunningham, E. L. and D. A. Agard (2003). "Interdependent folding of the N- and C-terminal domains defines the cooperative folding of alpha-lytic protease." Biochemistry **42**(45): 13212-9.
- Cunningham, E. L. and D. A. Agard (2004). "Disabling the folding catalyst is the last critical step in alpha-lytic protease folding." Protein Sci. **13**(2): 325-31.
- D'Aquino, J. A., J. Gómez, et al. (1996). "The magnitude of the backbone conformational entropy change in protein folding." Proteins **25**(2): 143-56.
- Davis, J. H. (1996). NMR Studies of Proteins: Assignments, Dynamics and Unfolding of alpha-lytic Protease, and Solution Structure of omega-conotoxin GVIA, University of California San Francisco.
- Day, R., B. J. Bennion, et al. (2002). "Increasing temperature accelerates protein unfolding without changing the pathway of unfolding." J. Mol. Biol. **322**(1): 189-203.
- Day, R. and V. Daggett (2005). "Ensemble versus single-molecule protein unfolding." Proc. Natl. Acad. Sci. U.S.A. **102**(38): 13445-50.
- Day, R. and V. Daggett (2005). "Sensitivity of the folding/unfolding transition state ensemble of chymotrypsin inhibitor 2 to changes in temperature and solvent." Protein Sci. **14**(5): 1242-52.
- DeLano, W. L. (2002). The PyMOL Molecular Graphics System. Palo Alto, CA, DeLano Scientific.
- Dill, K. A., S. Bromberg, et al. (1995). "PRINCIPLES OF PROTEIN-FOLDING - A PERSPECTIVE FROM SIMPLE EXACT MODELS." Protein Science **4**(4): 561-602.

- Dill, K. A. and H. S. Chan (1997). "From Levinthal to pathways to funnels." Nat. Struct. Biol. **4**(1): 10-9.
- Dill, K. A., K. M. Fiebig, et al. (1993). "Cooperativity in protein-folding kinetics." Proceedings of the National Academy of Sciences of the United States of America **90**(5): 1942-1946.
- Duda, C. T. and A. Light (1982). "Refolding of bovine threonine-neochymotrypsinogen." J. Biol. Chem. **257**(16): 9866-71.
- Fersht, A. R. (2000). "Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism." Proc. Natl. Acad. Sci. U.S.A. **97**(4): 1525-9.
- Fersht, A. R., A. Matouschek, et al. (1992). "The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding." J. Mol. Biol. **224**(3): 771-82.
- Fuhrmann, C. N., B. A. Kelch, et al. (2004). "The 0.83 Å resolution crystal structure of alpha-lytic protease reveals the detailed structure of the active site and identifies a source of conformational strain." J. Mol. Biol. **338**(5): 999-1013.
- Fulton, K. F., E. R. Main, et al. (1999). "Mapping the interactions present in the transition state for unfolding/folding of FKBP12." J. Mol. Biol. **291**(2): 445-61.
- Gsponer, J. and A. Caflisch (2002). "Molecular dynamics simulations of protein folding from the transition state." Proc. Natl. Acad. Sci. U.S.A. **99**(10): 6719-24.
- Higaki, J. N. and A. Light (1986). "Independent refolding of domains in the pancreatic serine proteinases." J. Biol. Chem. **261**(23): 606-609.

- Ho, B. K. and D. A. Agard (2008). "Identification of new, well-populated amino-acid sidechain rotamers involving hydroxyl-hydrogen atoms and sulfhydryl-hydrogen atoms." BMC Struct Biol **8**: 41.
- Ho, B. K. and D. A. Agard (2009). "Probing the Flexibility of Large Conformational Changes in Protein Structures through Local Perturbations." PLoS Comput Biol **5**(4): e1000343.
- Itzhaki, L. S., D. E. Otzen, et al. (1995). "The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding." J. Mol. Biol. **254**(2): 260-88.
- Ivankov, D. N., S. O. Garbuzynskiy, et al. (2003). "Contact order revisited: influence of protein size on the folding rate." Protein Sci. **12**(9): 2057-62.
- Jaswal, S. S. (2000). Thermodynamics, Kinetics and Landscapes in alpha-lytic Protease: A Role for Pro Regions and Kinetic Stability, University of California San Francisco.
- Jaswal, S. S., J. L. Sohl, et al. (2002). "Energetic landscape of alpha-lytic protease optimizes longevity through kinetic stability." Nature **415**(6869): 343-6.
- Jaswal, S. S., S. M. Truhlar, et al. (2005). "Comprehensive analysis of protein folding activation thermodynamics reveals a universal behavior violated by kinetically stable proteases." J. Mol. Biol. **347**(2): 355-66.
- Jemth, P., R. Day, et al. (2005). "The structure of the major transition state for folding of an FF domain from experiment and simulation." J. Mol. Biol. **350**(2): 363-78.

- Jorgensen, W. L., J. Chandrasekhar, et al. (1983). "Comparison of simple potential functions for simulating liquid water." The Journal of chemical physics **79**(2): 926-935.
- Kazmirski, S. L., A. Li, et al. (1999). "Analysis methods for comparison of multiple molecular dynamics trajectories: applications to protein unfolding pathways and denatured ensembles." J. Mol. Biol. **290**(1): 283-304.
- Kelch, B. A. and D. A. Agard (2007). "Mesophile versus thermophile: insights into the structural mechanisms of kinetic stability." J. Mol. Biol. **370**(4): 784-95.
- Kelch, B. A., K. P. Eagen, et al. (2007). "Structural and mechanistic exploration of acid resistance: kinetic stability facilitates evolution of extremophilic behavior." J. Mol. Biol. **368**(3): 870-83.
- Kossiakoff, A. A., J. L. Chambers, et al. (1977). "STRUCTURE OF BOVINE TRYPSINOGEN AT 1.9 A RESOLUTION." Biochemistry **16**(4): 654-664.
- Krantz, B. A., R. S. Dothager, et al. (2004). "Discerning the structure and energy of multiple transition states in protein folding using psi-analysis." J. Mol. Biol. **337**(2): 463-75.
- Ladurner, A. G., L. S. Itzhaki, et al. (1998). "Synergy between simulation and experiment in describing the energy landscape of protein folding." Proc. Natl. Acad. Sci. U.S.A. **95**(15): 8473-8.
- Lazaridis, T. and M. Karplus (1998). ""New view" of protein folding reconciled with the old through multiple unfolding simulations." Science **278**(5345): 1928-31.
- Lesk, A. M. and W. D. Fordham (1996). "Conservation and variability in the structures of serine proteinases of the chymotrypsin family." J. Mol. Biol. **258**(3): 501-37.

