MECHANICAL FOLDING/UNFOLDING OF PROTEINS PROBED BY SINGLE MOLECULE ATOMIC FORCE MICROSCOPY

by

Qing Peng

M. Sc., Wuhan University, China, 2004
B. Sc., Wuhan University, China, 2001

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Chemistry)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2011

© Qing Peng, 2011
Abstract

The mechanical folding/unfolding of proteins is involved in many biological processes. However, the molecular mechanism underlying the mechanical folding/unfolding of proteins remains an open question. Most of the current knowledge about the protein folding is from ensemble measurements. In the study of the molecular mechanism underlying the mechanical folding/unfolding of proteins, single molecule atomic force microscopy (AFM) has its unique advantages. Although many endeavors have been made by using single molecule AFM to study the mechanical folding/unfolding of proteins and numerous interesting details have been revealed, the underlying mechanism of protein mechanical folding/unfolding remains largely unknown.

The main objective of this thesis is to study the mechanical folding/unfolding of some model proteins using single molecule AFM. First, we studied the mechanical unfolding pathways of two domain-insertion proteins: a natural one, T4-lysozyme (T4L), and an artificially designed one, GL5/T4L (GL5: a mutant of protein GB1). Our study on T4L provided the first direct evidence of the kinetic partitioning assumption for protein folding at the single molecule level. Our study on GL5/T4L revealed its mechanical unfolding pathway with a reversed mechanical unfolding hierarchy. The designing of domain-insertion proteins also presented a new concept to program the mechanical unfolding pathway of multi-domain proteins.

Second, we studied the mechanical folding/unfolding of TNfn3 domain by combining single molecule AFM with the steered molecular dynamics (SMD) simulation and protein engineering. The mechanical design of TNfn3 was found robust and the backbone H-bonds of TNfn3 were found critical for its mechanical stability. Our results
showed the first direct evidence that the mechanical folding pathways of TNfn3 are governed by kinetic partitioning.

Third, we studied the folding/unfolding kinetics and mechanics of an artificially designed mutually exclusive protein GL5/I27w34f (I27w34f: a tryptophan removed mutant of I27). The mutually exclusive protein GL5/I27w34f is designed to mimic the natural domain-insertion proteins which are typically difficult to study directly. Our study provided the first direct evidence that protein folding can generate sufficient mechanical strain to unravel a host protein and the folding of mutually exclusive proteins involving a tug-of-war. Mutually exclusive proteins provide a new system for manipulating protein folding.
Preface

A version of chapter 2 has been published. [Peng Q., Li H. (2008) Single molecule atomic force microscopy reveals kinetic partitioning of the mechanical unfolding pathway of T4-lysozyme. Proceedings of the National Academy of Sciences, USA, 105(6):1885-1890. I designed this project together with my supervisor Prof. Hongbin Li. I engineered the polyprotein chimera (GB1)$_4$/T4-lysozyme/(GB1)$_4$ and carried out both the single molecule AFM experiments and data analysis. I wrote the manuscript together with Prof. Li.

A version of chapter 3 has been published. [Peng Q., Li H. (2009) Domain insertion effectively regulates the mechanical unfolding hierarchy of elastomeric proteins: towards engineering multi-functional elastomeric proteins. Journal of the American Chemical Society, 131(39):14050-6. I designed this project together with my supervisor Prof. Hongbin Li. The domain-insertion protein GL5/T4L was engineered by Ms. Dingyue Khor under my instruction. I carried out both the single molecule AFM and the stopped-flow experiments. I analyzed all of the experimental data. I wrote the manuscript together with Prof. Li.

performed part of the single molecule AFM experiment on \textit{wt} TNfn3. Zhuang S. performed all of the SMD simulations on \textit{wt} TNfn3 and the corresponding data analysis. I carried out most of the single molecule AFM experiments. I analyzed all of the experimental data. I wrote the manuscript together with Prof. Li.

A version of chapter 5 has been submitted for publication. [Peng Q.], Fang J., Wang M., Li H. Kinetic partitioning mechanism governs the folding of the third FnIII domain of Tenascin-C: evidence at the single molecule level. I designed this project together with my supervisor Prof. Hongbin Li. Wang M. performed part of the single molecule AFM experiments. I carried out part of the single molecule AFM experiments. I analyzed all of the experimental data. I wrote the manuscript together with Prof. Li.

A version of chapter 6 has been published. [Peng Q.], Li H. (2009) Direct observation of tug-of-war during the folding of a mutually exclusive protein. Journal of the American Chemical Society, 131(37):13347-54. I designed this project together with my supervisor Prof. Hongbin Li. I engineered the hybrid protein GL5/I27w34f. I carried out both the single molecule AFM and the stopped-flow experiments. I analyzed all of the experimental data. I wrote the manuscript together with Prof. Li.

Check the first pages of these chapters to see footnotes with similar information.
Chapter 1: Introduction .................................................................................................................. 1
  1.1 The theory of protein folding ................................................................................................. 2
    1.1.1 Basic concepts of protein and protein folding ............................................................... 2
    1.1.2 The theories of protein folding ....................................................................................... 5
    1.1.3 The protein folding in vivo and in vitro: a comparison ................................................... 11
  1.2 Application of single molecule atomic force microscopy in protein mechanical folding/unfolding studies ......................................................................................................................... 13
    1.2.1 Why should we study the protein folding/unfolding at the single molecule level? .................................................................................................................................................. 13
    1.2.2 Principles of single molecule AFM ............................................................................... 14
      1.2.2.1 What is single molecule AFM? ............................................................................... 14
      1.2.2.2 How does the single molecule AFM work? .............................................................. 16
    1.2.3 Why should we study the protein folding/unfolding mechanics? .................................. 22
      1.2.3.1 Is the protein folding/unfolding mechanics biological relevant? ............................ 22
      1.2.3.2 The mechanical stability of a protein is different from its thermodynamic stability ................................................................................................................................. 24
      1.2.3.3 A unique anisotropic property for mechanical unfolding of a protein ..................... 29
    1.2.4 Accomplishments in the field of protein folding/unfolding study made by using single molecule AFM .................................................................................................................. 30
      1.2.4.1 The deviation from the two-state model of protein unfolding detected by using single molecule AFM .................................................................................................................. 30
      1.2.4.2 The detection of the hidden multiple unfolding pathways of the same protein .......... 32
      1.2.4.3 Studying the misfolding behavior of proteins and the mechanism related .................. 34
      1.2.4.4 Monitoring the protein folding process in real time ................................................... 36
      1.2.4.5 Revealing the insights on the energy landscape underlying the mechanical folding/unfolding of proteins using single molecule AFM ......................................................... 40
    1.2.5 The combination of the single molecule AFM with computer simulations .................. 42
  1.3 Objectives ............................................................................................................................. 44

Chapter 2: Single molecule atomic force microscopy reveals kinetic partitioning of the mechanical unfolding pathway of T4-lysozyme ........................................................................ 46
  2.1 Synopsis ............................................................................................................................... 46
  2.2 Introduction .......................................................................................................................... 47
  2.3 Results .................................................................................................................................. 50
    2.3.1 T4-lysozyme exhibits distinct multiple unfolding pathways .......................................... 50
2.3.2 Two-state unfolding of T4-lysozyme is characterized by a long unfolding distance to the transition state ................................................................. 55
2.3.3 Three-state unfolding of T4-lysozyme shows diversity in unfolding pathways ................................................................................................. 56
2.3.4 Possible pathways for three-state unfolding .................................................. 59
2.3.5 Interactions between helix A and reminders of the C-terminal lobe plays critical role in determining the mechanical stability of T4-lysozyme .......... 61

2.4 Discussion ................................................................................................. 65
2.4.1 Kinetic partitioning in the mechanical unfolding and folding of T4-lysozyme ................................................................................................. 65
2.4.2 Possible mechanisms for kinetic partitioning ........................................ 66

2.5 Experimental section ............................................................................ 69
2.5.1 Protein engineering ................................................................................ 69
2.5.2 Single molecule atomic force microscopy ........................................... 69

Chapter 3: Domain insertion effectively regulates the mechanical unfolding hierarchy of elastomeric proteins: towards engineering multi-functional elastomeric proteins ........................................................................ 71

Chapter 4: Mechanical design of the third FnIII domain of tenascin-C ........ 72
4.1 Synopsis .................................................................................................. 72
4.2 Introduction ............................................................................................. 73
4.3 Results ..................................................................................................... 77
4.3.1 The mechanical unfolding of TNfn3 is an apparent two-state process .... 77
4.3.2 The mechanical unfolding of TNfn3 is characterized by a long unfolding distance from the native state to the transition state ............................. 80
4.3.3 The SMD simulations of the mechanical unfolding of TNfn3 ............... 81
4.3.4 Constant velocity SMD simulations ...................................................... 89
4.3.5 Using site-directed mutagenesis to probe the nature of unfolding transition state observed in single molecule AFM ........................................ 89
4.3.6 The design of proline mutants of TNfn3 ................................................. 91
4.3.7 Phenotypic effects of proline mutations on A-strand of TNfn3 ............. 92
4.3.8 Phenotypic effects of proline mutations on G-strand of TNfn3: F88 is the Achilles heel of TNfn3 ................................................................. 97
4.3.9 Proline substitutions do not affect the mechanical unfolding distance of TNfn3 ............................................................................................... 100
4.4 Discussion ............................................................................................... 102
4.4.1 Mechanical unfolding of TNfn3: an FnIII domain of a robust mechanical design ................................................................................................. 102
4.4.2 Comparison of the mechanical unfolding of TNfn3 versus FNfn10: similar structure but different unfolding behaviors .................................... 104
4.4.3 Single-molecule AFM versus SMD: similarities and discrepancies ....... 106

4.5 Experimental section .............................................................................. 109
4.5.1 Protein engineering ........................................................................... 109
4.5.2 Single molecule atomic force microscopy ........................................... 110
4.5.3 SMD simulation .................................................................................. 111
4.5.4 Monte Carlo simulation ...................................................................... 111
Chapter 5: Kinetic partitioning mechanism governs the folding of the third FnIII domain of tenascin-C: evidence at the single molecule level

5.1 Synopsis
5.2 Introduction
5.3 Results
  5.3.1 Mechanical folding dynamics of TNfn3
  5.3.2 Folding intermediate states are detected in the mechanical unfolding/refolding cycles of TNfn3
  5.3.3 Folding of TNfn3 is influenced by its neighboring TNfn3 domains: misfolded superfolds
5.4 Discussion
  5.4.1 Single molecule AFM results provide supporting evidence for kinetic partitioning mechanism of protein folding
  5.4.2 Possible structure of the folding intermediate states
  5.4.3 Misfolding behavior of neighboring TNfn3 domains
5.5 Experimental section

Chapter 6: Direct observation of the tug-of-war during the folding of a mutually exclusive protein

Chapter 7: Summary and prospects

7.1 Summary
7.2 Prospects
  7.2.1 Revealing the molecule mechanism of protein misfolding
  7.2.2 Revealing the folding mechanism of the transmembrane proteins
  7.2.3 Designing novel molecular indicators/probes

References

Appendix A: Polyprotein engineering

A1. Sequences of proteins and the encoding cDNAs
  A1.1 pseudo wild type T4-lysozyme (T4L*)
  A1.2 PERM1
  A1.3 wild type GB1
  A1.4 GB1-L5 (GL5)
  A1.5 GL5/T4L
  A1.6 wild type TNfn3
  A1.7 S6P(TNfn3)
  A1.8 E9P(TNfn3)
  A1.9 K11P(TNfn3)
  A1.10 T14P(TNfn3)
  A1.11 A84P(TNfn3)
  A1.12 E86P(TNfn3)
  A1.13 F88P(TNfn3)
  A1.14 T90P(TNfn3)
  A1.15 F88A(TNfn3)
  A1.16 I27w34f
  A1.17 GL5/I27w34f
A2. Engineering the polyprotein from the gene level ............................................... 164
A3. Expression of the protein .................................................................................... 169
Appendix B: The mechanical unfolding of T4-lysozyme using single molecule AFM
..................................................................................................................................... 171
B1. Multiple unfolding pathways of T4-lysozyme are not originated from the
        heterogeneity due to different attachment sites of the cantilever ..................... 171
B2. Two-state unfolding of T4-lysozyme may involve unfolding intermediate state that
        unfold at forces below 20 pN .............................................................................. 172
B3. Classification of unfolding events of T4-lysozyme ............................................. 173
List of tables

Table 4.1  Unfolding force and kinetic parameters for the mechanical unfolding of TNfn3 and its mutants ................................................................. 97
List of figures