- Li, A. and V. Daggett (1994). "Characterization of the transition state of protein unfolding by use of molecular dynamics: chymotrypsin inhibitor 2." Proc. Natl. Acad. Sci. U.S.A. **91**(22): 10430-4.
- Li, A. and V. Daggett (1996). "Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations." J. Mol. Biol. **257**(2): 412-29.
- MacKerell, A. D., D. Bashford, et al. (1998). "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins." J. Phys. Chem. B.
- Martínez, J. M. and L. Martínez (2003). "Packing optimization for automated generation of complex system's initial configurations for molecular dynamics and docking." J. Comput Chem **24**(7): 819-25.
- Matouschek, A., J. T. Kellis, et al. (1989). "Mapping the transition state and pathway of protein folding by protein engineering." Nature.
- Mills, F. C., M. L. Johnson, et al. (1976). "OXYGENATION-LINKED SUBUNIT INTERACTIONS IN HUMAN HEMOGLOBIN - EXPERIMENTAL STUDIES ON CONCENTRATION-DEPENDENCE OF OXYGENATION CURVES." Biochemistry **15**(24): 5350-5362.
- Nall, B. T. and E. H. Zuniga (1990). "Rates and energetics of tyrosine ring flips in yeast iso-2-cytochrome c." Biochemistry **29**(33): 7576-7584.
- Oroguchi, T., M. Ikeguchi, et al. (2007). "Unfolding Pathways of Goat  $\alpha$ -Lactalbumin as Revealed in Multiple Alignment of Molecular Dynamics Trajectories." J. Mol. Biol. **371**: 1354-64.

- Ota, N. and D. A. Agard (2001). "Enzyme specificity under dynamic control II: Principal component analysis of alpha-lytic protease using global and local solvent boundary conditions." Protein Sci. **10**(7): 1403-14.
- Park, C., S. Zhou, et al. (2007). "Energetics-based Protein Profiling on a Proteomic Scale: Identification of Proteins Resistant to ...." Journal of Molecular Biology.
- Peters, R. J., A. K. Shiau, et al. (1998). "Pro region C-terminus:protease active site interactions are critical in catalyzing the folding of alpha-lytic protease." Biochemistry **37**(35): 12058-67.
- Phillips, J. C., R. Braun, et al. (2005). "Scalable molecular dynamics with NAMD." J Comput Chem **26**(16): 1781-1802.
- Phillips, J. C., R. Braun, et al. (2005). "Scalable molecular dynamics with NAMD." J Comput Chem **26**(16): 1781-802.
- Plaxco, K. W., K. T. Simons, et al. (1998). "Contact order, transition state placement and the refolding rates of single domain proteins." J. Mol. Biol. **277**(4): 985-94.
- Rutherford, K., E. Alphantery, et al. (2008). "The V108M mutation decreases the structural stability of catechol O-methyltransferase." Biochimica Et Biophysica Acta-Proteins and Proteomics **1784**(7-8): 1098-1105.
- Rutherford, K. and V. Daggett (2008). "Four human thiopurine S-methyltransferase alleles severely affect protein structure and dynamics." Journal of Molecular Biology **379**(4): 803-814.
- Rutherford, K. J. and V. Daggett "A Hotspot of Inactivation: The A22S and V108M Polymorphisms Individually Destabilize the Active Site Structure of Catechol O-Methyltransferase." Biochemistry **0**(ja).

- Sauter, N. K., T. Mau, et al. (1998). "Structure of alpha-lytic protease complexed with its pro region." Nat. Struct. Biol. **5**(11): 945-50.
- Scott, K. A., L. G. Randles, et al. (2006). "The folding pathway of spectrin R17 from experiment and simulation: using experimentally validated MD simulations to characterize States hinted at by experiment." J. Mol. Biol. **359**(1): 159-73.
- Shannon, P., A. Markiel, et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks, Cold Spring Harbor Lab Press, Publications Dept.
- Silen, J. L., D. Frank, et al. (1989). "Analysis of prepro-alpha-lytic protease expression in Escherichia coli reveals that the pro region is required for activity." J Bacteriol **171**(3): 1320-5.
- Sohl, J. L., S. S. Jaswal, et al. (1998). "Unfolded conformations of alpha-lytic protease are more stable than its native state." Nature **395**(6704): 817-9.
- Sosnick, T. R., R. S. Dothager, et al. (2004). "Differences in the folding transition state of ubiquitin indicated by phi and psi analyses." Proc. Natl. Acad. Sci. U.S.A. **101**(50): 17377-82.
- Steward, R. E., R. S. Armen, et al. (2008). "Different disease-causing mutations in transthyretin trigger the same conformational conversion." Protein Engineering Design & Selection **21**(3): 187-195.
- Theobald, D. L. and D. S. Wuttke (2006). "THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures." Bioinformatics **22**(17): 2171-2.

- Theobald, D. L. and D. S. Wuttke (2008). "Accurate Structural Correlations from Maximum Likelihood Superpositions." PLoS Comput Biol.
- Truhlar, S. M. and D. A. Agard (2005). "The folding landscape of an alpha-lytic protease variant reveals the role of a conserved beta-hairpin in the development of kinetic stability." Proteins **61**(1): 105-14.
- Truhlar, S. M., E. L. Cunningham, et al. (2004). "The folding landscape of Streptomyces griseus protease B reveals the energetic costs and benefits associated with evolving kinetic stability." Protein Sci. **13**(2): 381-90.
- Voelz, V. A., M. S. Shell, et al. (2009). "Predicting peptide structures in native proteins from physical simulations of fragments." PLoS Comput. Biol. **5**(2): e1000281.
- Young, T. A., E. Skordalakes, et al. (2007). "Comparison of proteolytic susceptibility in phosphoglycerate kinases from yeast and E-coli: Modulation of conformational ensembles without altering structure or stability." J. Mol. Biol. **368**(5): 1438-1447.



## Appendix 1: Supplemental Material for Chapter 1

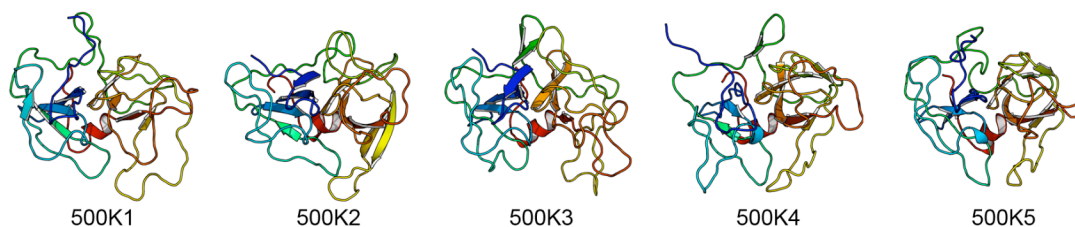
**Table A1.1: Parameter loadings for the  $\alpha$ LP Principal Components Analysis landscape.**

Parameter	PC1	PC2
$C\alpha$ RMSD	0.344	-0.061
Native Intra-Domain Contacts	-0.350	0.009
Native Inter-Domain Contacts	-0.329	-0.108
Non-Native Intra-Domain Contacts	0.323	0.280
Non-Native Inter-Domain Contacts	0.268	0.351
Radius of Gyration	0.312	-0.364
Non-Polar Solvent Accessible Surface Area	0.332	-0.267
Polar Solvent Accessible Surface Area	0.284	-0.534
Native Backbone Hydrogen Bonds	-0.346	-0.052
Non-Native Backbone Hydrogen Bonds	0.259	0.539

**Table A1.2: Selected properties of the  $\alpha$ LP crystal structure and TSE.**

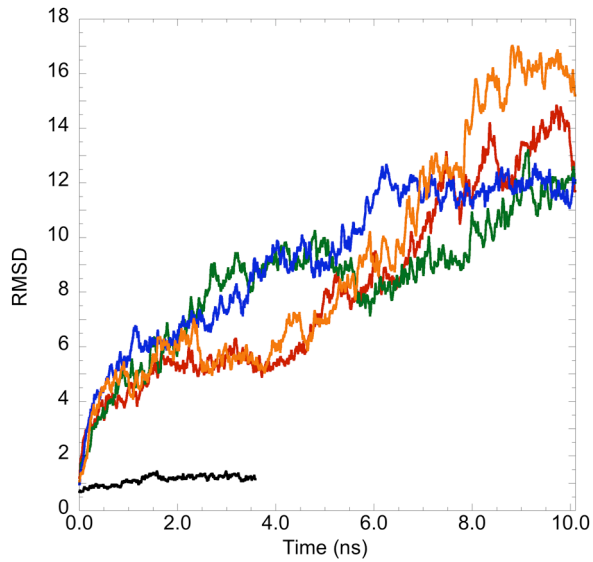
Simulation	$C\alpha$ RMSD ( $\text{\AA}$ )	NPSASA ( $\text{\AA}^2$ )	Native Contacts
Native	0.00	4005	771
500K1	$4.39 \pm 0.15$	$6110 \pm 110$	$458 \pm 6$
500K2	$4.93 \pm 0.06$	$5680 \pm 110$	$451 \pm 8$
500K3	$5.98 \pm 0.12$	$5690 \pm 170$	$462 \pm 10$
500K4	$5.04 \pm 0.11$	$6220 \pm 130$	$442 \pm 6$
500K5	$5.23 \pm 0.12$	$5820 \pm 160$	$445 \pm 9$
ALL	$5.1 \pm 0.5$	$5900 \pm 300$	$451 \pm 11$

Means  $\pm$  1 standard deviation are shown for each TSE.



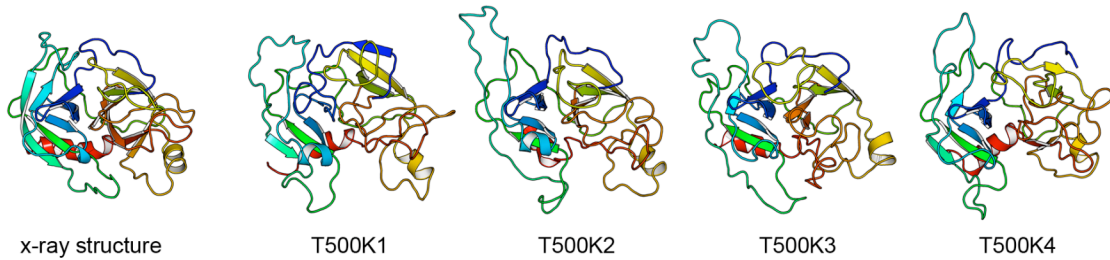
**Figure A1.1: Representative conformations of the  $\alpha$ LP TSE from each simulation show both the similarity and diversity of the TSE.**

The structures are colored blue at the N-terminus and progressing to red at the C-terminus.



**Figure A1.2: Ca RMSD for trypsin control and unfolding simulations.**

Traces correspond to black, T298K; red, T500K1; green, T500K2; blue, T500K3; orange, T500K4.



**Figure A1.3: X-ray structure of trypsin and members of its unfolding TSE from each simulation.**

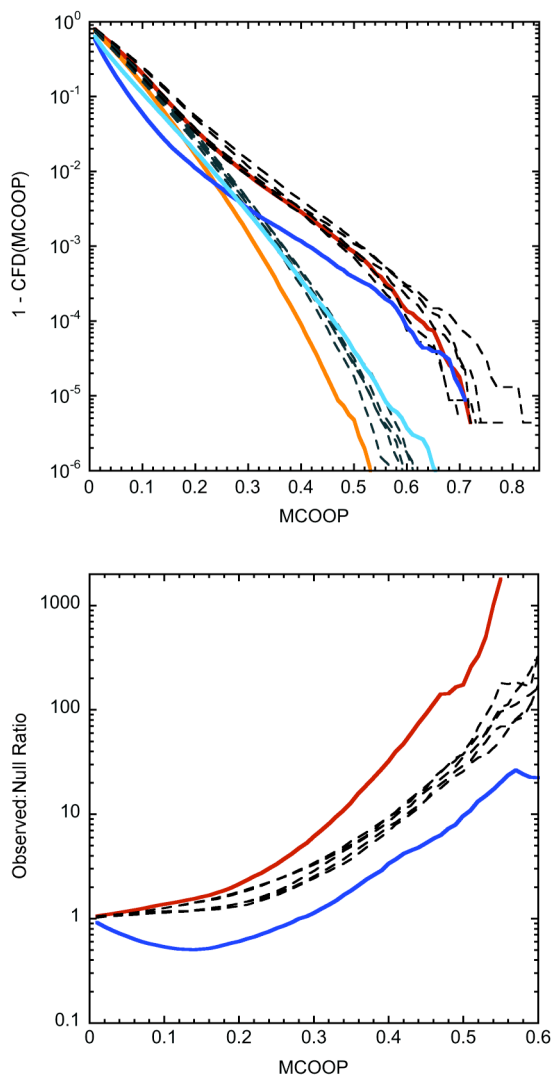
Some similarities are seen with  $\alpha$ LP, particularly the maintenance of the  $\beta$ -sheet in the N-terminal domain and the C-terminal  $\alpha$ -helix and the disruption of the domain interface both near the active site and at the “top” of the molecule as pictured.

**Table A1.3: Properties of the trypsin TSE.**

Simulation	Time at native cluster exit (ns)	TSE C $\alpha$ RMSD (Å)	TSE Fraction Native Contacts
T500K1	3.64	5.16 $\pm$ 0.12	0.515 $\pm$ 0.008
T500K2	1.41	5.11 $\pm$ 0.10	0.553 $\pm$ 0.010
T500K3	1.04	5.91 $\pm$ 0.04	0.560 $\pm$ 0.011
T500K4	1.57	5.59 $\pm$ 0.09	0.457 $\pm$ 0.012
ALL		5.4 $\pm$ 0.3	0.52 $\pm$ 0.04

The trypsin TSE was generated using the conformational clustering method due to the heterogeneity of the unfolding simulations.