Figure 1.1  The schematic of “protein folding funnel” ................................. 7
Figure 1.2  The schematic for the kinetic partitioning mechanism (KPM) ......... 10
Figure 1.3  Schematics of single molecule AFM setup .................................. 16
Figure 1.4  Using single molecule atomic force microscopy to probe the mechanical unfolding of the individual protein. The three different modes for single molecule AFM experiment ..................................... 20
Figure 1.5  The energy landscape of protein folding and unfolding in the two-state model ................................................................. 26
Figure 2.1  The two state mechanical unfolding of T4-lysozyme .................... 53
Figure 2.2  T4-lysozyme can unfold via multiple distinct three-state unfolding pathways .............................................................. 54
Figure 2.3  Unfolding forces of T4-lysozyme show a weak dependency on the pulling speed ................................................................. 56
Figure 2.4  Histogram of $\Delta L_{C1}$ (A) and $\Delta L_{C2}$ (B) during three-state unfolding trajectories ...................................................... 59
Figure 2.5  Mechanical unfolding behaviors of T4-lysozyme circular permutant PERM1 ............................................................................. 64
Figure 4.1  The three-dimensional structure of the third fibronectin type III domain of tenascin-C (TNfn3) .................................................... 79
Figure 4.2  Typical force-extension curves of polyprotein (TNfn3)............... 79
Figure 4.3  Unfolding force of TNfn3 and its dependence on the pulling speeds ................................. 81
Figure 4.4  Constant force and constant velocity SMD simulations of the mechanical unfolding of TNfn3 .................................................. 84
Figure 4.5  Snapshots of TNfn3 during its simulated mechanical unfolding .... 85
Figure 4.6  Profiles of hydrogen bond energy of inter-strand hydrogen bond in A-B and F-G strands versus time in two representative SMD unfolding trajectories of TNfn3 ...................................................... 88
Figure 4.7  Mechanical unfolding of proline mutants of TNfn3 .................... 93
Figure 4.8  Typical force-extension curve of polyprotein chimera (GB1-E9P) that shows missing unfolding force peaks of E9P .................. 94
Figure 4.9  The unfolding force histograms of TNfn3 proline mutants as well as F88A .............................................................................. 95
Figure 4.10  Pulling speed dependence of the mechanical unfolding force of TNfn3 mutants ........................................................................... 101
Figure 4.11  Using Monte Carlo simulation to estimate the unfolding rate constant $a_0$ and unfolding distance $\Delta x_u$ ..................................... 113
Figure 5.1  The majority of TNfn3 folds in a two-state fashion .................... 119
Figure 5.2  Mechanical folding experiments revealed that TNfn3 can fold into multiple distinct folding intermediate states ................. 123
Figure 5.3  The misfolding behaviors of TNfn3 involving neighboring TNfn3 domains ............................................................................ 125
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 5.4</td>
<td>Mechanical properties of the misfolded superfold of TNfn3 domains</td>
<td>126</td>
</tr>
<tr>
<td>Figure 5.5</td>
<td>Misfolded skips of TNfn3 domains can unfold in two steps</td>
<td>128</td>
</tr>
<tr>
<td>Figure 5.6</td>
<td>Possible structures of the observed folding intermediate states</td>
<td>132</td>
</tr>
<tr>
<td>Figure 5.7</td>
<td>The rare mechanical unfolding events detected by single molecule AFM</td>
<td>133</td>
</tr>
<tr>
<td>Figure A1</td>
<td>The restriction sites we used for protein engineering</td>
<td>165</td>
</tr>
<tr>
<td>Figure A2</td>
<td>General procedure for engineering the gene of polyprotein</td>
<td>167</td>
</tr>
<tr>
<td>Figure A3</td>
<td>The schematic of the gene of the constructed polyprotein chimera (GB1)4/T4-lysozyme/(GB1)4</td>
<td>168</td>
</tr>
<tr>
<td>Figure A4</td>
<td>The photos of DNA electrophoresis in the agarose gel for the genes encoding the (A) monomer, (B)dimer, (C) tetramer and (D) octamer of the protein TNfn3</td>
<td>169</td>
</tr>
<tr>
<td>Figure B1</td>
<td>Pulling angle may affect the measured unfolding force and contour length increment of T4-lysozyme</td>
<td>172</td>
</tr>
</tbody>
</table>
## List of symbols and abbreviations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>loading rate of force</td>
</tr>
<tr>
<td>AFM</td>
<td>atomic force microscopy (microscope)</td>
</tr>
<tr>
<td>ATP</td>
<td>adenosine triphosphate</td>
</tr>
<tr>
<td>BR</td>
<td>bacteriorhodopsins</td>
</tr>
<tr>
<td>C</td>
<td>concentration of protein samples (in M)</td>
</tr>
<tr>
<td>CD</td>
<td>circular dichroism</td>
</tr>
<tr>
<td>d</td>
<td>path length of the cuvette (in cm)</td>
</tr>
<tr>
<td>ddFLN4</td>
<td>the fourth domain of distyostelium discoideum filamin</td>
</tr>
<tr>
<td>deg</td>
<td>degree</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DSE</td>
<td>denatured state ensemble</td>
</tr>
<tr>
<td>EB</td>
<td>ethidium bromide</td>
</tr>
<tr>
<td>E. Coli</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>ECM</td>
<td>extra-cellular matrix</td>
</tr>
<tr>
<td>EGF</td>
<td>epidermal-growth-factor</td>
</tr>
<tr>
<td>E_x</td>
<td>extension of the molecule</td>
</tr>
<tr>
<td>F</td>
<td>force</td>
</tr>
<tr>
<td>FES</td>
<td>free energy surface</td>
</tr>
<tr>
<td>FnIII</td>
<td>fibronectin type III</td>
</tr>
<tr>
<td>FNfn10</td>
<td>the tenth fibronectin type III domain of fibronectin</td>
</tr>
<tr>
<td>FRET</td>
<td>forster resonance energy transfer</td>
</tr>
<tr>
<td>F_u</td>
<td>unfolding force</td>
</tr>
</tbody>
</table>
GB1  the B1 binding domain of protein g from *streptococcus*
GdmCl  guanidinium chloride
GFP  green fluorescent protein
GL5  a loop insertion mutant of GB1 in which five residues (GGGLG) were inserted into the second loop of GB1
$h$  Planck constant
I27  the 27th Ig domain of human titin
I1  the first unfolding intermediate of TNfn3 along the mechanical unfolding pathway observed in the SMD simulation
I2  the second unfolding intermediate of TNfn3 along the mechanical unfolding pathway observed in the SMD simulation
I3  the third unfolding intermediate of TNfn3 along the mechanical unfolding pathway observed in the SMD simulation
Ig  immunoglobulin
IgG  immunoglobulin g
IPTG  isopropyl-1-β-d-thiogalactoside
$\kappa$  proportionality constant
$k_B$  Boltzmann constant
$k_c$  spring constant of the cantilever
$K_{eq}$  equilibrium constant
KPM  kinetic partitioning mechanism
LB  Luria-Bertani broth
$L_C$  contour length
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_{\text{Extension}}$</td>
<td>extension</td>
</tr>
<tr>
<td>M</td>
<td>molar per liter</td>
</tr>
<tr>
<td>MBP</td>
<td>maltose-binding protein</td>
</tr>
<tr>
<td>mdegree</td>
<td>milidegree</td>
</tr>
<tr>
<td>MRE</td>
<td>mean residue ellipticity</td>
</tr>
<tr>
<td>N</td>
<td>native state</td>
</tr>
<tr>
<td>NBA</td>
<td>native basin of attraction</td>
</tr>
<tr>
<td>nm</td>
<td>nanometer</td>
</tr>
<tr>
<td>NMR</td>
<td>nuclear magnetic resonance</td>
</tr>
<tr>
<td>OD</td>
<td>optical density</td>
</tr>
<tr>
<td>$p$</td>
<td>persistence length</td>
</tr>
<tr>
<td>PBS</td>
<td>phosphate buffer saline</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PDB</td>
<td>protein data bank</td>
</tr>
<tr>
<td>PERM1</td>
<td>a circular permutant of T4-lysozyme in which the first 11 amino acid residues of T4-lysozyme were relocated to its C-terminus</td>
</tr>
<tr>
<td>PID</td>
<td>proportional, integral and differential amplifier</td>
</tr>
<tr>
<td>R</td>
<td>gas constant</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>$R_{\text{NC}}$</td>
<td>distance between the two termini</td>
</tr>
<tr>
<td>s</td>
<td>second</td>
</tr>
<tr>
<td>S.D.</td>
<td>standard deviation</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>sodium dodecyl sulfate polyacrylamide gel electrophoresis</td>
</tr>
</tbody>
</table>
SMD  steered molecular dynamics

STM  scanning tunneling microscopy

T  temperature

TNfnALL  a recombinant protein fragment with all the fifteen fnIII domains from human tenascin-C

t  time

TEM  transmission electron microscopy

TNfn3  the third fibronectin type III domain of human tenascin-C

T4L  T4-lysozyme

U  unfolded state

UV  ultraviolet

v  pulling speed

WLC  worm-like chain

\( W T^* \)  pseudo wild type

\( \alpha_0 \)  unfolding rate at zero force

\( \alpha(F) \)  unfolding rate at force \( f \)

\( \beta_0 \)  folding rate at zero force

\( \theta_{\text{obs}} \)  the observed ellipticity (in deg)

\( \Phi \)  the partition factor

\( \phi \)  \( \phi \) value, \( \phi = \frac{(\Delta G_{W}^{T S-D} - \Delta G_{M}^{T S-D})}{(\Delta G_{W}^{N-D} - \Delta G_{M}^{N-D})} = \frac{\Delta \Delta G^{T S-D}}{\Delta \Delta G^{N-D}} \)

\( \Delta G \)  the free energy change during the folding/unfolding transition

\( \Delta G_{N\cdash T} \)  free energy difference between the transition state and the native state
$\Delta G_{T\text{-}U}$ free energy difference between the transition state and the unfolded state

$\Delta G_{N\text{-}U}$ free energy difference between the native state and the unfolded state

$\Delta L_C$ contour length increment

$\Delta L_{C1}$ contour length increment upon the transition from native state to intermediate state

$\Delta L_{C2}$ contour length increment upon the unfolding of the intermediate

$\Delta L_{C\text{-}skip}$ contour length increment upon the unfolding of the misfolded TNfn3 domains

$\Delta L_{C\text{total}}$ contour length increment upon the fully unfolding of the hybrid domain-insertion protein GL5/T4L

$\Delta x_c$ displacement of the cantilever

$\Delta x_p$ movement of the piezoelectric positioner

$\Delta x_f$ folding distance, the distance between the unfolded and the transition state

$\Delta x_u$ unfolding distance, the distance between the native and the transition state

3D three-dimensional

$[\text{denaturant}]^{1/2}$ the midpoint of the protein unfolding transition at which 50 percent of the population is folded and 50 percent is unfolded
Acknowledgements

First of all, I would like to express my earnest appreciation to my supervisor, Dr. Hongbin Li. Dr. Li opened the door of biophysics for me and brought me into this flourishing field. Throughout my PhD study, his supervision, guidance, and support always helped me conquer the challenges I encountered. His solid expertise, acute instinct, dedication to work and passion on research make him my role model to be a real scientist. After working with Dr. Li’s for more than five years, I would say he is not only an excellent mentor, but also a good friend. The experience of working with Dr. Li is definitely the invaluable fortune for my future.

Secondly, I would also like to thank all of my groupmate from Dr. Li’s laboratory. I thank you for all of your technical and intellectual assistance. Your friendship will always be cherished.

In particular, thank you to Yi Cao and Deepak Sharma for your training on my molecular biology techniques and single molecule AFM techniques at the beginning of my PhD study. Thank you to Shunlin Zhuang, Meijia Wang, Ying Guo, Ashlee Jollymore, Eileen Wang, Dingyue Khor, Anderson Chang, Yuanai Khor, Martha Ma, M.M. Balamurali, Peng Zheng, Shanshan Lv, Kai Shih Er, Tao Shen, Chengzhi He, Devin Li, Na Kong and Jie Fang for your friendship that have accompanied me during the past five years.

Thirdly, I would say thank you to Dr. David D.Y.Chen, Dr. Pierre Kennepohl and Dr. Takamasa Momose for being on my supervisory committee. Your technical and intellectual guidance is truly appreciated. I especially thank Dr. David D.Y.Chen for his
careful evaluation on the draft of my thesis. I also thank Dr. Brian R. James for serving as the chair of my Comprehensive Exam.

I am also greatly indebted Dr. Brian W. Matthews of University of Oregon, Dr. Martin Sagermann of UCSB and Dr. Harold Erickson of Duke University. I received the plasmids that encode the pseudo wild type T4-lysozyme protein and its circular permutant PERM1 from Dr. Brian W. Matthews and Martin Sagermann as generous gifts. I received the plasmids that encode TNfnALL from Dr. Harold Erickson as a generous gift.

Next, I would like to thank all my friends and family. In particular, thank you to Ying Li and Rui Hua for your invaluable friendship. I am so fortunate that I can have your friendship.

Last but not least, I would like to thank my parents for their enormous and selfless love.
Dedication

This dissertation is dedicated to members of my family, including my mother, Bende JIN, my father, Xinglong PENG, and my wife, Jiehua ZHOU. My parents give me my life and all their love and my wife supported me to finish this thesis with her tolerance and encouragement.
Chapter 1: Introduction

Understanding the molecular mechanism underlying protein folding is one of the most essential questions in Biochemistry and remains an open question. Currently, most of the knowledge about the protein folding/unfolding has been extracted from experiments based on the traditional ensemble techniques. From those experiments, the thermodynamic/kinetic information about proteins in solution is obtained based on the average of the properties from a large number of molecules. Combined with powerful computer simulation techniques, many theories and models have been proposed (1-22). However, the knowledge from those traditional ensemble studies is averaged from the observation of over $10^{14}$ molecules at a time (23). However, because molecular details about the folding and unfolding dynamics at the single molecule level are masked, the folding/unfolding process of an individual protein molecule cannot be probed in an ensemble measurement.

Fortunately, the rapid development of single molecule techniques in the last two decades enables one to manipulate a single protein molecule each time with millisecond to second temporal resolution and nanometer spatial resolution. By directly observing the folding/unfolding reaction of an individual protein molecule under native conditions, either *in vitro* or *in vivo*, many important questions in protein folding could be answered. Single molecule techniques have their unique advantages in the study of protein folding (23-25).

The mechanical folding/unfolding of proteins is one of the important aspects of protein folding/unfolding. Mechanical folding/unfolding of proteins could be involved in
many biological processes, such as transportation of proteins across membranes (the import machinery of organelles(26)), protein degradation through proteasomes (26, 27), and catalyzed protein folding (28), etc(29). Therefore, to study the mechanical folding/unfolding mechanism of proteins at the single molecule level, single molecule force spectroscopy techniques(30, 31) are critically important. In this thesis, I will only focus on the single molecule atomic force microscope (AFM) studies on protein mechanical folding/unfolding problems.

In this chapter, I will give a review on this relatively new research field. The review will cover the following information: 1) The basic concepts and the current protein folding theories. In this part, I will mainly focus on the Kinetic Partitioning Mechanism (KPM) theory of protein folding – one of the currently widely accepted protein folding theories. 2) Application of single molecule atomic force microscopy in protein mechanical folding/unfolding studies. In this part, I will explain in detail what single molecule AFM is and why single molecule AFM is important for protein folding/unfolding studies. I will also present some examples to emphasize the advantages of using single molecule AFM to study protein folding. 3) The objectives of this thesis.

1.1 The theory of protein folding

1.1.1 Basic concepts about proteins and protein folding

Proteins are essential parts of all type of organisms and are involved in virtually every physiological process. Numerous efforts have been carried out to explore the properties of proteins. Unlike most other linear polymers, under physiological conditions, almost all of the natural proteins undergo a spontaneous transition from a disordered state
to an ordered state, which is called “folding.” The polypeptide chain is considered to be random coil when unfolded, similar to the other polymer molecules in their theta solvent, but spontaneously forms its unique native, three-dimensional (3D) structure in physiological media. Within cells, immediately after synthesis, or even during the translation, the polypeptides can fold into the specific tertiary structure efficiently.

The unique 3D structures of proteins not only single proteins out from the regular polymers, but are also critical for their biological functions. The first three-dimensional structure of proteins was revealed by X-ray crystallography in 1958 and the protein structures were believed firm and rigid.(32-34) After that, thanks to the emergence of new techniques and instruments, such as nuclear magnetic resonance (NMR)(35-38), low-temperature flash photolysis(39-42) and hydrogen exchange (HX) techniques(43-47), we have understood much more about protein structures/conformational dynamics. With more than 65000 protein structures having been deposited so far (from the Protein Data Bank (www.rcsb.org)), numerous details on protein structures have been revealed. Most of the known proteins, in their native states, are built upon a molecular skeleton of hydrogen-bonded structural elements, α-helix and β-sheet, which are interconnected by tight turns and flexible loops. These elements are held together through hydrogen bonds between the adjacent elements and the hydrophobic interactions between the side-chains of amino acid residues. Anfinsen’s experiments which won him the Nobel Prize proved that the amino acid sequence alone is adequate to define a protein’s tertiary structure(1). However, the molecular mechanism underlying the folding of proteins remains an open question, probably the most essential open question in Biochemistry.
The passion on searching for answers of the question – “how does a polypeptide chain fold”, has spurred the major developments on protein studies, both experimental (48-50) and theoretical (21, 51, 52). The “protein-folding(unfolding) problem” is, in fact, several separated but not independent questions. The first question is essentially about the thermodynamics: Why can a newly synthesized polypeptide chain fold spontaneously? The second is also about the thermodynamics: How does an amino acid sequence identify its three-dimensional (3D) structure? In other words, how does the protein determine a structure with the lowest free energy under physiological media as its native structure? The third is a kinetic question: Why/How can the polypeptide chain acquire its native state (N) efficiently within a reasonable timescale? Finally, as a long linear molecule, how does a protein obviate the possible pathways which lead to misfolding? Trying to spell out the mechanism responsible for protein folding, considerable efforts have been dedicated to the field of protein folding. Before the prosperity of the single molecule study on protein folding during the last two decades, the major tool for the protein folding/unfolding study is the thermodynamic/kinetic measurements in solution. From those experiments, fundamental concepts that the folded and unfolded forms of a protein are in a dynamic equilibrium were built up. In other words, most proteins can reversibly unfold/fold according to the environmental conditions. Current experimental knowledge comes mainly from those studies. The combination of powerful computer simulation techniques and experiments has successfully provided insights into the mechanisms of protein folding (53, 54) and many theories have been proposed(1-22).
1.1.2 The theories of protein folding

For most natural proteins, the folding is an intrinsic tendency encoded in their primary sequence. As long as the external conditions (pH, temperature, ionic strength, and cosolutes) allow, the proteins tend to exist in their native state (conformation). Theorists have explained this phenomenon by speculating that the native structure of proteins represents the minimum of their global free energy landscape. This speculation has been validated in plenty of protein thermodynamic experiments. Based on this speculation, the transition state theory has been adopted to depict the thermodynamics of protein folding, such as the equilibrium between the native state and unfolded state of the protein. So far, the transition state theory can successfully explain most of the observed thermodynamic experimental data. From a macroscopic level, the kinetics of protein folding can also be well explained using transition state theory. However, those studies did not address the conundrum of how a free unfolded protein molecule efficiently reaches its native state within a biologically relevant time scale. As Levinthal argued (55, 56), the kinetic accessibility of the native structure of a protein molecule cannot be based on a random research of all available conformations. This suggests that favorable pathways must exist which can essentially confine the search of the conformation within a limited space. Therefore, the proteins will reach the native structure within a biologically relevant time scale. Although the Levinthal paradox is over simplified, it still can help us to understand protein folding at the single molecule level.