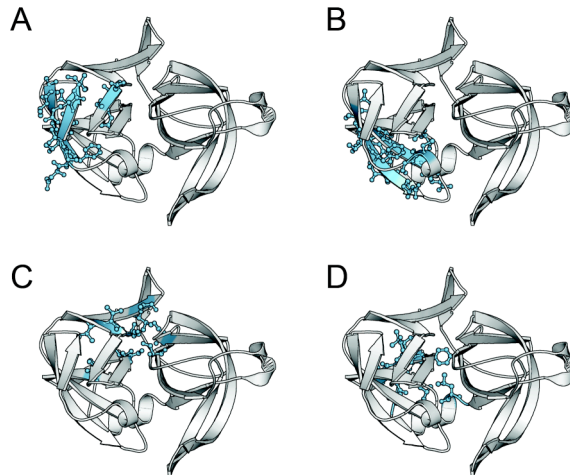
## Appendix 2: Supplemental Material for Chapter 2



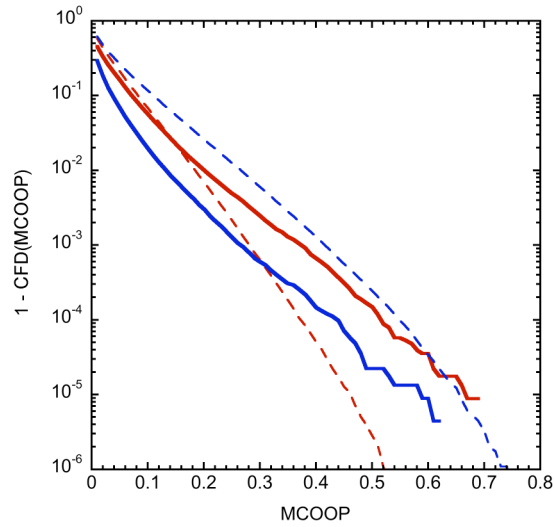
**Figure A2.1: The large cooperativity difference between  $\alpha$ LP and trypsin is not solely due to simulation number.**

(A) Distribution plot is similar to that of Figure 2.3.  $\alpha$ LP observed (red) and null distributions (orange) and trypsin observed (dark blue) and null (light blue) distributions are reproduced from Figure 2.3. MCOOP distributions of all five combinations of four  $\alpha$ LP simulations were calculated. Observed distributions for these five are shown in black dashed lines, and the corresponding null distributions as gray dashed lines. Both the observed and null distributions move to the right, as expected from decreasing the number of independent simulations used to calculate MCOOP. (B) The ratio of the observed to null curves from (A) is plotted. The higher the ratio at large values of MCOOP, the more actual information is available from the combination of simulations.

The ratio at high MCOOP values for  $\alpha$ LP (red) is substantially higher than that for trypsin (blue), an indication of how reliable the observed cooperativity is. The ratios for all five combinations of four  $\alpha$ LP simulations are also much greater than the trypsin ratio, indicating the number of simulations is not the principal factor for the weaker cooperativity observed in trypsin.



**Figure A2.2: Additional clusters identified in Figure 2.4.**  
Representation is as in Figure 2.5. (A) C5A, (B) C6A, (C) C7A, and (D) C8A.



**Figure A2.3: MCOOP distributions for abbreviated simulations.**

The  $\alpha LP_{\text{early}}$  observed (red solid line) is significantly greater than the  $\alpha LP_{\text{early}}$  null distribution (red dashed line). This is not true for trypsin<sub>early</sub> observed (blue solid line) and null (blue dashed line). Because trypsin<sub>early</sub> had such low cooperativity, and in fact worse than random, we did not pursue it further.

### Appendix 3: A new kinetically stable protease alignment

#### *Description*

Seven kinetically stable proteases from the PDB (1SSX, 2PFE, 2EA3, 2OUA, 4SGB, 1HPG, and 2SFA) were aligned manually using PyMOL. This structural alignment was used as a seed alignment for a multiple sequence alignment using ClustalW for sequence retrieved from the nr database. The alignment was further tweaked manually as necessary. It includes 51 sequences and 212 positions. 19 positions are absolutely conserved (\*), and 17 positions are almost completely conserved (+).

#### *Alignment*

```
1SSX-aLP/1-212 ANIVGGIEYSINNASLCSVGFSVTR-GATKGFVTAGHCGT
115374572/1-212 AEIIGGAAYYIGGTSRCSIGFSVT-----GGFVTAGHCGR
111068941/1-212 ATVRGGDAYLINRGGRCSVGFSVTT-----GFVTAGHCGT
145595461/1-212 YDIRGGDQFVINSRLICSVGFAVA-----GGFVTAGHCGN
395199/1-212 YDLVGGDAYYM-GGGRCSVGFSVTQ-GSTPGFATAGHCGT
2PFE-TFPA/1-212 AAIIGGNPYYF-GNYRCSIGFSVRQ-GSQTGFATAGHCGS
2EA3/1-212 FDVIGGNAYTIGGRSIRCSIGFAVN-----GGFITAGHCGR
134097115/1-212 ADVIGGDAYYIGSGSRCSVGFSVQG-----GFVTAGHCGN
126348002/1-212 AGTVGGDPYYT-GNVRCISIGFSVH-----GGFVTAGHCGG
119883589/1-212 YDIRGGDQYVIDNRLICSVGFAVA-----GGFVTAGHCGD
115374515/1-212 YDTRGGDAYYP-GNARCSIGFPVN-----GGFVTAGHCGG
115374484/1-212 YDVRGGDPCFI-GGARCTVGFSVN-----GGFITSGHCGS
145611928/1-212 VTIRGGDAYRI-GSSRCSVGFSVTT-----GFVSAGHCGN
5042248/1-212 ATVQGGDVYYINRSSRCSIGFAVTT-----GFVSAGHCGG
115376981/1-212 YDVRGGDPYYF-SNARCSIGFSVN-----GGFVTAGHCGG
134102939/1-212 YNVVGGDAYYM--GGRCSVGFSVRSSSGQAGFVTAGHCGT
1709805/1-212 ADIRGGDAYYMNGSGRCSVGFSVTR-GTQNGFATAGHCGR
395197/1-212 YDLVGGDAYYI-GNGRCSIGFSVRQ-GSTPGFVTAGHCGS
21225659/1-212 AGTVGGDPYYT-GNVRCISIGFSVH-----GGFVTAGHCGR
```