Over the last two decades, through the interplay of sophisticated experiments (23, 25, 46, 57-59) and the study of protein folding models(8, 17-19, 21, 51, 60), a widely
accepted framework for understanding folding kinetics has emerged which can be described as the “protein-folding funnel” model\( (3, 8, 14, 15) \). The essential concept is that there is a funnel-shaped energy landscape underlying protein folding. The protein is folding along a continuum trajectory toward a global free-energy minimum and the folding process is explicitly amino acid sequence dependent — “Every unique sequence has its own funnel”\( (61) \). In the “protein-folding funnel” model, protein folding is described as a process that advances from a high free energy (also highly disordered) state with few, if there is any, intramolecular interactions (open end), to a low free energy (also organized) state with native intramolecular interactions (bottom) (Fig. 1.1). In the unfolded state, the polypeptide behaves as a random coil with the maximum entropy and can sample the huge conformational space; under the folding condition, the protein is folded into its native structure and conformation(s) with the relatively lowest entropy. The surface of the funnel could be smooth, but all of the current theoretical studies are assuming a rugged underlying free energy landscape for proteins \( (3, 18, 19, 22) \). The rugged free energy surface (FES) is consisting of many local minima separated by barriers with different heights. However, despite their complicated FES, each protein has an accessible dominant basin which can be reached within the biological time scale. Within this theoretical framework, the major challenge is how to depict the process by which a polypeptide chain finds a path to the global free energy minimum in the complicated FES. The other challenges include the description of the nature of the intermediates and the structural features of the transition state ensemble (TSE). To answer these questions, D. Thirumalai and his colleagues have developed a conceptual framework which is complementary to the “protein-folding funnel” model, the kinetic
partitioning mechanism (KPM)(21, 51, 52).

![Image]

**Figure 1.1** The schematic of “protein folding funnel”. The funnel represents a rugged energy landscape with multiple kinetic traps, energy barriers and folding pathways. The folding kinetics of protein is likely multiple-exponential (adapted from the reference(19)). The shape of energy landscape is a bumpy bowl-like funnel. Each point on the landscape represents a possible conformation with certain free energy that can be adopted by the polypeptide. The free energy corresponding to each point is a function of the conformational freedom.

The groundwork of the KPM originated from the assumption that there are many local low-energy minima (in which the proteins are partially folded or misfolded) located on the FES besides the global dominant basin (native state). According to the KPM, those local low-energy minima are explained by introducing the concept of topological frustration. For a typical globular protein, the fraction of the hydrophobic amino acid (a.a.)
residues within its primary sequence is slightly more than 50 percent (62). Obviously, the distribution of those hydrophobic a.a. residues should be uniform throughout the whole contour of the polypeptide chain, at least roughly; otherwise the proteins would be unstable and tend to aggregate (52). This characteristic could have been selected by nature through evolution. Furthermore, the uniform distribution of the hydrophobic a.a. residues will be independent of the length of the polypeptide chain. Therefore, within the contour of a polypeptide chain, the hydrophobic a.a. residues from any continuous part of the chain with a length \( l \) (\( l < L_C \), \( L_C \): contour length) will have a tendency to form tertiary contacts as long as the environment allows. In all probability, the resulting conformations, modeled by interactions between local hydrophobic residues, would be incompatible with the unique native structure under folding conditions. This incompatibility between the structures at shorter length scales (regionally and partially) and the native conformation (full length scale) is defined as topological frustration. Considering the competing interactions between different parts of the polypeptide chain and the continuously connected a.a. residues, the topological frustration is an inevitable effect for proteins with any primary sequence. Therefore, topologically, all proteins are frustrated. The longer the polypeptide chain is, the more it is frustrated. The apparent deduction from proteins’ topological frustration is the complicated underlying topography of the FES with many separated minima. (52)

In the KPM theory, the numerous local minima have clear physical meaning. Qualitatively, the local minima are the kinetic intermediate states (traps); either partially folded structures containing native-like contacts or totally misfolded conformations with non-native contacts. According to the topological frustration, for a polypeptide chain
consisting of a limited number of a.a. residues, the possible ways to self-assemble into a sub-stable conformation would be astronomical (52). Most of those conformations would be incompatible with the global fold of the native protein. Since only part of the whole polypeptide sequence is involved in those incompatible conformations, the majority of those structures would have relatively high free energies. It would be hard for those conformations to survive through the thermal fluctuations and they could only transiently trap the polypeptide sequence in the corresponding local minima. Nevertheless, if the free energy barrier which separated a local minimum and the global minimum apart is high enough, in other words, when certain structure has a thermodynamic stability which is comparable with the native state, such a structure could potentially trap the polypeptide molecule in it for a realistic time. These “stable” non-native structures are known as kinetic traps. During the folding process, when some protein molecules are trapped into the kinetic traps, the folding rate of those protein molecules will be slowed down.
Figure 1.2 The schematic for the kinetic partitioning mechanism (KPM). A fraction of the unfolded population, which is given by \( \Phi \) (the partition factor) (21), goes directly to the native state. This folding process is defined as the “fast process”. The remaining fraction of the molecules, \((1-\Phi)\), gets trapped in intermediate states or misfolded structures. This folding process is defined as the “slow process”.

The topological frustrations are predicted as an inherent property for every primary sequence of proteins (51). Therefore, the kinetic traps are also in all likelihood existing on every FES. When the folding process is initiated, all of the unfolded polypeptide chains are synchronized to explore the native basin of attraction (NBA) on the FES. As the assumption of all current protein folding theories, including KPM, the unfolded protein molecule exists as the random-coil with great freedom to adopt any possible conformation. Thus, at the “starting point” of the folding process, an individual molecule could start its folding process from any unique initial conformation, which would also lead to a unique route crossing the FES in search of the NBA. Statistically, a certain fraction (\( \Phi \)) of the unfolded population would directly fold into the NBA and successfully evade any kinetic traps. This subpopulation will behave as faster folders who
take the fast process. The remaining fraction \((1-\Phi)\) would be doomed to struggle through one or more low-energy traps and experience some “intermediate(s)” (Fig. 1.2) (21). Therefore, it will take a longer time scale for them to navigate to the native structure and those molecules are apparent slow folders. Thus, the existence of the kinetic traps on the rugged FES can partition the whole population of the denatured molecules into two classes: fast folders and slow folders. In the framework of KPM, the existence of multiple folding(unfolding) pathways has become inevitable, especially at the single molecule level. The distribution of the whole population of the protein molecules in different folding(unfolding) pathways is mainly determined by the relative height and number of the kinetic traps along each pathway. Therefore, this phenomenon is the so called “kinetic partitioning”. Although this concept is currently popular, supporting experimental proof remains limited. So far, most of the limited reported experimental evidence are indirect proofs (63-66). (We will discuss this topic in more detail in Chapter 2)

1.1.3 The protein folding in vivo and in vitro: a comparison

Most theories of protein folding, including the KPM theory, and the majority of current protein folding studies are focusing on the protein folding in vitro. Those theories and studies are all based on the well-known Anfinsen’s hypothesis in which the folding process of the protein has been predicted as a self-assembly process. However, for the protein folding in vivo, the Anfinsen’s hypothesis, in the strict sense, may have to be reframed as an assisted self-assembly process. It has been found, in vivo, a significant fraction of proteins fold with the assistance of molecular chaperones — a family of unrelated classes of protein. Molecular chaperones are proteins which can mediate the
proper assembly of other polypeptides while they are not involved into the final functional assembled structures. The molecular chaperones has been proposed to package their target protein into correct conformation via two possible mechanisms: 1) inhibiting the unproductive assembly pathways which could act as kinetic traps and lead to misfolded structures; 2) unfolding the kinetically trapped misfolded structures. No matter which mechanism is actually carried out in vivo, the mechanical interaction between the chaperone protein and its target polypeptide chain is predicted important and critical. Understanding the mechanism of the chaperone assisted protein folding in vivo continues to be an important task which requires further experimental and theoretical efforts. In this thesis, I will only focus on the protein folding/unfolding in vitro, both experimentally and theoretically.

Among the many protein folding theories, the KPM theory has led to some testable hypotheses. Many of these predictions have been verified which is owing to the invention of novel experimental methods that has allowed us to probe the protein folding process with microsecond temporal resolution(23, 25, 46, 57-59). Of course, this theory is still far from being perfect and there are still some aspects of the problem of protein folding that remain to be elucidated. Some open questions which remain to be answered are: Can this funnel landscape theory explain the folding mechanism for downhill barrierless folders? What is the mechanism underlining the folding process so that a fully unstructured polypeptide chain can fold into its native structure in an incredibly efficient way? To tackle these questions, further developments of both experimental and theoretical tools will be needed.
1.2 Application of single molecule atomic force microscopy in protein mechanical folding/unfolding studies

1.2.1 Why should we study the protein folding/unfolding at the single molecule level?

Thanks to the advances in single molecule instrumentation, which is critical to the key progress made in the field, single-molecule research has flourished over the last decade and cultivated much in-depth interdisciplinary research collaboration, involving elements from Chemistry, Physics, Biology and Electronic Engineering (24, 67-78). Recently, single molecule techniques have been introduced into the protein folding study by monitoring the folding of one individual molecule at a time (24, 79, 80). Several advantages make single-molecule techniques especially alluring and powerful for the study of protein folding (23-25). Here, I am only focusing on the application of the single molecule techniques in the field of protein mechanical folding/unfolding. The single molecule techniques are so powerful that we are enabled to picture the FES landscape and the protein folding process in greater detail than ever before. First, protein folding is measured one molecule at a time in a single-molecule experiment. Second, dynamics of protein folding/unfolding can be repetitively measured under equilibrium conditions for single molecules (or small ensembles). Third, by manipulating a single molecule at a time, single-molecule force spectroscopy allows the direct measurement of forces associated with protein folding/unfolding at a molecular level, as well as the molecular conformational and functional responses to mechanical forces. Finally, the extremely low detection limit of single-molecule methods also directly benefits the protein folding/unfolding study. One day, single molecule techniques will ultimately enable us to
decipher the protein folding mechanism, both the thermodynamic folding and mechanical folding, at the single molecule level.

Currently, two principal single molecule spectroscopy methods are being used: force based techniques and fluorescence based microscopy methods. The fluorescence based methods, especially the Forster resonance energy transfer (FRET)(25, 57), has produced a lot of important, interesting and previously inaccessible insights on protein folding. For mechanical folding/unfolding, single molecule force spectroscopy techniques are required. Amongst single molecule force spectroscopy techniques, single molecule atomic force microscopy (AFM) and optical tweezers are two most widely used methods. In this thesis, I will only focus on single molecule AFM.

1.2.2 Principles of single molecule AFM

1.2.2.1 What is single molecule AFM?

The first atomic force microscope (AFM) in the world was built in 1986 by Binnig, Quate and Gerber(81). In principle, AFM can provide real atomic resolution in both ultra-high vacuum and liquid environments. Since its birth, AFM has been mainly utilized in imaging samples(82, 83). Single molecule AFM technique (single molecule force spectroscopy) is one of the relatively newer major applications of AFM(84, 85) besides imaging. As shown in Figure 1.3, single molecule AFM typically consists of two key systems: force sensing/controlling system and movement sensing/controlling system. AFM gathers the information of the samples by "feeling" the force between the sample and the sharp mechanical probe (tip). The sharp probe, with a typical radius of curvature 20 nm (for Veeco probe, model: MLCT), is connected to the end of a soft cantilever,
which is typically constructed from silicon or silicon nitride. If the tip is brought close enough to the sample surface, forces between the tip and the sample will be strong enough to cause a deflection of the cantilever. Generally, the deflection of the cantilever is measured precisely using the laser deflection method (Fig. 1.3). The laser light from a diode laser source is reflected off from the back surface near the very end of the cantilever and collected by a position sensitive photo diode, leading to the accurate measurement of the deflection of the cantilever, which in turn can be converted to the force experienced by the cantilever. As for the displacement measurement, the relative movement between the sample and the force probe is manipulated by a piezoelectric positioner. Typically, the sample is mounted on a coverslip attached to the piezoelectric positioner which can move the sample in the x, y and z directions. In most cases, there is a feedback mechanism employed to control the tip-to-sample distance.
In a typical single molecule AFM experiment, a tandem modular polyprotein is studied. The polyprotein is adsorbed to the surface of a coverslip which is typically golden or glass and mounted onto the Piezoelectric positioner. The polyprotein is picked up randomly by the cantilever tip through non-specific adhesion and stretched from the substrate. The cantilever and the protein molecules are completely immersed in buffer. The backsurface of the cantilever is typically coated with a reflective golden layer. The deflection of the cantilever is detected by measuring the movement of the laser spot on the photodiode.

1.2.2.2 How does the single molecule AFM work?

There are also different modes of operation for single molecule AFM: constant velocity mode(86), force-clamp mode(87, 88) and force-ramp mode(87, 89). For the constant velocity mode, the object molecule is stretched at a constant pulling speed; whereas in the force-clamp mode, a constant force is applied to the target molecule while the end-to-end distance of the molecule is monitored as the function of time; as to the
force-ramp mode, the force applied to the molecule is ramped at a constant rate (from low to high) which means the pulling force is a linear monotonic increasing function of time.

In constant velocity mode, when performing the experiment, the force is monitored as a function of the end-to-end distance of the stretched molecule (extension $L_{\text{Extension}}$). The collected data is plotted as the force-extension curve, which directly records the mechanical unfolding process of the target protein at the single molecule level. Figure 1.4A shows the schematic of the constant-velocity experiment and the characteristic force-extension curve from pulling a polyprotein (TNfn3)$_4$. The force-extension curve shows the typical periodical saw-tooth pattern. It can be easily identified that the elongation of the polyprotein molecule is nonlinearly correlated with the force (Fig. 1.4A, from stage 1 to stage 2). The non-linear increase of the force ($F$) versus the extension ($L_{\text{Extension}}$) can be well depicted with the worm-like chain (WLC) model of polymer elasticity (1.1)(90).

$$F(x) = \frac{k_B T}{p} \left\{ \frac{1}{4} \left[ \left( 1 - \frac{x}{L_C} \right)^2 - 1 \right] + \frac{x}{L_C} \right\}$$

(1.1)

where $F(x)$ is the force at extension $x$, $k_B$ is Boltzmann constant, $T$ is absolute temperature, $p$ is the persistence length and $L_C$ is the contour length of the polymer. For a given molecule with known parameters, the entropic restoring force can be theoretically calculated at any given extension of the molecule. When the force is high enough, it will stochastically trigger the unfolding of one of the protein domains of the polyprotein. The unfolded domain will lose force resistance immediately and be extended by the stretching force. Consequently, the effective contour length of the whole polyprotein chain increases greatly and the entropic restoring force generated by the molecule quickly drops
back to a low value (Fig. 1.4A, from stage 2 to stage 3). If the polyprotein is further pulled before the molecule detaches from the tip or the substrate, the cycle of increase/decrease of the force will repeat until all of the domains in the polyprotein are unfolded (Fig. 1.4A, from stage 3 to stage 4). The sequential unfolding of protein domains in the polyprotein gives rise to a saw-tooth like pattern in the force-extension curve. Each peak in the force-extension curve corresponds to the unfolding of a TNfn3 domain in the polyprotein chain and the peak height is a direct measurement of the mechanical stability of TNfn3 at the single molecule level.

In the force-ramp mode, the end-to-end distance of the protein picked up is monitored as a function of force. The deflection of the cantilever is precisely controlled through an electronic feedback loop (by manipulating the displacement of the piezoelectric positioner) so that the force increases linearly as a function of time (Fig. 1.4B). The unfolding of each individual module will lead to a spike in the force-time curve. The extension-time curve exhibits a stair case-like appearance. Each stair in the curve represents the unfolding of a protein domain in the polyprotein and the extension as a function of time at each stage is not constant. Figure 1.4B shows the schematic of the force-ramp experiment and the correlated characteristic step-wise extension-time curve and force-time curve from stretching the polyprotein (TNfn3)$_4$. The length of the molecule increases with time following the WLC model of polymer elasticity.

The force-clamp mode, to some extent, is similar to the force-ramp mode, in which the force is regulated by a sensitive electronic feedback loop. In the force-clamp mode, the force applied to the protein is kept constant through the electronic feedback loop and the end-to-end distance of the protein picked up is monitored as a function of
Upon the unfolding of one of the domains in the polyprotein chimera, the effective contour length of the whole molecule will increase. The sequential unfolding of the domains in the polyprotein chimera will result in a stepwise extension-time curve in which each step corresponds to the unraveling of an individual protein module. Figure 1.4C shows the schematic of the force-clamp experiment and the correlated characteristic extension-time curve from pulling the polyprotein (TNfn3)$_4$. The extension-time curve shows the typical stepwise pattern with a stochastic nature. The unfolding of each domain in the polyprotein molecule will result in an identical contour length increment ($\Delta L_c$) for the whole molecule while the dwell time for each individual domain is scattered and not apparently related to the stretching force (Fig. 1.4C). However, the unfolding traces are definitely correlated with the amplitude of the force. By averaging multiple normalized recordings into one time course, the unfolding time course of the polyprotein can be readily described by a single exponential which is consistent with a Markovian process where the probability of unfolding at any given time is independent of the previous history(91). It is clear that the unfolding kinetics is directly controlled by the mechanical force. In the work reported in this thesis, I performed most of the single molecule AFM experiments in the constant velocity mode.
Figure 1.4 Using single molecule atomic force microscopy to probe the mechanical unfolding of the individual protein. The three different modes for single molecule AFM experiment: (A) constant velocity mode, (B) force-ramp mode and (C) force-clamp mode.

A) The schematic of the constant velocity mode of single molecule AFM measurements. The end-to-end distance of the polyprotein is manipulated by the piezoelectric positioner and increases at a constant velocity. The force applied on the polyprotein is a function of the end-to-end distance of the polyprotein obeying the WLC model and can be measured from the deflection of the AFM cantilever.
The stretching and subsequently unfolding a tandem modular polyprotein in the constant velocity mode can be recorded as a force-extension curve with the characteristic saw-tooth pattern appearance. Except for the last peak, every peak in the force-extension corresponds to the mechanical unfolding of one protein domain in the polyprotein. The last peak represents the detachment of the polyprotein molecule either from the cantilever tip or from the substrate after it is fully extended. In the “fishing” mode, the polyprotein is picked up by the cantilever tip from the substrate randomly along its contour. Hence, the number of protein modules being stretched can vary in different recordings. In the constant velocity mode, the mechanical stability of a protein domain is measured by its unfolding force. Upon stretching, the end-to-end distance of the polyprotein increases which diminishes the entropy of the polyprotein. Therefore, the polyprotein molecule will generate an entropic restoring force in response to its end-to-end distance increase following the WLC model (from stage 1 to stage 2). When one domain is unfolded, the contour length of the whole polyprotein molecule will increase immediately and drastically and the force acting on the cantilever will drop to a low level simultaneously (stage 3). The force will increase again following the WLC model upon the stretching until another domain unfolds (stage 4). The red curve is the WLC fit to the polymer elasticity.

B) The schematic of the force-ramp mode of single molecule AFM measurements. The stretching force, \( F \), is monitored as a linear function of time, \( t \), (the bottom panel, \( F = at \), where \( a \) is the ramp rate and \( t \) is time). At the same time, the end-to-end distance of the polyprotein is also monitored as a function of time, shown in the middle panel. The mechanical unfolding of a polyprotein will result in a staircase like extension-time trace with a round corner. The unfolding of each individual domain in the polyprotein will result in a sudden contour length increment of the whole molecule and correspond to one stair step in the extension-time trace. Upon the unfolding of one protein domain, there will be a break in the linear increasing of the force and the force will first transiently relax to a low value and then quickly adjusted to the normal value which will leave a spike in the force-time trance. This spike is due to the response time of the electronic loop and the sampling frequency. The red curve is the WLC fit to the polymer elasticity.

C) The schematic of the force-clamp mode of single molecule AFM measurements. The stretching force, \( F \), is set as a constant force and the end-to-end distance of the polyprotein is monitored as a function of time. The mechanical unfolding of a polyprotein will result in a staircase like extension-time trace with a sharp corner. The unfolding of each individual domain in the polyprotein will result in a sudden contour length increment of the whole molecule and correspond to one stair step in the extension-time trace.
In most single molecule AFM experiments, the target protein domain of interest is constructed into a polyprotein. There are two ways to build the polyproteins. One way is to construct the polyprotein from the gene level through recombinant DNA techniques (Appendix A2). Polyprotein genes are first engineered by multiple-step cloning and the polyprotein is then expressed. The other way is to express the monomer of the protein of interest first. The expressed monomer domains are then chemically cross-linked via intermolecular disulfide bonds between the rationally designed cysteines located at different monomeric proteins (93, 94).

### 1.2.3 Why should we study the protein folding/unfolding mechanics?

#### 1.2.3.1 Is the protein folding/unfolding mechanics biologically relevant?

Since the single molecule AFM technique can only investigate the mechanical aspect of protein folding/unfolding problem, an immediate question will raise: Is the protein folding/unfolding mechanics biological relevant? The answer is simple, YES. A big proportion of the proteins in nature have mechanical functions. Mechanical processes are involved in almost every aspect of the cell metabolism and life cycle (29, 95-100). Force is either a product or an inducer of all physiological reactions mentioned above. Researchers are trying to directly apply external mechanical forces onto those processes to disturb the course or alter the kinetics or even the fate of these reactions, in order to determine their underlying molecular mechanisms. In all of these physiological processes, proteins are included as the key components without exception. The mechanical folding/unfolding of the corresponding protein is often necessary for the cellular machinery to fulfill their bio-functions. The obvious examples include muscle
proteins(101, 102), extra-cellular matrix (ECM) proteins(103), and cytoskeletal proteins(104). There are two major classes of proteins which are mechanically involved into the physiological reactions: “active” molecular motors and “passive” elastomeric proteins. The “active” molecular motors include ATP synthase, polymerases, myosin and kinesin, etc(105-109). Those molecular motors can automatically consume the ATP within the cell and convert the chemical energy stored in ATP into mechanical work. Significant conformational change or partial unfolding is typically accompanying the fulfillment of the bio-functions for the molecular motors. While for the “passive” elastomeric proteins, they function as molecular springs and passively bear the mechanical stresses imposed onto them. By undergoing conformation transitions or even unfolding, elastomeric proteins can dissipate the applied mechanical energy into heat, in which way the whole cellular machinery could be protected from breakage. When the external mechanical forces are removed, the “passive” elastomeric proteins can refold or return to their original conformations efficiently. In this way, the “passive” elastomeric proteins work as fundamental building blocks for cells and tissues to offer them mechano resistance, elasticity, and extensibility. The giant protein titin(110), ECM protein tenascin(111), fibronectin(112), spectrin(113), and elastin(114) are all well known elastomeric proteins.

Except for their indispensable physiological functions, mechano proteins, both “active” and “passive”, could also have promising applications in material science. Molecular motors have already been successfully adopted into nano-scaled mechanical devices(115, 116), showing the great potential of incorporating protein-based mechanical elements into nano-devices. Another example is resilin, an elastomeric protein found in
specific regions of the cuticle of most insects which plays critical roles in insect flight. Resilin has fantastic resilience properties and can repetitively store/release the energy with almost zero waste(117). A hydrogel based on resilin has been reported to have potential medical application(118). Such mechanical proteins have aroused tremendous interest from a nanoscience and nanotechnology point of view(115, 116). They are ideal building blocks for constructing nanomechanical devices from bottom-up and thus could have great potential in nanotechnology. With the development of nanotechnology, it is imaginable that more diverse protein-based functional components, such as molecular springs, switches, sensors and motors, will be incorporated into nano mechano-devices in the future. Besides those naturally occurring ones, the rationally designed and synthesized protein-based functional components can also be adopted to satisfy specific needs. The mechanical properties of these amazing mechanical proteins are ‘encoded’ in their primary sequences, three-dimensional structures as well as their unique arrangement from separated modules into organized complexes or materials.

1.2.3.2 The mechanical stability of a protein is different from its thermodynamic stability

The mechanical stability of a protein is the intrinsic property measured by its unfolding force. However, different from its thermodynamic stability, the mechanical unfolding force for a protein is not an invariable and it is only a kinetic stability. The mechanical stability of a protein is determined by the height of the mechanical unfolding energy barrier $\Delta G_{N-T}$ (the free energy difference between the native state and the mechanical transition state, Fig. 1.5) and the unfolding distance $\Delta x_u$ (distance between the
native state and the mechanical unfolding transition state)(85), while the thermodynamic stability of a protein is defined as the free energy difference between the unfolding state and the native state ($\Delta G_{N-U}$, Fig. 1.5). There is no directly correlation between the mechanical stability and the thermodynamic stability of a protein.

When free from the mechanical forces, the native state is energetically more favorable for most proteins. The dominant species is the folded protein under the normal experimental conditions which mimic physiological conditions. Typically, the spontaneous unfolding rate of the protein is negligible in such an environment. In order to study the protein folding/unfolding process, the protein must be artificially shifted away from its stable native state, first, to trigger the folding/unfolding dynamics. In this way, the energy landscape underlying will be tilted. The species which can affect the folding/unfolding energy landscape are defined as denaturants, such as guanidinium hydrochloride (GdmCl), urea, pH and heat, etc. Just like the other usual denaturant, the mechanical force can also be utilized as a “special” denaturant which directly acts on the protein 3D structure, one molecule a time. Single molecule AFM involves mechanical force as a “denaturant” to tilt the folding/unfolding energy landscape. The effect of the mechanical force on protein folding/unfolding is well described by the Bell-Evans model(119, 120):

$$k_u(F) = \alpha_0 \exp\left(\frac{F\Delta x_u}{k_B T}\right)$$

(1.2)

$$k_f(F) = \beta_0 \exp\left(-\frac{F\Delta x_f}{k_B T}\right)$$

(1.3)

where $F$ is the applied force, $\Delta x_u$ is the distance between the native state and the transition state (unfolding distance) and $\Delta x_f$ is the distance between the denatured state
and the transition state (folding distance). The mechanical unfolding and folding rate constants at zero force are represented as $\alpha_0$ and $\beta_0$, separately.

**Figure 1.5** The energy landscape of protein folding and unfolding in the two-state model. N, T and U stand for the native state, the transition state and the unfolded state of protein, respectively. $\Delta G_{N-T}$ reflects the height of the free energy barrier for protein unfolding reaction. Similarly, $\Delta G_{T-U}$ defines the kinetic energy barrier for protein folding. The difference between $\Delta G_{N-T}$ and $\Delta G_{T-U}$ is defined as $\Delta G_{N-U}$ which is the free energy difference between the native state and the denatured state. The thermodynamic stability of a protein is determined by its $\Delta G_{N-U}$ while the folding/unfolding kinetics of a protein depends on its $\Delta G_{N,T}/\Delta G_{T,U}$. $\Delta x_u$ and $\Delta x_f$ are the distances that the transition state is away from the native state and the unfolded state along the reaction coordinate, respectively.

The mechanical stability of a protein is regulated by its $\alpha_0$ and $\Delta x_u$. At a given pulling speed (a given unfolding rate), a higher unfolding free energy barrier and a smaller unfolding distance $\Delta x_u$ will lead to a higher average unfolding force. But the overall mechanical stability of a protein cannot be predicted based on $\alpha_0$ or $\Delta x_u$ alone. A
protein A with a smaller $\alpha_0$ is not necessarily more stable than a protein B with a larger $\alpha_0$ if their $\Delta x_0$ are different. Yet for a known protein with constant $\Delta x_0$, the faster the pulling speed is, the higher the average unfolding force will be. This was predicted by Evans(120, 121) and verified experimentally(92, 122). This behavior indicates that the mechanical unfolding of the protein is a non-equilibrium process. For the equilibrium process, the transition happens reversibly, meaning the rates of the transition for both directions are comparable. In other words, the free energy changes ($\Delta G$) for the transition of both directions are similar. However, in the case of protein mechanical unfolding, the refolding is completely inhibited by the force if the force is big enough to trigger the unfolding.

This non-equilibrium unfolding is mainly resulted from the relatively high pulling speed used in the experiments. Since the extension rate is fixed in the experiment, the molecule must elongate a corresponding distance within a certain time. This means the mechanical unfolding rate for the protein domains being stretched is also largely fixed. For example, given a pulling speed of 400nm$\cdot$s$^{-1}$, the unfolding rate of TNfn3 domain in the polyprotein $(T\text{Nfn3})_8$ will be $\sim$13 s$^{-1}$ (400 nms$^{-1}$/29.3 nm = 13.7 s$^{-1}$). The average unfolding force for TNfn3 is $\sim$120 pN and the folding rate at zero force ($\alpha_0$) is $\sim$1.1 s$^{-1}$(123). If the folding distance ($\Delta x_f$) of TNfn3 is similar to its unfolding distance ($\Delta x_u$) which is 0.42 nm (typically $\Delta x_f > \Delta x_u$), the folding rate is only $6.7 \times 10^{-6}$ s$^{-1}$ (calculated through equation (1.3)). The folding rate at 120 pN could be even smaller if a bigger $\Delta x_f$ is used. The artificially controlled unfolding rate is way faster than the folding rate of the protein under the unfolding force. The best known exception is the myosin II
coiled coil protein which can fold and refold in equilibrium under ~25 pN force at a pulling speed of 40nm/s – 130nm/s(124). But the myosin II coiled coil protein does not fold in the two-state mode and its secondary-structure-dominated simple topology makes it special among all of the proteins studied so far. For the other proteins with a modular, topologically more complex tertiary structure, their mechanical unfolding processes are all found as non-equilibrium processes obeying a two-state model (at the relatively high pulling speed used in real experiments)(24).

In principle, if the pulling speed is slow enough, it is possible that the folding rate and the unfolding rate can be equal or at least comparable. However, the drift of the cantilever and the piezoelectric positioner make those ultra-slow pulling speed experiments extremely challenging. For most of the current single molecule AFM instrumental settings, this ultra slow pulling speed are still unreachable. Therefore, the dominant proportion of the single molecule AFM experiments in the constant velocity mode is still performed under high pulling speed and the studied protein mechanical unfolding mechanics is still the non-equilibrium process. Due to the technical progress, it is only very recently that the ultra slow pulling speed single molecule AFM experiment became realistic. A very recent work has proved that the mechanical unfolding of calmodulin is reversible, at least very close to the equilibrium, under a very slow pulling speed when Ca$^{2+}$ is present(125).

For some proteins, their mechanical folding rate at zero force is found to be similar to the chemical folding rate with no denaturant(92). This consistency is more likely a coincidence since more other examples emphasized the differences between these two kinetics with different reaction coordinators(91, 122, 126-128). Even for a same
protein domain, the mechanical unfolding pathway is different upon the pulling from
different directions(129, 130). It is evident that there is no correlation between
mechanical kinetics and chemical kinetics(84, 85). Consequently, the mechanical stability
of a protein cannot be directly derived according to its available kinetic and
thermodynamic data previously measured in ensemble experiments.