20UA-NAP/1-212	ADIIGGLAYTMGG--RCSVGFAATNASGQPGFVTAGHCGS
29827541/1-212	FDIRGGDAYYIDNTARCSVGFSVTK-GNQQGFATAGHCGR
108757299/1-212	YDLRGGDPYYF-SNYRCSVGFPVN-----GGFVTAGHCGG
117165030/1-212	EDLVGGDAYYIDGQARCSIGFSVTK-DQQQGFATAGHCGK
21219350/1-212	EDLVGGDAYYIDDQARCSIGFSVTK-DDQEGFATAGHCGD
21954476/1-212	EDLVGGDAYYIDDQARCSIGFSVTK-DDQEGFATAGHCGD
62857300/1-212	YDIRGGDAYHMGGGGRCSVGFAVTK-GTQHGAFATAGHCGR
57335304/1-212	YDIRGGDAYYMGGGGRCSVGFAVTK-GTQHGAFATAGHCGR
21223051/1-212	YDLRGGEAYYINSSRCSIGFPITK-GTQQGFATAGHCGR
21954478/1-212	YDLRGGEAYYINSSRCSIGFPITK-GTQQGFATAGHCDR
84495524/1-212	ANVYGGQQIEF-SGYVCSLGFNATK-AGAPVFITAGHCGE
145595474/1-212	--IAGGEAIWG-GGGVCSLGFNVRS-GSNYYFLTAGHCTD
28894463/1-212	--LSGGDGIHSTTGLRCSAGVNVQS-GTTYFVVTAGHCTD
21220240/1-212	--IQGGDAIYA-SSWRCSLGFNVRTSSGAEYFLTAGHCTD
29833094/1-212	--ITGGDAIYG-GGYRCSLGFNVHS-GSTYYFLTAGHCGE
4SGB-SGPB/1-212	--ISGGDAIYS-STGRCSLGFNVRS-GSTYYFLTAGHCTD
2329851/1-212	--IAGGDAITG-NGGRCSLGFNVTK-GGEPHFLTAGHCTE
474022/1-212	--IQGGDAIYA-SSWRCSLGFNVRSSSGVDYFLTAGHCTD
21220235/1-212	--VAGGDAITG-GGGRCSLGFNVTK-GGEPYFITAGHCTE
119884788/1-212	--ITGGERITGASGGTCSLGFNVRS-GSNYYFLTAGHCTD
22416397/1-212	--ISGGDAIYA-SSWRCSLGFNVQDSSGNYYFLTAGHCTD
1HPG/1-212	--VLGGGAIYG-GGSRCSAAFNVTK-GGARYFVTAGHCTN
2SFA/1-212	--IAGGEAIYAAGGGRCSLGFNVRSSSGATYALTAGHCTE
29833095/1-212	--IAGGDAITG-GGGRCSLGFNVVK-GGQPYFITAGHCTE
21219276/1-212	--ASGGDAIFG-GGARCSLGFNVTAGDGSPAFLTAGHCGV
29834039/1-212	--VSGGDAIFG-GGARCSLGFNVTAGDGSPAFLTAGHCGV
117164940/1-212	--LNGAEPIRS-TAGRCSAGFNVTG-GRSEFILTAGHCGP
117164958/1-212	--VSGGDAIFG-GGARCSLGFNVTAGDGAPAFLTAGHCGV
21219257/1-212	--LNGAEPILS-TAGRCSAGFNVTG-GTSDFILTAGHCGP
1709806/1-212	--IAGGDAIWG-SGSRCSLGFNVVK-GGEPYFLTAGHCTE
730737/1-212	--VAGGDAIYG-GGSRCSAAFNVTK-NGVRYFLTAGHCTN
1742917/1-212	--ASGGDAIFG-GGARCSLGFNVTAGDGSAAFLTRGHCGG

\*+                   \*+ ++                   +\*\*\*

1SSX-aLP/1-212	VNATARIG---GAVVGTFAARVFPGN-DRAWVSLTS-AQT
115374572/1-212	SGAAASGA---SGGAGTFAGSSFPGN-DYAWVRATS-NWT
111068941/1-212	AGAAASTTGG--ASTGTFSGSSFPGN-DYAFVIRSTS-GNT
145595461/1-212	VGEPTTGS---AAQGVIRGSSFPGD-DLAWVETNA-SWI
395199/1-212	VGTSTTGYN--QAAQGTFEESFPGD-DMAWVSVNS-DWN
2PFE-TFPA/1-212	TGTRVSS-----PSGTVAGSYFPGR-DMGWVRIITS-ADT
2EA3/1-212	TGATTAN-----PTGTFAGSSFPGN-DYAFVRTGA-GVN
134097115/1-212	QGDSTSQ-----PSGTFEGSSFPGN-DYGWVRTAS-GEN
126348002/1-212	AGAGVSGWD--RSHIGTFQGSSFPEN-DYAWVSVGS-GWW
119883589/1-212	VGEPTSGSG---VAQGTVRGSSFPGD-DYGWVQTNA-TWT
115374515/1-212	VGTNTSGSN--GVAQGTVRGSSFPTN-DYGWVQTNG-SWV
115374484/1-212	AGATVTGYN--GVVMGTVQASVFPKG-DYAWVATNS-SWT
145611928/1-212	VGTAVQTSTG--ASLGSFAGKVFPGSADMAFIRTVS-GHQ



5042248/1-212 SGASATTSSG--EALGTFSGSVFPGSADMA YVRTVS-GTV  
115376981/1-212 AGTATTGFN--GVALGTIRASTFPTN-DWG W VATNG-SWT  
134102939/1-212 RGTAVSGYN--QVAMGSFQGS SFPNN-DYAWVSVNS-NWT  
1709805/1-212 VGTTTNGVN--QQAQGT FQGSTFPGR-DIAWVATNA-NWT  
395197/1-212 VGNATTGFN--RVSQGTFRGS WFPGR-DMAWVAVNS-NWT  
21225659/1-212 AGAGVSGWD--RSYIGTFQGS SFPDN-DYAWVSVGS-GWW  
20UA-NAP/1-212 VGTQV SIG---NGRGVFERSVF PGN-DAAFVRGTS-NFT  
29827541/1-212 AGAPTAGFN--EVAQGT VQASVFP GH-DMAWVGVNS-DWT  
108757299/1-212 AGTPTTGHN--GVALGTIRGS VWPGS-DYGWVATHG-SWT  
117165030/1-212 PGATTTGFN--QADQGT FQASTFP GK-DMAWVGVNA-DWT  
21219350/1-212 PGATTTGYN--EADQGT FQASTFP GK-DMAWVGVNS-DWT  
21954476/1-212 PGATTTGYN--EADQGT FQASTFP GK-DMAWVGVNS-DWT  
62857300/1-212 VGTSTSGYN--QVAQGT FQGSTFP GR-DMAWVAANT-NWR  
57335304/1-212 VGTSTSGYN--QVAQGT FQGSTFP GR-DMAWVTANT-NWR  
21223051/1-212 AGSSTGAN--RVAQGT FQGSIFP GR-DMAWVATNS-SWT  
21954478/1-212 AGSSTGAN--RVAQGT FQGSIFP GR-DMAWVATNS-SWT  
84495524/1-212 GYQTF SKNG---TTLGKTQA FSFP GN-DYAYSTLAS-SWT  
145595474/1-212 VISNWYSNSSQTNYL GSTAGSSFP GN-DYGIVSLSG---Y  
28894463/1-212 AAPTWYTGSDATTPVGST TATSFP GN-DYGVVRYTNTAVP  
21220240/1-212 GAGAWRASSG-GTVIGQT AGSSFP GN-DYGIVQYTG-SVS  
29833094/1-212 VASTWYSNSGQTTTLG TNVSYSFP TN-DFALVRYTNTSVA  
4SGB-SGPB/1-212 GATTWWANSARTTVLGT TSGSSFP NN-DYGIVRYTNTTIP  
2329851/1-212 GISTWSDSS--GQVIGENA ASSFP GD-DYGLVKYTA-DVA  
474022/1-212 GAGTWYSNSARTTAIGS TAGSSFP GN-DYGIVRYTG-SVS  
21220235/1-212 SISTWSDSS--GNVIGENA ASSFP DN-DYGLVKYTA-DVD  
119884788/1-212 VVSSWYDN---GSL LGPTAGSSFP GD-DYGIVRLNN--GY  
22416397/1-212 GAGTWWSNSSHTTTLGT TAGSSFP GN-DYGIVRYTNSVA  
1HPG/1-212 ISANWSASSG-GSVVGV REGTSFP TN-DYGIVRYTD-GSS  
2SFA/1-212 IASTWYTNSGQTSLLG TRAGTSFP GN-DYGLIRHSN-ASA  
29833095/1-212 SISTWSDSS--GSQIGTNE QSSFP GN-DFGLVKYTS-NAD  
21219276/1-212 AADQWSDAQG-GQPIATVDQ AVFP GEGDFALVRYDDPATE  
29834039/1-212 AAAAWSDSQN-GQPIATVDQ ATFP GEGDFSLVKYDDPNTQ  
117164940/1-212 TGSVWFGDGG-GDQVGETV AGSFP GD-DFSLVEYADGKAG  
117164958/1-212 ADDQWSDAQG-GQPIATVDQ AVFP GEGDFALVRYDDPATE  
21219257/1-212 TGSVWFGDRPGDGQVGR TVAGSFP GD-DFSLVEYANGKAG  
1709806/1-212 SVTSWSDTQG-GSEIGANE GSSFP EN-DYGLVKYTS-DTA  
730737/1-212 LSSTWSSTSG-GTSIGV REGTSFP TN-DYGIVRYTT-TTN  
1742917/1-212 GATMWSDAQG-GQPIATVDQ AVFP PEGDFGLVRYDGPSTE

+ +\* \*

1SSX-aLP/1-212 LLPRVANGS--SFVTVRGSTEA AVGAAVCRSGRTTG YQCG  
115374572/1-212 STNKVA----GISSRVAGS TEAGVGASICRSGSTTG VYCG  
111068941/1-212 YQGVVNNYS-GGTIAISG STAATGASVCRSGSTTG VF CG  
145595461/1-212 PRPWVSTYD-GNVVTVTGS QEAAVGA AVCRSGRTTG WKCG  
395199/1-212 TTPTVN----EGEVTVSGS TEAAVGASICRSGSTTG WHCG  
2PFE-TFPA/1-212 VTPLVNRYN-GGTVTVTGS QEAA TGSSVCRSGATTG WR CG  
2EA3/1-212 LLAQVNNYS-GGRVQVAGHTAAPVGS AVCRSGSTTG WHCG

134097115/1-212 PVPLVNDYQ-GGTVGVAGSSEAAEGASICRSGSTTGWHCG  
126348002/1-212 TVPVVLGWGTVSDQLVRGSNEAPVGASICRSGSTTRWHCG  
119883589/1-212 PRPWVSTHD-GNVVTVTGSQEAAVGASVCRSGRTTGWRCG  
115374515/1-212 SQPWVNNYA-GGVDIVAGSNEAGVGASICRSGSTTGKRCG  
115374484/1-212 PQPWVNTYG-GGNVIVTGAQAAVVGASVCRAGPTTGWRCG  
145611928/1-212 LTGTINGYG-RGNLPSVSGSTQAGVGSSICRSGSTTGVCYCG  
5042248/1-212 LRGYINGYG-QGSFPVSGSSEAAVGASICRSGSTTQVHCG  
115376981/1-212 PQPWVYSYN-NANVTVAGSQEAGVGASICRSGYTTGWRCG  
134102939/1-212 PQPWVNLYN-GSARVVSGSSAAPVGSSICRSGSTTGWHCG  
1709805/1-212 PRPLVNGYG-RGDVTVAGSTASVVGASVCRSGSTTGWHCG  
395197/1-212 PTSLVRNSG--SGVRVTGSTQATVGSSICRSGSTTGWRCG  
21225659/1-212 TVPVVLGWGTVSDQLVRGSNVAPVGASICRSGSTTHWHCG  
20UA-NAP/1-212 LTNLVSRYNSGGYATVSGSSTAPIGSQVCRSGSTTGWYCG  
29827541/1-212 ATPDVAGAA-GQNVSIAGSVQAIVGAAICRSGSTTGWHCG  
108757299/1-212 PQPWVNNYS-GGNVTVAGSQEAPVNASICRSGYTTGWRCG  
117165030/1-212 ATPDVKAQN-DQKVQVAGSVEALVGASVCRSGSTTGWHCG  
21219350/1-212 ATPDVKAEG-GEKIQLAGSVEALVGASVCRSGSTTGWHCG  
21954476/1-212 ATPDVKAEG-GEKIQLAGSVEALVGASVCRSGSTTGWHCG  
62857300/1-212 STPYVKGAG-GQNVQVTGSTQAVVGASVCRSGSTTGWHCG  
57335304/1-212 STPYVRGAG-GQNVQVTGSTQAVVGASVCRSGSTTGWHCG  
21223051/1-212 ATPYVLGAG-GQNVQVTGSTASPVGASVCRSGSTTGWHCG  
21954478/1-212 ATPYVLGAG-GQNVQVTGSTASPVGASVCRSGSTTGWHCG  
84495524/1-212 GIGAVDLWT-GSARAVTGSSNAAVGTAICKSGRTTYWTCG  
145595474/1-212 EPGYVYLYN-GNYQDITTAGNAFVGQSVQSRGRTTGLHSG  
28894463/1-212 HPGTVG-----TVDITGTATAYVGQVCRRGATTGVRCG  
21220240/1-212 RPGTAN-----GVDITRAATPSVGTTVIRDGSTTGTHSG  
29833094/1-212 HPSAVG-----SQTISSAATPSVGTTVYRRGSTTGTHSG  
4SGB-SGPB/1-212 KDGTVG-----GQDITSAANATVGMVTRRGSTTGTHSG  
2329851/1-212 HPSQVNLYD-GSSQSIGAAEAAVGMQVTRSGSTTQVHSG  
474022/1-212 RPGTAN-----GVDITRAATPSVGTTVIRDGSTTGTHSG  
21220235/1-212 HPSEVNLYN-GSSQAISGAAEATVGMQVTRSGSTTQVHDG  
119884788/1-212 EPGYVYLYN-GGYQDITTAGNAFVGQSVRRSQTTGLHSG  
22416397/1-212 KSGAVG-----SQDITSAATPSVGTTVYRRGSTTGTHSG  
1HPG/1-212 PAGTVDLYN-GSTQDISSAANAVVGQAIKKSSTTKVTSG  
2SFA/1-212 ADGRVYLYN-GSYRDITGAGNAYVGQTVQVRSGSTTGLHSG  
29833095/1-212 HPSEVDLYN-GSTQPIKAGDATVGQKVTRSGSTTQVHSG  
21219276/1-212 APSEVDLGD--QTLPISGAAEAAVGQEVFRMGSTTGLADG  
29834039/1-212 APSEVNVGN-GQTVQISQAAEATVGQQVLRMGSTTGLNDG  
117164940/1-212 DGADVAVGDGKGV RITGLGEPAVGQVRVFRSGSTSGLRDG  
117164958/1-212 APSEVNLGD--QTVQISQAAEATVGQQVFRMGSTTGLADG  
21219257/1-212 DGADVAVGDGKGV RITGAGEPAVGQVRVFRSGSTSGLRDG  
1709806/1-212 HPSEVNLYD-GSTQAITQAGDATVGQAVTRSGSTTQVHDG  
730737/1-212 VDGRVNLYN-GGYQDIASAADAVVGQAIKKSSTTKVTSG  
1742917/1-212 APSEVDLGD--QTLPISGAAEASVQEVFRMGSTTGLADG