1.2.3.3 A unique anisotropic property for mechanical unfolding of a protein

The thermodynamic stability of a protein is isotropic. However, the mechanical
stability of a protein is an anisotropic property. When stretched, the protein shows an
anisotropic deformation response to the force. The same protein domain can choose very
different unfolding pathways with drastically distinct energy barrier heights when it is
pulled from different directions. This anisotropic property first attracted people’s
attention when two inconsistent mechanical stabilities of the same protein ubiquitin were
reported by two groups independently(129, 131). When the force is applied to its N-/C-
termini of ubiquitin, ubiquitin will be unfolded at ~200 pN; when ubiquitin is torn from
its C-terminus and the residue Lys48, only ~80 pN is sufficient to unfold it. The
anisotropic nature of the protein mechanical unfolding reveals the ruggedness of the
underlying energy landscape and suggests the existence of multiple unfolding pathways.
Along different coordinates defined by the various directions of pulling forces, the
mechanical unfolding pathways are distinct. This characteristic has been utilized by
Rief’s group to explore the energy landscape underlining the mechanical unfolding of
green fluorescent protein (GFP)(132). They substituted the rationally chosen residues of
GFP into cysteine and designed a protocol to crosslink the single cysteine mutated GFP
domain into polyprotein via disulfide bond. In this way, they can apply the force onto the GFP domain in various directions. They successfully studied the anisotropic mechanical response of GFP to different pulling directions. Upon pulling GFP in different orientations, the average unfolding forces range from ~100 pN to over 600 pN. Based on their results, an energy landscape for the mechanical unfolding of GFP was constructed.

1.2.4 Accomplishments in the field of protein folding/unfolding study made by using single molecule AFM

1.2.4.1 The deviation from the two-state model of protein unfolding detected by using single molecule AFM

The majority of the known proteins studied in the traditional methods adopt a two-state folding/unfolding mode. However, in some cases, there is some indirect evidences indicating the existence of an intermediate state(s) along the protein folding/unfolding pathways, in which the protein might fold/unfold via more than one step(133, 134). The indirect observation of the intermediates can hardly reveal the structural information and formation mechanism of the intermediates.

Single molecule AFM is a perfect tool to study the folding/unfolding processes of proteins which might involve intermediates along the folding/unfolding pathways. As we know, the contour length increment ($\Delta L_C$) upon mechanically unfolding a protein domain can be precisely measured in single molecule AFM experiments. Since the $\Delta L_C$ upon an unfolding event reflects the number of amino acids that have been unraveled during the unfolding event, the length of $\Delta L_C$ can be used as a direct indicator of the intermediate state. If an unfolding intermediate exists along the mechanical unfolding pathway of a
protein, the unfolding of the whole domain will be split into two steps, which means only part of the amino acid sequence can be released from the folded structure in each step. Therefore, the $\Delta L_C$ for each step will be significantly shorter than that of the full unfolding of the whole protein domain. Since only one intermediate is involved, the sum of the two short $\Delta L_C$ should equal to the $\Delta L_C$ corresponding to the full unfolding of the whole protein domain. In this way, the existence of the mechanical unfolding intermediates can be easily identified and the structural information of the intermediates can also be extracted by combining the $\Delta L_C$ of the intermediate with the protein 3D structure. Some natural elastomeric proteins are found as three-state unfolders, such as FNfn10(135), I27 domain from titin(136), ubiquitin(91) and ddFLN4(137). The structure of those mechanical unfolding intermediates is typically found having a native-like structure — the partially folded region maintains most of the native interactions while the other region is unstructured and loses all of the native interactions. For example, I27 domain (the 27th Ig domain of human titin) unfolds in two steps when being stretched. Referring to its crystal structure, the three-state unfolding of I27 has been explained as following: the A-strand in I27 will be detached from the main structure of I27 first under the relatively low forces while the other part of I27 will be unfolded in its entirety under higher forces. To further validate their explanation, Fernandez and his co-workers adopted a contour length sensitive glycine insertion protocol(138). Since the $\Delta L_C$ upon the unfolding of a folded structure is sensitive to the number of the amino acids involved in the structure, the insertion of a certain number of glycine residues into the loops of I27 will lead to the $\Delta L_C$ alteration. The effect of the insertion is dependent on the location where the glycine residues are inserted. If the loop elongated by glycine insertion is
unfolded at the intermediate state, the $\Delta L_C$ upon the transition from native state to intermediate state ($\Delta L_{C1}$) will increase but the $\Delta L_C$ upon the unfolding of the intermediate ($\Delta L_{C2}$) will remain unchanged. On the contrary, if the inserted glycines are located in the folded structure of the intermediate, the $\Delta L_{C1}$ will be the same whereas the $\Delta L_{C2}$ will be longer. Fernandez and co-workers used this method to confirm that the structure for the unfolding intermediate state of I27 is native-like without the short $\beta$-strand A. This protocol has also been applied to explore the mechanical unfolding pathway for other proteins including ddFLN(137), GFP(139), and MBP(140).

Another protocol which introduces a disulfide bond to selectively block the unfolding of certain region of the target protein domain can potentially also be used to explore the structural information of intermediate states(141).

1.2.4.2 The detection of the hidden multiple unfolding pathways of the same protein

As predicted by the kinetic partitioning theory, the folding and unfolding processes are complex and there should be multiple folding/unfolding pathways underlying the protein folding/unfolding transition. However, before single molecule techniques arose, the existence of the multiple folding/unfolding pathways could not be directly detected in traditional ensemble experiments. There is only limited indirect experimental evidence showing parallel folding pathways(63-65). With the progress in single molecule detection and manipulation techniques, the direct observation of parallel folding/unfolding pathways on the same protein molecule becomes possible in recent. With the superb ability to trigger, record and characterize the mechanical
unfolding of individual protein molecules, single molecule AFM has proven its power in revealing the co-existence of multiple protein folding/unfolding pathways. One of the best examples is the bifurcation in the unfolding pathways of GFP revealed by Rief’s group using single molecule AFM(139). In the previous study on GFP from the same group, the anisotropy of the GFP mechanical unfolding has been revealed. The lifetime of the obligatory unfolding intermediate of GFP exhibited a multi-exponential distribution, which has been considered consistent with a simple reaction scheme with a single pathway (130). Combining computer simulations with protein engineering and single molecule AFM, the authors re-sampled the energy landscape of GFP mechanical unfolding. The complicated energy landscape underlying the GFP mechanical unfolding was constructed and, at least, two parallel mechanical unfolding pathways were identified. When GFP is mechanically unfolded from its native state, it will first reach an obligatory intermediate with the N-terminal α-helix ruptured. After that, the remaining structure can unfold along two distinct pathways: either from the N-terminus or from the C-terminus. The majority of the GFP molecules (~78%) will adopt the N-terminal pathway involving the two additional intermediates, while ~22% of the population will unfold via the C-terminal pathway with only one additional intermediate. By engineering cross-linked mutants (introducing intra-molecular disulfide bonds into selected regions of the GFP structure), the preference of the GFP molecules over the two mechanical unfolding pathways can be fine tuned. When the N-terminus of GFP is blocked by covalently linking two adjacent β-strands together via a disulfide bond, the whole molecule will be forced to mechanically unfold along the minor pathway. While if the C-terminus is blocked by a disulfide bond, GFP will arbitrarily unfold via the major pathway. These
results revealed that the energy landscape of GFP is not only rugged, but can also be manipulated.

However, because of the lack of refolding data, there is still an ambiguity. Based on the mechanical unfolding data alone, we cannot distinguish the following two scenarios: A) an individual GFP molecule can sample both major and minor pathways. For each independent unfolding process, the choice of the pathway is only determined by the probability; B) there are two sub-populations for GFP. Each one of the populations is restricted to only one particular pathway leaving the other pathway forbidden. To further elucidate this ambiguity, refolding experiments on a specific protein molecule will be needed. Furthermore, according to the kinetic partitioning theory, there should be abundant possible pathways for the folding/unfolding process of a protein. However, in the case of GFP, only two well defined pathways were identified. It looks like these results are still supporting the concept that there are clear and well-defined folding/unfolding pathways existing along the energy landscape, which is conflicting with the kinetic partitioning assumption. To further validate the kinetic partitioning theory experimentally, more experimental efforts are needed.

1.2.4.3 Studying the misfolding behavior of proteins and the mechanism related

Protein misfolding has been proposed as an inevitable phenomenon during the folding process(142). Especially, the very first and critical step in development of various conformational diseases, such as Parkinson’s disease and bovine spongiform encephalopathy in cattle, is the misfolding of protein. Thus, in order to prevent protein misfolding mediated pathologies with rational approaches, it is important to investigate
the properties of misfolded proteins and understand the mechanisms of their misfolding process. To achieve this, the rationale guiding the protein misfolding at the single molecule level is desirable.

The misfolding behavior of proteins can be studied by single molecule AFM at the single molecule level. The presence of the misfolding structure in a polyprotein can be identified by single molecule AFM in a straightforward fashion, given the misfolded protein exhibiting abnormal mechanical properties ($\Delta L_c$ or unfolding forces). Fernandez and coworkers detected rare misfolding events of adjacent protein domains upon repetitively stretching and relaxing polyprotein I27 and native tenascin using single-molecule AFM experiments(143). The I27 domain or FNIII domains in tenascin were found to misfold with one of its neighbouring domains into a dimeric fold with a contour length increment slightly longer than twice of $\Delta L_c$ of the single domains. The misfolding could be induced by the weak domain-domain interaction of the neighboring domains during the folding process which tilt the landscape of the protein. This finding provides important insight in the design of natural tandem modular elastomeric proteins. The evolutionary pressure to avoid misfolding might answer for the sequence diversity of different domains in a naturally occurring elastomeric protein with tandem arrangement. The frequency of the misfolding events is only ~2% in all unfolding traces. Such misfolding events can be easily averaged out in an ensemble experiment.

Another example using single molecule AFM to study the misfolding of proteins has been reported in 2005 by Daniel J. Muller’s group. Na$^+$/H$^+$ antiporter NhaA from Escherichia coli is a membrane protein. In the report, the researchers applied single molecule AFM to directly probe the stepwise folding of NhaA in vitro. The folding
kinetics of different structural segments of NhaA was investigated in detail. In most cases, the unfolded NhaA molecules can fold back into its native structure. Occasionally, the unfolding trace of refolded polypeptide showed an extra force peak comparing to the native NhaA, which indicates the unfolding of a distinct structure. This means that the folding chain can be kinetically trapped into stable and non-native structures, which could be assigned to misfolding events of NhaA. In principle, the misfolding can arise either within the membrane or at the interface of the membrane–water region. Considering that surface misfolding may affect the spontaneous folding of protein (into its native structure) and the following membrane insertion, this process could be assumed as a competition with the insertion of transmembrane domains. These observations demonstrate the ability of single molecule AFM to detect different folding states and behaviors of a membrane protein.

**1.2.4.4 Monitoring the protein folding process in real time**

So far I have discussed the mechanical unfolding studies of proteins using single molecule AFM, but the folding process of the protein can also be probed using the same method. The first endeavour was from Fernandez’s group who employed the force-clamp technique to explore the folding dynamics of the protein ubiquitin under low forces(88). In their experiment, a polyubiquitin molecule was first picked up and all the ubiquitin domains being stretched were unfolded at a high constant force. The N-/C- termini of the unfolded polyubiquitin were then quickly quenched to a low constant force and the end-to-end distance of the polyubiquitin was traced as the function of time. The clamping force was kept low for a period of time and increased to the high value again to unfold
the protein domains which succeeded in folding back during the low force time window. The unfolding events presented in the second high force time window confirmed that the individual ubiquitin domains can manage to fold back into their native state when clamped by low force. Therefore, the recorded trajectory during the low force time window indeed reflects the end-to-end distance fluctuation of the polyubiquitin during the folding. The complex pattern of the trajectories revealed rich information for the ubiquitin folding: four distinct stages can be identified during the folding and all ubiquitin domains fold in a continuous fashion instead of the predicted stepwise folding for individual domains. This observation is different from the traditional views of protein folding and new physical models are needed to explain these results. The amplitude of the clamping force is reverse correlated to the folding rate of ubiquitin. The higher the force clamping the unfolded ubiquitin, the longer the time is needed for ubiquitin to fold. When the force is higher than a certain threshold, the folding of ubiquitin will be prevented.

In addition to the force-clamp mode, the normal constant velocity mode can also probe the folding pathway of proteins. The first observation that the unfolded protein can generate force during refolding was reported by Bennett’s group and Marszalek’s group together in 2006 using single molecule AFM with constant velocity mode (144). As one of the most common amino-acid motifs, a structure of 24 ankyrin-B repeats was stretched and found to behave as a linear and fully reversible spring when it is stretched using single molecule AFM. During the relaxation, the ankyrin repeats are trying to refold and the clear refolding force peak of ankyrin was detected. The average mechanical refolding force for ankyrin can be as high as 32 pN. This is the first direct measurement of the
refolding force of a protein domain using single molecule AFM.

The mechanical folding process of ubiquitin has also been investigated using single molecule AFM with constant velocity mode(145). The mechanical folding of ubiquitin shows a simple two-state manner which is surprising considering its complicated mechanical unfolding behaviors. The mechanical folding mechanism of ubiquitin also demonstrates the isotropic response to the stretching forces as long as the force does not change the polypeptide elasticity. The folding rate and force are independent of the direction of the applied force. This is also in contrast with the mechanical unfolding of ubiquitin which exhibits obvious anisotropy(129).

In another study, Schwaiger et al. analyzed the folding mechanism of ddFLN4 domain which is relatively more complex. The ddFLN4 domain has been found to unfold through a stable mechanical unfolding intermediate state(137). By exploring its folding kinetics and folding pathway with a mechanical single-molecule protocol mimicking a double-jump stopped-flow experiment, Schwaiger et al. found there is also an obligatory and productive intermediate on the folding pathway of the domain(146). The refolding intermediates show identical mechanical properties as the unfolding intermediates, suggesting that they are closely related. The folding process of ddFLN4 is divided into two consecutive steps and this could be the reason why the overall folding rate of ddFLN4 is ten times faster than all other ddFLN homologous domains lacking of the folding intermediate.

In recent work, the mechanical folding/refolding of calmodulin was studied under a condition close to the equilibrium when its binding ligand Ca\(^{2+}\) is present(125). Using a custom-built low-drift AFM, the researchers successfully observed the force-induced
conformational equilibrium fluctuations of eukaryotic calmodulin at the single molecule level. The full energy landscape of mechanical folding/unfolding of calmodulin was constructed. They found that calcium ions can accelerate the folding kinetics of the individual calmodulin domains. It was also revealed that a wasp venom peptide can bind noncooperatively to full length calmodulin with 2:1 stoichiometry whereas a target enzyme peptide can bind to calmodulin cooperatively with 1:1 stoichiometry. The real-time binding transitions were detected when mechanical force is directly applied to the target peptide.

Most of the proteins studied by single molecule AFM are two-state folders. The best known example which deviates from the two-state folding pattern is the myosin II coiled coil protein. By manipulating the myosin II coiled coil protein with constant velocity mode under slow pulling speed (40nm/s – 130nm/s), Rief’s group revealed that the mechanically unfolded myosin II coiled coil protein can manage to refold under a burden of ~25 pN and the refolding is an equilibrium process (124). The refolding of myosin II happens gradually, instead of the all-or-none manner, with the contour length of the whole molecule shortening steadily while the loading force remains constant. All of the insights obtained from the studies discussed here would have been impossible without utilizing the single molecule force microscopy techniques. Single molecule AFM is definitely an ideal choice for such applications.
1.2.4.5 Revealing the insights on the energy landscape underlying the mechanical folding/unfolding of proteins using single molecule AFM

To characterize the energy landscape underlying the mechanical folding/unfolding process of a protein, the parameters in equations 1.2 and 1.3 must be extracted. To achieve a meaningful measurement of those parameters, repetitive independent observations on the same or different target molecules are necessary to build a representative data pool with a large number of data points.