+ \* \* \*

1SSX-aLP/1-212 TITAKNVTANYAE-----GAVRGLTQGNACMGRGDSGGSW  
115374572/1-212 TVQAKNATVNYSQ-----GSVSGLTRTNVCAEPGDSGGSW  
111068941/1-212 TVRALGATVNYAE-----GRVTGLTQTNVCAEPGDSGGSF  
145595461/1-212 TITAKNVTVNYSY-----GPVYGMVRSTACAQPGDSGGSF  
395199/1-212 TIQQHNTSVTYPE-----GTITGVTRTSVCAEPGDSGGSY  
2PFE-TFPA/1-212 TIQSKNQTVRYAE-----GTVTGLTRTTACAEGGDSGGPW  
2EA3/1-212 TITALNSSVTYPE-----GTVRGLIRTTVCAEPGDSGGSL  
134097115/1-212 TVEAKNQTVRYPQ-----GTVEGLTRTNVCAEPGDSGGSW  
126348002/1-212 TVLAKNETVNYSQ-----GAVRQMTKTSVCAEGGDSGGSF  
119883589/1-212 TITATNVTVNYSG-----QLVHGLVRSTACAQPGDSGGPF  
115374515/1-212 SIQAKNITVNYSN-----GPVYGLTQTNVCAEPGDSGGSW  
115374484/1-212 TVLARNATVNQAQ-----GSVTGLVRTNVCAEPGDSGGPW  
145611928/1-212 TVGALGATVNYAQ-----GSVTGLTRTSVCAEPGDSGGSF  
5042248/1-212 TIGAKGATVNYPQ-----GAVSGLTRTSVCAEPGDSGGSF  
115376981/1-212 TLLAKNITVNYSN-----GPVYGMSHTNACANGGDSGGSV  
134102939/1-212 SVQALNQTVRYAE-----GTVYGLTRTNVCAEPGDSGGSF  
1709805/1-212 TIQQLNNTSVTYPE-----GTISGVTRTSVCAEPGDSGGSY  
395197/1-212 TIQQHNTSVTYPQ-----GTITGVTRTSACAQPGDSGGSF  
21225659/1-212 TVLAHNETVNYSYG-----SVVHQLTKTSVCAEGGDSGGSF  
2OUA-NAP/1-212 TIQARNQTVSYPQ-----GTVHSLTRTSVCAEPGDSAGSF  
29827541/1-212 TVEEHDTSVTYEE-----GTVDGLTRTTVCAEPGDSGGSF  
108757299/1-212 VLQAKNITVNYSV-----GPVYGLHKTNACADGGDSGGSV  
117165030/1-212 TVQQHDTSVNYAE-----GTVDGLTETTVCAPGDSGGPF  
21219350/1-212 TIQQHDTSVTYPE-----GTVDGLTETTVCAPGDSGGPF  
21954476/1-212 TIQQHDTSVTYPE-----GTVDGLTGTTVCAPGDSGGPF  
62857300/1-212 TIQQHNTSVTYPE-----GTISGVTRTTVCAEPGDSGGSY  
57335304/1-212 TIQQHNTSVTYPE-----GTISGVTRTTVCAEPGDSGGSY  
21223051/1-212 TVTQLNTSVTYQE-----GTISPVTRTTVCAEPGDSGGSF  
21954478/1-212 TVTQLNTSVTYQE-----GTISPVTRTTVCAEPGDSGGSF  
84495524/1-212 SVQAKNVTVNVDNGDGTSSVSGLTKSNTCTEGGDSGGSW  
145595474/1-212 SVTGLNATVNYSY-----GTVRGLIRTNVCAERGDSSGSL  
28894463/1-212 QVIALNATVNYGGG-----DVVSGLIQTNICAEPGDSGGPL  
21220240/1-212 RVTALNATVNYGGG-----DVVGGLIQTTVCAEPGDSGGSL  
29833094/1-212 RVTALNATVNYGSG-----DVVYGMIQTTVCAEGGDSGGPL  
4SGB-SGPB/1-212 SVTALNATVNYGGG-----DVVYGMIRTNVCAEPGDSGGPL  
2329851/1-212 TVTGLDATVNYGNG-----DIVNGLIQTDVCAEPGDSGGSL  
474022/1-212 RVTALNATVNYGGG-----DIVSGLIQTTVCAEPGDSGGPL  
21220235/1-212 TVTGLDATVNYGNG-----DIVNGLIQTDVCAEPGDSGGSL  
119884788/1-212 SVTGLNATVNYVE-----GTVYGLIRTNVCAERGDSSGSL  
22416397/1-212 RVTALNATVNYGNG-----EIVYGLIQTTVCAEPGDSGGPL  
1HPG/1-212 TVTAVNVTVNYGD-----GPVYNMVRTTACSAGGDSGGAH  
2SFA/1-212 RVTGLNATVNYGGG-----DIVSGLIQTNVCAEPGDSGGAL  
29833095/1-212 TVTGLDATVNYGNG-----DIVNGLIQTDVCAEPGDSGGSL  
21219276/1-212 QVLGLDATVNYPE-----GMVTGLIQTDVCAEPGDSGGSL  
29834039/1-212 NVTGLDATVNYPE-----GTVTGLIQTDVCAEPGDSGGSL  
117164940/1-212 RVTALDATVNYPE-----GTVTGLIETDVCAEPGDSGGPM  
117164958/1-212 QVLGLDATVNYPE-----GTVTGLIQTDVCAEPGDSGGSL

21219257/1-212	RVTALDATVNYPE-----GTVTGLIETDVCAEPGDSGGPM
1709806/1-212	EVTALDATVNNG-----DIVNGLIQTTVCAEPGDSGGAL
730737/1-212	TVSAVNVTVNYS-----GPVYGMVRTTACSAGGDSGGAH
1742917/1-212	QVLGLDVTVNYPE-----GTVTGLIQTDVCAEPGDSGGSL