To extract unfolding kinetic parameters from the data, Monte Carlo simulations or numeric fitting of the unfolding force distribution is often used(91, 92, 127, 147, 148). The Monte Carlo simulations are performed under exact experimental conditions(149). Given that the virtual polypeptide has a length of $L$ (nm) and is stretched at a certain pulling speed $a$ (nms$^{-1}$), we can stretch the very molecule in the simulation from the zero extension. In every time interval $t$ (s), the extension $x$ is increased by $at$, and the force acting on the polypeptide chain at the current extension is given by the WLC interpolation formula (equation 1.1). The unfolding rate $k_u$ is exponentially dependent on the acting force which is described by equation 1.2. Based on these two equations, we can generate the unfolding force histogram corresponding to any given pair of $\Delta x_u$ and $\alpha_0$. By generating a histogram which matches the experimental unfolding force histogram via a trial-and-error manner, the $\Delta x_u$ and $\alpha_0$ can be estimated.

In principle, the $\Delta x_u$ and $\alpha_0$ can be estimated accurately. Nevertheless, the data from a real experiment might be contaminated by many sources of noise, such as calibration error of the spring constant of the cantilever, instrument drift, different number of the domains in different recordings, temperative fluctuations, etc. These
factors will make the distribution of unfolding forces broader and affect the precision of
the estimated \( \Delta x_u \) and \( \alpha_0 \) values. Indeed, the \( \Delta x_u \) and \( \alpha_0 \) are not totally independent
parameters. For one, given an unfolding force histogram, the Monte Carlo simulation can
regenerate it with multiple combinations of \( \Delta x_u \) and \( \alpha_0 \). The degeneracy of these \( \Delta x_u \) and
\( \alpha_0 \) combinations cannot be removed if there is only one unfolding force histogram. In
order to estimate the \( \Delta x_u \) and \( \alpha_0 \) more accurately, we can perform the experiments under
the constant velocity mode with different pulling speeds. Given that the \( \Delta x_u \) and \( \alpha_0 \) are
invariable at different pulling speed, the “real” set of the \( \Delta x_u \) and \( \alpha_0 \) should be used to
successfully simulate the experimental results at all pulling speeds. Theoretically, the \( \Delta x_u \)
and \( \alpha_0 \) can be directly extracted from the pulling speed dependent experiments.

In most mechanical unfolding experiments on proteins, the experiments are
performed under a condition far away from the equilibrium. Usually, signatures of
folding reactions are not accessible in these experiments (91, 92, 127, 147, 148). Very
recently, an ultra slow pulling speed can be realized using an AFM setup with minimal
instrument drift(145). With such slow pulling speeds, the mechanical refolding of
proteins can be directly observed. Similarly to the way we extract the kinetic parameters
of mechanical unfolding, the unfolding force histogram from Monte Carlo simulations
can be used to compare to the experimental refolding force distribution to extract the
zero-force folding rate \( \beta_0 \).(125) For the folding distance \( \Delta x_f \) (the potential width for
folding), as shown in the case of equilibrium fluctuations of single calmodulin
molecules(125), it is not a free parameter but is given by the contour length of the
unfolded protein under the given force, minus the distance from the native state to the
transition state and the distance between the N-/C- termini in the native state. Thus, the $\Delta x_j$ can be calculated from the contour length increase and the force (125).

1.2.5 The combination of the single molecule AFM with computer simulations

Single molecule AFM experiments have provided new insights into protein mechanical folding/unfolding dynamics. However, it is difficult to work out detailed molecular mechanisms only from single molecule AFM experiments. Molecular dynamics simulations offer a powerful tool to explain the observed experimental phenomenon at an atomic level(150). The predictions made by molecular dynamics simulations could be used to define further single molecule AFM studies(151).

Different molecular dynamics simulation methods have been employed to investigate the mechanical unfolding of proteins(150, 152-168), in which the steered molecular dynamics simulation (SMD) has been most popular. In the SMD simulations, one terminus of the protein domain is fixed and a virtual force is applied to the other terminus to extend the protein. The mechanical responses of the protein to the force are then simulated and recorded. In many cases, the simulated mechanical unfolding process of a protein can explain the single molecule AFM results well. For example, a mechanically stable unfolding intermediate was observed in the SMD simulation of mechanical unfolding of the tenth FNIII domain in fibronectin (FNfn10)(150). This simulation result matches well with the single molecule AFM study on FNfn10 in which a mechanical unfolding intermediate with similar mechanical properties was also detected(135). The SMD simulation on mechanical unfolding of protein I27 also nicely explained the single molecule AFM data. An obligatory mechanical unfolding
intermediate state of I27 is observed in both SMD simulations(153) and the single molecule AFM experiments(136). The structure of the unfolding intermediate state of I27 predicted in the SMD simulations is very native-like with only the first several residues (β-strand A) extended. These predictions were largely verified by single molecule AFM experimental data(153).

Molecular dynamics simulations reveal that the mechanical stability of a protein domain is not an integral stability. The mechanical resistance of a protein (or the main energy barrier of the protein mechanical unfolding) is highly localized. The interactions within the force bearing region of the protein determine its mechanical stability(169, 170). Just like the catalytic center of an enzyme, the force bearing region a protein can be treated as a “mechanical active site”. The mechanical unfolding process of a protein is regulated by the breakage of its “mechanical active site”. Upon the disruption of the “mechanical active site”, the main resistance of the protein to forces dissappears and the protein will be unraveled. Although the mechanical resistance of a protein is highly localized, the stability of the “mechanical active site” is not an isolated issue. The “mechanical active site” is also stabilized by interactions with other regions of the protein(171).

One of the ultimate aims of the molecular dynamics simulation is trying to understand the general principles underlining the protein folding/unfolding. If all of the principles are known, the folding/unfolding process and characteristics of a protein with known structure should be fully predictable. Such efforts are underway and predictions based on current knowledge are proposed(164, 172). Some of the predictions match with actual results but some do not(164, 173, 174). However, due to the computational
constraints, the simulated biological time is limited within 1 μs, which is significantly shorter than most real biological processes taking place on time scales between 10 μs and 1 ms. Such limitation has strongly hindered the progress of SMD simulations. Recently, there is a landmark work reported by David E. Shaw’s group(175). They successfully simulated the equilibrium folding/unfolding of a small protein FiP35 for as long as 100 μs. They also simulated the dynamics of protein BPTI at its native state for 1 ms. This study is a big progress in the field of SMD simulations. Of course, there is still a long way to go before we can address the puzzle of protein folding/unfolding.

1.3 Objectives

In this thesis, I intend to use single molecule AFM to study the complex folding/unfolding behaviors of proteins, which can be either single domain proteins or complex proteins with multiple domains. My experimental efforts can be summarized into the following three aspects:

1) To study the mechanical unfolding pathways of two domain-insertion proteins.

One is a natural domain-insertion protein T4-lysozyme (T4L) and the other one is an engineered artificial domain-insertion protein GL5/T4L (GL5: a loop insertion mutant of protein GB1).

T4-lysozyme is a small protein with only 164 amino acid residues while it has two sub-domains with the N-terminal subdomain inserted into the sequence of the C-terminal subdomain. Our study on T4-lysozyme is the first investigation on the mechanical unfolding/folding of a natural domain inserted protein from its N-/C- termini using single molecule AFM. Having investigated the folding/unfolding mechanism of T4 lysozyme, I
will extend my study to an engineered artificial domain-insertion protein GL5/T4L to investigate its mechanical unfolding hierarchy.

2) To study the folding/unfolding kinetics and mechanics of the artificially designed mutually exclusive protein GL5/I27w34f.

As a natural continuation, I will discuss the design of a special domain-insertion protein -- a mutually exclusive protein GL5/I27w34f, and study its folding/unfolding kinetics. In this part, I will use a combination of experimental techniques, including single molecule AFM, circular dichroism (CD), fluorescence, stopped-flow, to investigate the complex folding/unfolding dynamics of this designed mutually exclusive protein.

3) To study the parallel mechanical folding/unfolding pathways of TNfn3 domain

TNfn3 is a small single domain protein with a normal $\beta$-sandwich structure. In the SMD simulation, multiple mechanical unfolding pathways have been predicted but the mechanical folding pathway is still left unstudied. The mechanical unfolding barrier of TNfn3 is proposed to be localized in its hydrophobic core and the mechanical folding of TNfn3 is predicted to be a simple two-state transition. To test those predictions and study the mechanical folding/unfolding mechanism of TNfn3, I performed single molecule AFM studies.
Chapter 2: Single molecule atomic force microscopy reveals kinetic partitioning of the mechanical unfolding pathway of T4-lysozyme*

2.1 Synopsis

Kinetic partitioning is predicted to be a general mechanism for proteins to fold into their well-defined native three-dimensional structure from unfolded states following multiple folding pathways. However, experimental evidence supporting this mechanism is still rare. Using single-molecule atomic force microscopy, we observed one of the first two direct evidences for the kinetic partitioning theorem. We observed that upon stretching from its N- and C-termini, T4-lysozyme unfolds via multiple distinct unfolding pathways: the majority of T4-lysozymes unfold in an all-or-none fashion by overcoming a dominant unfolding kinetic barrier; and a small fraction of T4-lysozymes unfold in three-state fashion involving unfolding intermediate states. The three-state unfolding pathways do not follow well-defined routes, instead they display great variability and diversity in their individual unfolding pathways. Our results also indicate that the coupling between the two subdomains, which is mediated by α-helix A, plays critical roles in defining the complex mechanical unfolding behavior of T4-lysozyme. These results provide direct evidence for the kinetic partitioning of the mechanical unfolding pathways of T4-lysozyme, and the complex unfolding behaviors of T4-lysozyme reflect

* A version of this chapter has been published as “[Peng Q.]. Li H. (2008) Single molecule atomic force microscopy reveals kinetic partitioning of the mechanical unfolding pathway of T4-lysozyme. Proceedings of the National Academy of Sciences, USA, 105(6):1885-1890”.
the stochastic nature of kinetic barrier rupture in mechanical unfolding processes. Our results demonstrate that single molecule AFM is an ideal tool to investigate the folding/unfolding dynamics of complex multi-module proteins that are otherwise difficult to study using traditional methods.

2.2 Introduction

The folding and unfolding of proteins are fundamental processes inside the cell. From an unfolded and presumably random coil-like state, a protein must fold into a well-defined three-dimensional structure, which is unique to each protein, to be biologically functional. The folding and unfolding processes are complex and may involve multiple pathways (18, 176), which are believed to be governed by the general kinetic partitioning mechanism (52, 177), although experimental proof supporting this mechanism is still limited (66, 178-180). Currently, most studies on protein folding/unfolding dynamics are using small, single domain proteins as model systems. In this chapter, we are going to explore the mechanical unfolding pathways of proteins with complex, multiple independent submodules, which are anticipated to present more complex folding/unfolding dynamics (60, 181, 182). The coupling between modules plays important roles in defining the overall conformational dynamics of these proteins (182-184). T4-lysozyme, a small natural protein adopting a domain-insertion arrangement, is an excellent model system in this aspect (185) (Fig. 2.1A). It has been widely studied for more than 30 years and the availability of high resolution structures of hundreds of T4-lysozyme mutants makes it especially appealing (186). Although traditional ensemble studies showed that T4-lysozyme unfolds in an apparent two-state
fashion(187-189), it has been well recognized that there exist two subdomains in T4-lysozyme: an α/β N-terminal subdomain and an all α C-terminal subdomain. The unique feature of T4-lysozyme is that the N-terminal helix A forms part of the C-terminal subdomain, resulting in the coupling between the two subdomains. The two subdomains have distinct thermodynamic stability giving rise to the possibility for T4-lysozyme to unfold from at least two regions(190). Recent fragment studies confirmed the partial independence of the subdomain architectures of T4-lysozyme(191-193). Here we use single molecule AFM techniques to investigate the unfolding dynamics and pathways mediated by the domain-insertion of T4-lysozyme.

Single molecule AFM exploits the stretching force as a “denaturant” to destabilize the folded state and force proteins to undergo a force-induced unfolding reaction along the reaction coordinate defined by the stretching force(86, 92, 132, 194). Single molecule AFM has evolved over the last decade into a powerful tool to investigate the folding and unfolding dynamics of proteins at the single molecule level and to provide a glimpse of the cross-section of the energy landscape underlying protein unfolding(86, 92, 124, 148, 195-197). Most of the single molecule AFM studies carried out to date focused on small single domain proteins(86, 124, 127, 195, 197-199), and many of them were considered as two-state folders in traditional ensemble studies. Despite the relative simplicity of these model systems, single molecule AFM revealed novel information about the conformational dynamics of proteins, such as stable unfolding intermediate states and distinct alternative unfolding pathways, which are invisible in ensemble studies(91, 130, 135-137, 200). The unique domain-insertion topology of T4-lysozyme makes it an ideal model system to investigate its mechanical unfolding behaviors using single molecule
AFM. Using a solid-state polymerized T4-lysozyme polyprotein, the mechanical unfolding of T4-lysozyme has been investigated in detail by stretching T4-lysozyme from its two residues 21 and 124(93). In this chapter, we use single molecule AFM to stretch the pseudo wild type T4-lysozyme protein (referred as to T4L hereafter) and its circular permutant PERM1 (201) from their N- and C-termini to investigate their mechanical unfolding dynamics and the role of domain-insertion of T4L in their mechanical unfolding. Our results showed that, upon pulling from its N- and C-termini, the mechanical unfolding of T4L displays kinetic partitioning and T4L unfolds via multiple unfolding pathways: the majority of T4L unfold in an apparent two-state fashion, while ~13% of T4L molecules unfold in three-state fashion involving partially unfolded intermediate states. In addition, the three-state unfolding pathways show great diversity in the individual unfolding trajectory. We demonstrate that the outside domain (the N-terminal helix A plus the C-terminal helix bundle) protect the inner domain. The internal interaction of the outside domain (between the N-terminal helix A and the rest of C-terminal subdomain) is critical for the mechanical stability of T4L. The unfolding intermediate states are kinetic traps along the mechanical unfolding pathway and are likely to result from the residual structures present in the two subdomains after the crossing of the main unfolding barrier. Such complex unfolding pathways reflect the stochastic nature of kinetic barrier rupture in mechanical unfolding processes, and provide direct evidence for the kinetic partitioning of the mechanical unfolding pathways of T4L. Our results demonstrate that single molecule AFM is an ideal tool to investigate the folding/unfolding dynamics of complex domain-insertion proteins that are otherwise difficult to study using traditional methods.
2.3 Results

2.3.1 T4-lysozyme exhibits distinct multiple unfolding pathways.

To characterize the mechanical unfolding of T4L using single molecule AFM, we engineered the (GB1)₄-T4L-(GB1)₄ polyprotein chimera, in which T4L was flanked by (GB1)₄ at both ends (Fig. 2.1B, upper panel). In the polyprotein chimera, the well-characterized GB1 domains serve as fingerprints for identifying single molecule stretching events as well as discerning the signatures of the mechanical unfolding of T4L. The mechanical unfolding of GB1 is characterized by a contour length increment \( \Delta L_C \) of ~18 nm and unfolding force of ~180 pN(122, 198). Stretching the polyprotein chimera (GB1)₄-T4L-(GB1)₄ resulted in force-extension curves with characteristic saw-tooth pattern appearance corresponding to the mechanical unfolding of (GB1)₄-T4L-(GB1)₄ (Fig. 2.1B). Curve A and B are two examples that show the stretching and unfolding of the full length polyprotein chimera. In these two curves, we observed eight unfolding events occurring at ~180 pN with \( \Delta L_C \) of ~18 nm as measured by fitting the Worm-Like-Chain (WLC) model of polymer elasticity(202) to consecutive unfolding force peaks. These eight unfolding events can easily be identified as the mechanical unfolding of the eight GB1 domains that flank T4L in the chimera. Hence, the unfolding event that occurs at ~50 pN and precedes the unfolding of GB1 domains must correspond to mechanical unfolding of T4L. Since the polyprotein chimera was picked up randomly along the contour length of the molecule, the majority of the force-extension curves correspond to the stretching and unfolding of part of the polyprotein. As T4L was flanked by (GB1)₄ at both ends, if we observed five or more unfolding events of GB1 in a force-extension curve, we are certain that the unfolding event prior to the unfolding
events of GB1 domains must correspond to the stretching and unraveling of T4L (Fig. 2.1B, curves c-d). Furthermore, to obtain these force-extension curves, the polyprotein must have been picked up by attaching the AFM tip to one of the GB1 domains and there should be no direct interaction between the AFM tip and T4L, avoiding any potential modification of native state of T4L by the AFM tip (also see Appendix B1). Fitting the WLC model to the unfolding events of T4L measures an average $\Delta L_C$ of 59.0±4.0 nm (avg.±S.D., n=1269) for the mechanical unfolding of T4L from its N- and C-termini (Fig. 2.1C).