+ \* \* \* \* \*

1SSX-aLP/1-212	ITSAGQAQGVMSGGNVQSNGNNC-GIPASQRSSLFERLQP
115374572/1-212	ISG-TQAQGVTSGGSG-----NCSSG-----GTTYFQPVNE
111068941/1-212	YSG-SQAQGVTSGGSG-----NCNSG-----GVTYFQPVNE
145595461/1-212	VAG-SQAQGVTSGGSG-----NCSTG-----GSTVYQPVNE
395199/1-212	ISG-SQAQGVTSGGSG-----NCTSG-----GTTYHQPINP
2PFE-TFPA/1-212	LTG-SQAQGVTSGGTG-----DCRS-----GITFFQPINP
2EA3/1-212	LAG-NQAQGVTSGGSG-----NCRT-----GTTFFQPVNP
134097115/1-212	LSG-DQAQGVTSGGSG-----DCTSG-----GTTYFQPVNE
126348002/1-212	ISG-DQAQGVTSGGWG-----NCSSG-----GETWFQPVNE
119883589/1-212	VAG-SQAQGVTSGAGG-----DCAS-----GTTVYQPVNE
115374515/1-212	LSG-NQAQGVTSGGSG-----NCTSG-----GTTFFQPINP
115374484/1-212	LSG-SQAQGMTSGGSG-----NCTSG-----GQTYFQPVQP
145611928/1-212	YSG-AQGQGVTSGGSG-----NCAS-----GTTYFQPLNR
5042248/1-212	YSG-SQAQGVTSGGSG-----DCSR-----GTTYFQPVNR
115376981/1-212	ISG-NQAQGVTSGIAG-----GCDSSN---PQTFQPINP
134102939/1-212	ISG-NQAQGMTSGGSG-----NCSSG-----GTTYFQPVNE
1709805/1-212	ISG-SQAQGVTSGGSG-----NCSSG-----GTTYFQPINP
395197/1-212	ISG-TQAQGVTSGGSG-----NCSIG-----GTTFHQPVNP
21225659/1-212	ISG-DQAQGVTSGGWG-----NCSSG-----GETWFQPVNE
20UA-NAP/1-212	ISG-TQAQGVTSGGSG-----NCRT-----GTTFYQEVNP
29827541/1-212	VSG-SQAQGVTSGGSG-----DCTRG-----GTTYQPVNP
108757299/1-212	ISG-NQAQGVTSGVAG-----TCANGNP--PQTFYQPVNP
117165030/1-212	VAG-AQAQGTTSGGSG-----DCTNG-----GTTFYQPVNP
21219350/1-212	VSG-VQAQGTTSGGSG-----DCTNG-----GTTFYQPVNP
21954476/1-212	VSG-VQAQGTTSGGSG-----DCTNG-----GTTFYQPVNP
62857300/1-212	ISG-SQAQGVTSGGSG-----DCRT-----GTTYHQPLNP
57335304/1-212	ISG-SQAQGVTSGGSG-----DCRT-----GTTYHQPLNP
21223051/1-212	ISG-SQAQGVTSGGSG-----DCRT-----GETFFQPINA
21954478/1-212	ISG-SQAQGVTSGGSG-----DCRT-----GGTFFQPINA
84495524/1-212	MAG-NLAQGVTSGGAGYGSSGVCGEKVGQPNIAFYQPVGE
145595474/1-212	FSG-STALGLTSGGSG-----NCTWG-----GTTFFQPVVE
28894463/1-212	YAG-DKIIGILSGGSG-----DCAT-----GTTFYQPIQE
21220240/1-212	YGSNGTAYGLTSGGSG-----NCSSG-----GTTFFQPVTE
29833094/1-212	YGG-SVAYGLTSGGSG-----NCTSG-----GTTFFQPVTE
4SGB-SGPB/1-212	YSG-TRAIGLTSGGSG-----NCSSG-----GTTFFQPVTE
2329851/1-212	FSG-DKAVGLTSGGSG-----DCTSG-----GTTFFQPVTE
474022/1-212	YGSNGTAYGLTSGGSG-----NCSSG-----GTTFFQPVTE
21220235/1-212	FSG-DQAIGLTSGGSG-----DCTSG-----GETFFQPVTE
119884788/1-212	FSG-STALGLTSGGNG-----NCTFG-----GTTYFQPVIE
22416397/1-212	YGG-STAYGLTSGGSG-----NCTSG-----GTTFFQPVTE
1HPG/1-212	FAG-SVALGIHSGSSG-----C-SGTA--GSAIHQPVTE

2SFA/1-212	FAG-STALGLTSGGSG-----NCRTG----GTTFFQPVTE
29833095/1-212	FAA-DTAIGLTSGGSG-----DCTSG----GETFFQPVTE
21219276/1-212	FTRDGLAIGLTSGGSG-----DCTVG----GETFFQPVTT
29834039/1-212	FTQDGS AIGLTSGGSG-----DCTVG----GETFFQPVTT
117164940/1-212	FSE-GLALGVTSGGSG-----DCAKG----GTTFFQPLPD
117164958/1-212	FTQDGLAIGLTSGGSG-----DCAVG----GETFFQPVTT
21219257/1-212	FSE-GVALGVTSGGSG-----DCAKG----GTTFFQPLPE
1709806/1-212	FAG-DTALGLTSGGSG-----DCSSG----GTTFFQPVPE
730737/1-212	FAG-SVALGIHSGSSG-----C-TGTN--GSAIHQPVRE
1742917/1-212	FTRDGLAIRLTSGGTR-----DCTSG----GETFFQPVTT
	+ + ** + * + ++

1SSX-aLP/1-212	ILSQYGLSLVTG
115374572/1-212	ILSTYGLTLTR-
111068941/1-212	ILSAYGLTLVRG
145595461/1-212	ILSRYGLSLTTS
395199/1-212	LLSAYGLDLVTG
2PFE-TFPA/1-212	LLSYFGLQLVTG
2EA3/1-212	ILQAYGLRMITT
134097115/1-212	ILQAYGLTLLTQ
126348002/1-212	ILNRYGLTLHTA
119883589/1-212	ILSRYGLSLTTS
115374515/1-212	ILSTYGLSLTTN
115374484/1-212	VLSAYGLTLKTG
145611928/1-212	ILSTYGLTLVRG
5042248/1-212	ILQTYGLTLVTA
115376981/1-212	ILSTYGLTLRTG
134102939/1-212	ALSAYGLSLVRG
1709805/1-212	LLQAYGLTLVTS
395197/1-212	ILSQYGLTLVRS
21225659/1-212	ILNRYGLTLHTA
2OUA-NAP/1-212	MLNSWNLRLRT-
29827541/1-212	ILSTYGLTLKTS
108757299/1-212	ILGAYGLTLRRT
117165030/1-212	LLSDFGLTLKTT
21219350/1-212	LLSDFGLTLKTT
21954476/1-212	LLSDFGLTLKTT
62857300/1-212	LLQAYALTLTTT
57335304/1-212	LLQAYALTLTTT
21223051/1-212	LLQNYGLTLKTT
21954478/1-212	LLQNYGLTLKTT
84495524/1-212	ILSAYGLTLKTA
145595474/1-212	ALNVYGVNVY--
28894463/1-212	VLSAYGLTVY--
21220240/1-212	ALSAYGVSVY--
29833094/1-212	ALSYYGVSVG--
4SGB-SGPB/1-212	ALSAYGVSVY--

2329851/1-212	ALSATGTQIG--
474022/1-212	ALSAYGVSIVY--
21220235/1-212	ALSATGTQIG--
119884788/1-212	ALNRYGVDVY--
22416397/1-212	ALSAYGVHVY--
1HPG/1-212	ALSAYGVTIVY--
2SFA/1-212	ALSAYGVSII--
29833095/1-212	ALSTFGAQIG--
21219276/1-212	ALAAVGATLG--
29834039/1-212	ALEAVGATLG--
117164940/1-212	AMASLGVRLI--
117164958/1-212	ALEAVGATLG--
21219257/1-212	AMASLGVRLI--
1709806/1-212	ALAAYGAEIG--
730737/1-212	ALSAYGVNVY--
1742917/1-212	ALAAVGGTIG--


+

**Publishing Agreement**

*It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.*

***Please sign the following statement:***

*I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.*

  
\_\_\_\_\_  
Author Signature

6/18/09  
\_\_\_\_\_  
Date