T4L comprises 164 aa residues and is 59 nm long upon being unfolded and fully extended (164aa×0.36nm/aa). The distance between the N- and C-termini in the folded T4L is ~0.8 nm (PDB 1L63)(203). Hence, a complete mechanical unfolding of T4L will results in a $\Delta L_C$ of 58.2 nm (59.0 nm-0.8 nm), which is in close agreement with the experimentally determined $\Delta L_C$, corroborating our conclusion that the unfolding events of $\Delta L_C$ of ~59 nm correspond to the mechanical unfolding of T4L. The close match between the observed and predicted $\Delta L_C$ indicates that the unfolding events of T4L, within the force resolution of our AFM experiments (~20 pN) (see Appendix B2), correspond to the complete unfolding of T4L in an apparent all-or-none fashion.

In about 13% of the unfolding events of T4L, we observed that T4L domains unfolded in a complex three-state fashion involving unfolding intermediate state: instead of the all-or-none unfolding process (as those shown in Fig. 2.1B), the mechanical unfolding of some T4L occurred in two steps characterized by $\Delta L_{C1}$ and $\Delta L_{C2}$, respectively (Fig. 2.2A). Although different unfolding trajectories may show different patterns of $\Delta L_{C1}$ and $\Delta L_{C2}$, the sum of $\Delta L_{C1}$ and $\Delta L_{C2}$ gave a total $\Delta L_C$ of ~60 nm, which is
consistent with the complete unfolding of one folded T4L. These results indicate that T4L can unfold via multiple distinct pathways. For clarity, we will discuss the features of the two-state unfolding as well as three-state unfolding pathways separately.
Figure 2.1 The two state mechanical unfolding of T4-lysozyme. Majority of T4L molecules unfold in an apparent all-or-none fashion. A) Tertiary structure of WT* T4-lysozyme (PDB code: 1L63). T4L is 164 residues long and consists of 10 α-helices and four β-strands. The N-terminal sub-domain is colored in green, while the C-terminal sub-domain and the helix C are colored in yellow. Residues 1 and 164, from which T4L is being pulled to unfold, are shown in ball and sticks. Arrows indicate the force acting on T4L. B) Upper panel: schematic illustration of the polyprotein chimera (GB1)₄-T4L-(GB1)₄ used in single molecule AFM experiments. Lower panel: Typical force-extension curves of polyprotein (GB1)₄-T4L-(GB1)₄. The mechanical unfolding events of the well-characterized GB1 domains (colored in red) occurred at ~180 pN with ΔLₖ of ~18 nm and serves as fingerprints to identify single molecule stretching events and discern signatures of the mechanical unfolding of T4L. The unfolding of T4L always precedes the unfolding of GB1 domains and is characterized by unfolding forces of ~50 pN and ΔLₖ of ~60 nm, which corresponds to the complete unfolding of T4L (colored in green). Black broken lines correspond to WLC fits to the experimental data.
Figure 2.2 T4-lysozyme can unfold via multiple distinct three-state unfolding pathways. 
A) Typical force-extension curves of T4L showing three-state unfolding behaviors. The initial partial unfolding events, which correspond to the unfolding event from the native state to the unfolding intermediate state, are colored in blue. And the subsequent unfolding events, which correspond to the unfolding of the intermediate state to the fully unfolded state, are colored in cyan. WLC fits (black lines) to the experimental data reveal distinct patterns of $\Delta L_{C1}$ and $\Delta L_{C2}$. B) A series of force-extension curves of the same T4L measured during repeated stretching-relaxation experiments. The same T4L molecule exhibited distinct mechanical unfolding pathways, including the two-state pathway as well as multiple distinct three-state unfolding pathways. For all the events showing three-state unfolding behaviors, the sum of $\Delta L_{C1}$ and $\Delta L_{C2}$ is close to ~60 nm, which is in agreement with that expected from the complete unfolding of T4L.
2.3.2 Two-state unfolding of T4-lysozyme is characterized by a long unfolding distance to the transition state.

T4Ls that unfold in a two-state fashion show a narrow distribution in their unfolding forces (Fig. 2.1D) with an average of 50±13 pN (n = 1269) at a pulling speed of 400 nm/s. The unfolding force of T4L shows a weak dependency on different pulling speeds (Fig. 2.3): the unfolding force of T4L is 59 pN at a pulling speed of 1000 nm/s, which is similar to that when T4L is unraveled from its residues 21 and 124(93) and in good agreement with theoretical prediction(204). Using a standard Monte Carlo procedure(92), we fit the unfolding force histogram and the pulling speed dependence of the unfolding forces to estimate the unfolding rate constant at zero force $\alpha_0$ and the unfolding distance $\Delta x_u$ between the folded state and the transition state along the reaction coordinate. We found that the experimental data can be reproduced well using a $\alpha_0$ of 0.055 s$^{-1}$ and a $\Delta x_u$ of 0.75 nm. This result suggests that the mechanical resistance to unfolding is distributed over a distance of 0.75 nm, in contrast to the smaller $\Delta x_u$ and highly localized mechanical resistance observed for elastomeric proteins, such as I27(92) and ubiquitin(129). The measured $\Delta x_u$ is similar to the value reported by Yang et al.(93) for T4L being unraveled from residues 21 and 124.

The observed $\Delta L_C$ and $\Delta x_u$ for the two-state unfolding of T4L indicated that the mechanical unfolding barrier lies close to the N- and C-termini. Since helix A, which is at the N-terminus of the entire sequence of T4L, forms part of the C-terminal subdomain, the mechanical unfolding energy barrier for the two-state unfolding must correspond to breaking the interactions between helix A and the remainder of the C-terminal subdomain. Hence, the interactions between helix A and the remainder of C-terminal subdomain
constitute the main resistance to mechanical unfolding. After crossing the main unfolding barrier, the rest of T4L (N-terminal subdomain and the rest C-terminal subdomain) unfolds readily. This unfolding mechanism is the dominant unfolding pathway for T4L. However, to obtain detailed information about the molecular events during the unfolding process and pinpoint the exact location of the energy barrier, molecular dynamics simulations will be needed.

Figure 2.3 Unfolding forces of T4-lysozyme show a weak dependency on the pulling speed. Black solid line corresponds to Monte Carlo simulation results using an $\alpha_0$ of 0.055 s$^{-1}$ and a $\Delta x_u$ of 0.75 nm.

2.3.3 Three-state unfolding of T4-lysozyme shows diversity in unfolding pathways.

In contrast to the majority of T4Ls that unfold in a well-defined all-or-none fashion, about 13% of T4Ls unfold in three-state fashion involving an unfolding intermediate state. Such a three-state unfolding scheme does not follow a well-defined pathway, but instead showed great diversity and variability in their unfolding pathways.
Fig. 2.2A shows two force-extension curves in which T4L unfolds in two steps but with significantly different $\Delta L_{C1}$. For example, T4L in curve a) unfolds in two steps: the first step is a partial unfolding event of T4L resulting in $\Delta L_{C1}$ of 16 nm, and the second step corresponds to the subsequent unraveling of remainder of T4L resulting in $\Delta L_{C2}$ of 43 nm. The sum of $\Delta L_{C1}$ and $\Delta L_{C2}$ gave a total $\Delta L_{C}$ of ~60 nm, which corresponds to the $\Delta L_{C}$ resulted from the complete unfolding of one T4L. In contrast, the plot of the two-step unfolding of T4L results in b) $\Delta L_{C1}$ of 52 nm and $\Delta L_{C2}$ of 13 nm. Contour length increment upon unfolding is an intrinsic structural parameter that provides information about the location of kinetic barriers. The distinct patterns of contour length increments of the two unfolding events suggest the existence of distinct unfolding pathways as well as the different location of the kinetic barriers for unfolding.

The diversity in unfolding pathways for T4L was also observed in the same T4L molecule during repeated stretching-relaxation cycles. A few representative force-extension curves resulting from repeated stretching of the same polyprotein chimera are shown in Fig. 2.2B, where the T4L molecule exhibits both three-state and two-state unfolding behaviors. The choice of a given unfolding pathway is not deterministic, reflecting the stochastic nature of kinetic barrier crossing(205). For the events showing three-state unfolding behavior, all the T4L molecules show a combined $\Delta L_{C}$ of ~60 nm, while $\Delta L_{C1}$ and $\Delta L_{C2}$ showed different patterns. This observation indicates that the different unfolding pathways, both two-state unfolding and the diverse three-state unfolding schemes, are intrinsic properties of T4L defined by its underlying energy landscape and therefore the same T4L molecule can sample distinct unfolding pathways.
The diversity of the three-state unfolding pathways is further illustrated by the histogram of $\Delta L_{C1}$ and $\Delta L_{C2}$ compiled from 196 T4L molecules that unfold in three-state fashion (Fig. 2.4A and 2.4B). For comparison, the corresponding histogram for $\Delta L_{C1} + \Delta L_{C2}$ is also shown (Fig. 2.4C). It is clear that $\Delta L_{C1}$ and $\Delta L_{C2}$ show broad distributions that are clearly beyond the experimental error of the $\Delta L_C$ measurements in our experiments. Therefore, the broadness in the distribution of $\Delta L_{C1}$ and $\Delta L_{C2}$ reflects the intrinsic diversity of the three-state unfolding schemes manifested by T4L molecules. From limited three-state unfolding events, it seems that there exist three preferred pathways with $\Delta L_{C2}$ equals to $\sim 14$nm, $\sim 29$nm and $\sim 46$nm (as indicated by arrows), respectively, suggesting that the second kinetic energy barrier for the three-state unfolding is located roughly 45nm, 30nm and 13nm away from its resting length between N- and C-termini. However, due to the uncertainties of using the WLC model to fit the low unfolding force events of T4L, our resolution in $\Delta L_C$ measurements is rather limited (as indicated by the relatively large bin size of 3nm in Fig. 2.4A-C). Such a limited resolution made it difficult to accurately determine the distribution of unfolding pathways.

Despite the diversity in the three-state unfolding pathways, the unfolding forces of the two unfolding force peaks are surprisingly similar to each other, as well as to that for two-state unfolding. The unfolding forces for the two unfolding force peaks show a narrow distribution with an average force of $\sim 50$ pN (Fig. 2.4D and 2.3E).
Figure 2.4 Histogram of $\Delta L_{C1}$ (A) and $\Delta L_{C2}$ (B) during three-state unfolding trajectories. For comparison, the histogram of $\Delta L_{C1}+\Delta L_{C2}$, which was measured from the same unfolding trajectory of T4-lysozyme, is shown in C). It is evident that $\Delta L_{C1}$ and $\Delta L_{C2}$ show broad distribution, indicating broad distribution of unfolding pathways for T4L. The solid line in C) is the Gaussian fit to the experimental data with an average $\Delta L_{C1}+\Delta L_{C2}$ of 61.0±4.0 nm (n=196). D) and E) Unfolding force histograms of the first and second unfolding events of T4L observed in three-state unfolding pathways.

2.3.4 Possible pathways for three-state unfolding

T4L is made of two subdomains: the N-terminal subdomain is inserted into the middle of the C-terminal domain. Various intermediate states have been observed in chemical denaturation studies, along both folding and unfolding pathways(192, 206-210). However, in the mechanical unfolding experiments, the unfolding of T4L proceeds along
a well defined reaction coordinate set by the stretching force, which is quite different from that in chemical folding/unfolding studies. Hence, it is unlikely that the mechanical unfolding intermediates observed here directly correspond to those observed in the chemical folding/unfolding studies. Indeed, our single-molecule AFM studies show that mechanical unfolding pathways exhibit features that are quite different from that for the chemical unfolding pathway. The structure of T4L is unique in that the N-terminal subdomain is inserted into the C-terminal subdomain. Therefore, upon stretching T4L from its N- and C-termini, the C-terminal subdomain must be first unraveled: the helix A must be detached first from the remainder of C-terminal subdomain in order to extend T4L. Hence, regardless of the two-state or three-state unfolding pathways, disruption of the interaction between helix A and the remainder of the C-terminal subdomain constitutes the first barrier for mechanical unfolding of T4L. In this way, the C-terminal subdomain behaves as a switch to protect the N-terminal subdomain. The unfolding of the C-terminal subdomain is the pre-condition for the observation of the unfolding of the N-terminal subdomain. In contrast, the unfolding of the N-terminal subdomain occurs before the unfolding of the C-terminal subdomain in the chemical unfolding pathways of T4L.

In the mechanical unfolding of T4L, the observation of multiple three-state unfolding pathways indicates that, after crossing the first barrier, there are many potential contacts, being native or newly formed in the two subdomains, that can lead to local energy minima and potentially provide mechanical resistance to unfolding in a stochastic fashion. By using $\Delta L_C$ as a probe, it is possible to estimate the location of the kinetic barrier during the three-state unfolding(130). For example, the unfolding pathway of $\Delta L_{C1}$
and $\Delta L_{C2}$ of 30/29 nm is consistent with the unfolding of helix A plus the N-terminal subdomain followed by subsequent unraveling of the remaining C-terminal subdomain. However, because of the broad distribution of $\Delta L_{C1}$ and $\Delta L_{C2}$ and the degeneracy in $\Delta L_{C1}/\Delta L_{C2}$ due to the fact that different unfolding pathways may result in similar $\Delta L_{C1}/\Delta L_{C2}$, it remains difficult to accurately determine the location of kinetic barrier in three-state unfolding processes. Extended AFM studies with much improved resolution in $\Delta L_C$ measurements, in combination with complementary structural characterization experiments, will be required to characterize the unfolding intermediate states in detail.

2.3.5 Interactions between helix A and reminders of the C-terminal subdomain play a critical role in determining the mechanical stability of T4-lysozyme.

Our results indicated that the interaction between helix A with the C-terminal subdomain may hold the key to the mechanical stability of T4L. To further investigate the role of helix A on the mechanical unfolding of T4L, we use single molecule AFM to study the mechanical unfolding of a well-characterized T4L circular permutant PERM1. In PERM1, helix A is shuffled to the C-terminus of the overall sequence and it is no longer directly connected to the N-terminal subdomain, thus the two subdomains are decoupled(201). PERM1 was shown to have almost identical structure(201) with the wild type T4L (Fig. 2.5B), while the two subdomains are now arranged tandemly in sequence without the domain-insertion feature.

Using a similar strategy as for T4L, we characterized the mechanical unfolding of PERM1. If we observed five or more unfolding events of GB1 domains, we are certain that the given force-extension curve contains the signature of mechanical unfolding of
PERM1. In contrast to WT* T4L, the unfolding of PERM1 displays even greater diversity (Fig. 2.5B): about 20% of the force-extension curves (n=190) did not show any unfolding peak corresponding to the unfolding of PERM1, indicating that these PERM1 molecules unfold at forces below our detection limit (~20 pN); ~56% of force-extension curves showed a single unfolding force peak with a broad distribution of $\Delta L_C$ from 20 nm to 60 nm (Fig. 2.5C); and ~24% of unfold events showed a three-state unfolding (Fig. 2.5D).

In those unfolding events that only display one unfolding force peak, the majority of the unfolding events have a $\Delta L_C$ of ~60 nm, corresponding to the two-state complete unfolding of PERM1. However, 54 out of 107 molecules show a $\Delta L_C$ that is much smaller than 60 nm ($\leq$50 nm), suggesting that part of PERM1 unfolded at low forces prior to the observed unfolding event (Fig. 2.5C).

Compared with T4L, there is a significant increase in the number of PERM1 domains that unfold at forces below 20 pN, as well as the number of PERM1 domains that unfold from an already partially unfolded intermediate state. These results strongly indicate that the shuffling of helix A from the N-terminus to the C-terminus significantly weakens T4L. Although the two subdomains in the circular permutant PERM1 remain relatively intact, the coupling of the two subdomains by helix A is no longer present. And the mechanical unfolding barrier, which corresponds to the disruption of helix A from the rest of C-terminal lobe, is no longer the dominant unfolding barrier. Therefore, the C-terminal subdomain cannot protect the N-terminal subdomain any more and the N-terminal subdomain will directly face the external force in the structure of PERM1. The exposure of the labile N-terminal subdomain shifted the lowest mechanical unfolding barrier to the N-terminal subdomain and made it easier to unfold PERM1.
Similar to the mechanical unfolding of T4L, a significant percentage of PERM1 unfold in three-state fashion with a broad distribution of $\Delta L_C$. Together with the observation that the unfolding of many PERM1 only show partial unfolding, we conclude that many local interactions in T4L may serve as kinetic traps providing resistance to mechanical unfolding. These local traps are not well-defined, but are rather present in a statistical manner. These results highlight the critical importance of the protection effect of the outer C-terminal subdomain on the inner N-terminal subdomain when the two subdomains are in the domain-insertion arrangement or are, say, coupled. The coupling of helix A with the C-terminal domain is the key to the mechanical stability of T4L. This observation is consistent with the recent observation that retaining the domain-insertion arrangement (coupling helix A with C-terminal subdomain) is the key to the overall thermodynamic stability of T4L.(192, 211).
Figure 2.5 Mechanical unfolding behaviors of T4-lysozyme circular permutant PERM1. The sequence and three dimensional structure of PERM1 are shown in A) and B). In the circular permutant PERM1, the sequence of helix A is relocated to the C-terminus of the
whole sequence, and the two subdomains are decoupled. For clarity, the N-terminal subdomain is colored green, and the C-terminal subdomain is colored yellow. The new N- and C-termini, from which PERM1 is pulled to unfold, are shown in red and ball-and-stick representation. C) Representative force-extension curves of polyprotein chimera (GB1)$_4$-PERM1-(GB1)$_4$. The mechanical unfolding of PERM1 shows diverse unfolding behaviors. PERM1 in curve a) did not show clear unfolding force peaks, indicating that they unfold at low forces below our detection limit. PERM1 in curves b)-c) correspond to the two-state unfolding of fully folded or partially folded PERM1. Curve d) shows a three-state unfolding event of PERM1. D) Histogram of $\Delta L_C$ for the PERM1 molecules that unfold in two-state fashion (n=107). E) Histogram of $\Delta L_{C1}$ and $\Delta L_{C2}$ for PERM1 that unfold in three-state fashion (n=45).

2.4 Discussion

2.4.1 Kinetic partitioning in the mechanical unfolding of T4-lysozyme.

The statistical mechanics description of protein folding has provided tremendous insights into the protein folding mechanism(18, 212). Although there is only limited experimental data(66, 178, 179), it is now generally accepted that proteins can fold into their well-defined native structure following multiple folding pathways via kinetic partitioning mechanisms(52). Recent developments in single molecule techniques have made it possible to directly probe the folding and unfolding trajectories of proteins at the single molecule level, providing the possibility to directly probe the kinetic partitioning of folding and unfolding pathways. Here, our single molecule mechanical unfolding trajectories revealed that T4L unfolds via multiple distinct unfolding pathways: 87% of T4Ls unfold in an all-or-none fashion involving overcoming a dominant unfolding kinetic barrier, and ~13% of T4Ls unfold in three-state fashion and exhibit variability and diversity in their individual unfolding pathways. Similarly, during refolding experiments, 87% of T4Ls fold completely into their native conformations, while ~13% of molecules got trapped in partially folded states along their folding pathways. These mechanical
unfolding trajectories of T4L provide direct evidence for the kinetic partitioning of the mechanical unfolding pathways of T4L.

2.4.2 Possible mechanisms for kinetic partitioning

What is the origin for the kinetic heterogeneity observed for the mechanical unfolding of T4L? Two possible scenarios are feasible to explain the kinetic partitioning observed for the mechanical unfolding/folding of T4L. In the first scenario, there is one well-defined native conformation for T4L. The frustration on the free energy landscape results in direct (two-state unfolding) and indirect pathways to the unfolded state, leading to the kinetic partitioning observed here for the mechanical unfolding pathways. Along the indirect pathways, local energy minima can kinetically trap T4L into the various unfolding intermediate states, which are stable enough to be captured in our single molecule unfolding trajectories.

The second scenario is associated with the heterogeneity of the native conformations of T4L. Although there is no direct experimental evidence, it was suggested that the N-terminal domain of T4L probably undergoes hinge-bending motion in solution(213), which could potentially lead to more than one native conformations. If this scenario holds, the observed heterogeneity of the mechanical unfolding pathways of T4L could be explained by the unfolding from distinct native conformations of T4L. In addition, this model could also provide direct evidence for the kinetic partitioning mechanism for the folding pathways of T4L: if we assume the existence of multiple native conformations of T4L, the repetitive unfolding-refolding experiments shown in Fig. 2.2B would indicate that T4L, from the same unfolded and extended conformation,
folded into different native conformations following distinct folding pathways. Considering that there is no direct experimental evidence for the heterogeneity of native conformations for T4L, we think that the kinetic partitioning observed here is more likely to originate from the first mechanism. Nevertheless, the common theme of both scenarios is the frustration in the free energy landscape, which is the key to the kinetic partitioning mechanism. This result is also similar to the observations of different folding pathways for a RNA hairpin in optical tweezers experiments(214).

The broad distribution of unfolding pathways suggests that the three-state unfolding behavior is not producing well-defined unfolding intermediates. Instead, there are many local interactions/contacts in T4L that can trap the protein into an unfolding intermediate state along its unfolding pathway. Such complex multiple unfolding pathways reported here have much resemblance to that of the mechanical unfolding of *T. thermophila* ribozyme, where a wealth of unfolding and folding pathways observed in optical tweezers experiments are predominantly determined by local interactions(205, 214). This similarity between T4L and ribozymes provides good supporting evidence for the proposal that the kinetic partitioning mechanism is a common theme in the folding of proteins and RNAs(51, 52). Moreover, the complexity in the mechanical unfolding pathways of T4L is similar to the complexity of the folding/unfolding intermediate states observed in chemical folding and unfolding studies of T4L. During chemical unfolding/folding processes, folded/unfolded intermediate states were detected and reported with different structures(206, 209, 210, 215). It remains to be seen whether any of the unfolding intermediate states observed in single molecule AFM trajectories share
similar structural features with the intermediate species observed in chemical unfolding/folding studies.

The rich yet complex kinetic behaviors of T4L reflect the intrinsic properties of the underlying energy landscape. It is likely that the unique two subdomain structure of T4L and the associated domain-domain interactions play important roles in defining such a complex energy landscape that gives rise to the complex unfolding kinetics of T4L. Although T4L displays two-state unfolding behaviors probed by traditional methods, it has been well recognized that there exist two subdomains in T4L that have distinct thermodynamic stability and show possibilities of unfolding from at least two regions (190). Recent fragment studies confirmed the subdomain architecture of T4L (192, 211). In addition, a continuum of stability was observed by native state hydrogen exchange to occur throughout each subdomain which may give rise to a variety of folding and unfolding intermediate states (206, 209, 210, 215). In the mechanical unfolding of T4L, we observed that the domain-insertion arrangement of T4L is the reason for its major mechanical resistance. The interaction between the N-terminal helix A and the remainder of the C-terminal subdomain and the coupling between the two subdomains via helix A provide the dominant resistance to the mechanical unfolding force. Through this coupling, the outer C-terminal subdomain effectively protects the inserted N-terminal domain. The various unfolding intermediate states observed in three-state unfolding correspond to various partially unfolded structures of the two subdomains after the main barrier crossing event. These results highlight the critical importance of the domain-domain coupling resulting from the domain-insertion arrangement, as well as their interactions on the folding/unfolding kinetics of a multi-domain protein. We
anticipate that single molecule mechanical manipulation exemplified here will provide a
general tool and strategy to thoroughly investigate the mechanical unfolding kinetics of
complex domain-insertion proteins.

2.5 Experimental section

2.5.1 Protein engineering

The plasmids that encode the pseudo wild type T4-lysozyme protein (T4L) and its
circular permutant PERM1 were generous gifts from Prof. Brian W. Matthews of
University of Oregon and Prof. Martin Sagermann of UCSB. T4L and PERM1, flanked
with a 5’ BamHI restriction site and 3’ BgIII, KpnI restriction sites, were amplified by the
polymerase chain reaction, respectively. The genes of polyprotein chimera
pQE80L/(GB1)₄-T4L-(GB1)₄ and pQE80L/(GB1)₄-PERM1-(GB1)₄ were constructed by
using a previously described method(92) based on the identity of the sticky ends
generated by the BamHI and BgIII restriction enzymes (also see in the Appendix A2).
Polyproteins were overexpressed in DH5α strain and purified from supernatant using
Ni²⁺-affinity chromatography. The polyproteins were kept at 4°C in PBS buffer at a
concentration of ~200 μg/mL.

2.5.2 Single molecule atomic force microscopy

Single-molecule AFM experiments were carried out on a custom built atomic
force microscope, which was constructed as described(88). All the force-extension
measurements were carried out in PBS buffer. In a typical experiment, the polyprotein
sample (1 μL) was deposited onto a clean glass cover slip covered by PBS buffer (50 μL)
and was allowed to adsorb for approximately 5 minutes prior to the stretching experiments. The spring constant of each individual cantilever (Si$_3$N$_4$ cantilevers from Vecco, with a typical spring constant of 40 pNnm$^{-1}$) was calibrated in solution using the equipartition theorem before and after each experiment.
Chapter 3: Domain insertion effectively regulates the mechanical unfolding hierarchy of elastomeric proteins: towards engineering multi-functional elastomeric proteins

A version of Chapter 3 has been published as “[Peng Q.], Li H. (2009) Domain insertion effectively regulates the mechanical unfolding hierarchy of elastomeric proteins: towards engineering multi-functional elastomeric proteins. *Journal of the American Chemical Society*, 131(39):14050-6”.

DOI: 10.1021/ja903589t

According to the copyright policy of American Chemical Society (ACS), the republishing of ACS full articles in a thesis/dissertation on a website is not permissable. Therefore, Chapter 3 has been removed when the thesis is published on-line.

As an alternative, the link to the article's DOI is provided as following:

http://pubs.acs.org/doi/full/10.1021/ja903589t

* A version of this chapter has been published as “[Peng Q.], Li H. (2009) Domain insertion effectively regulates the mechanical unfolding hierarchy of elastomeric proteins: towards engineering multi-functional elastomeric proteins. *Journal of the American Chemical Society*, 131(39):14050-6”.
Chapter 4: Mechanical design of the third FnIII domain of tenascin-C*

4.1 Synopsis

By combining single molecule atomic force microscopy (AFM), proline mutagenesis and steered molecular dynamics (SMD) simulations, we investigated the mechanical unfolding dynamics and mechanical design of the third fibronectin type III domain of tenascin-C (TNfn3) in detail. We found that the mechanical stability of TNfn3 is similar to that of other constituting FnIII domains of tenascin-C, and that unfolding of TNfn3 occurs via an apparent two-state process. By employing proline mutagenesis to block the formation of backbone hydrogen bonds and introduce structural disruption in the β sheets, we revealed that in addition to the important role that the hydrophobic core packing plays in determining the mechanical stability of TNfn3, backbone hydrogen bonds in β hairpins are also responsible for the overall mechanical stability of TNfn3. Furthermore, proline mutagenesis revealed that the mechanical design of TNfn3 is robust and that the mechanical stability of TNfn3 is very resistant to structural disruptions caused by proline substitutions in the β sheets. Proline mutant F88P is one exception, as the proline mutation at residue 88 significantly reduced the mechanical stability of TNfn3 and led to unfolding forces below 20 pN. This result suggests that residue Phe88 is a weak point in the mechanical resistance for TNfn3. We also used SMD simulations to understand the molecular details underlying the mechanical unfolding of TNfn3.

*A version of this chapter has been published as “[Peng Q.], Zhuang S., Wang M., Cao Y., Khor Y., Li H. (2009) Mechanical design of the third FnIII domain of Tenascin-C. Journal of Molecular Biology, 386(5):1327-1342”.*
comparison between the AFM results and SMD simulations revealed similarities and discrepancies between the two. We also compared the mechanical unfolding and design of TNfn3 and its structural homologue, the tenth FnIII domain from fibronectin. These results revealed the complexity underlying the mechanical design of FnIII domains and will serve as a starting point for systematically analyzing the mechanical architecture of other FnIII domains in tenascins-C, and also help to gain a better understanding of some of the complex features observed for the stretching of native tenascin-C.

4.2 Introduction

In the chapter 2 & 3, we have investigated the mechanical unfolding of two domain-insertion proteins. For those domain-insertion proteins with multiple subdomains, their mechanical unfolding is typically complicated and involves multi-step unfolding or multiple unfolding pathways. However, most natural elastomeric proteins are composed of individual modular proteins with a single domain. The mechanical folding/unfolding of those natural elastomeric proteins is often necessary for the cellular machinery to fulfill their biological functions (101-104). Understanding the mechanism of the mechanical folding/unfolding of natural elastomeric proteins is therefore important for elucidating the related biological processes.

One of the fundamental functions involving the cellular elastomeric proteins is mechanotransduction. When an external mechanical force acts on cells, cells will respond to the force signal and trigger a variety of downstream processes (237-239). The extracellular matrix (ECM) proteins have been found play a critical role in cellular mechanotransduction. In living tissues, mechanical force is transmitted to cells through the
extracellular matrix (ECM), which serves as a mechanical scaffold for cells to adhere, migrate and differentiate\(^1\). ECM is linked to the intracellular cytoskeleton through the cell membrane receptors integrins to establish a mechanical continuum allowing the mechanical force be transmitted as a physiological signal between the interior and exterior of cells. A wide variety of ECM proteins are subject to mechanical tension in biological environments and many of them share a similar tandem modular architecture\(^1\). Mechanical tension not only alters the conformational states of such mechanical proteins, but may also modulate biological functions via force-modulated conformational changes. Therefore, understanding the relationship between the structure and mechanical properties of these proteins will be very important to understanding the underlying biology. Tenascin-C, a highly conserved oligomeric ECM glycoprotein (240-243), is one of the model systems for such studies.

Tenascin-C is an ECM protein and plays important roles in regulating the cell-matrix interactions\(^2\). Tenascin-C is a tandem modular protein and consists of a tenascin-assembly domain, a stretch of epidermal-growth-factor (EGF) like repeats, a fibronectin type III (FnIII) domain region that is composed of a series of FnIII domains, and a terminal knob domain that is homologous to the globular domain of fibrinogen. Tenascins are mainly expressed in regions that are subject to heavy tensile load\(^2\) or in tissues that undergo extensive structural re-modeling during pathological states, such as tissue injury and tumorigenesis (245-247). As tenascins are subject to mechanical stretching forces under physiological conditions, it is possible that the force-induced unfolding/refolding reactions of FnIII domains may be an important part of tenascins dynamics in vivo\(^4\).
Single molecule atomic force microscopy studies have provided insights into the mechanical design and functions of tenascin-C. It was revealed that tenascin-C is an elastic protein that can extend to several times its resting length via force-induced unfolding of FnIII domains(148, 248). It has been suggested that the mechanical unfolding of FnIII domains serves as a shock-absorber to prolong the lifetime of a tenascin-ligand bond(148). Recent studies revealed that some constituting FnIII domains display weakly populated folded microstates in addition to their native states, which may entail a possible mechanism for these FnIII domains to recover their mechanical resistance more rapidly after mechanical unfolding(249). Since these studies were carried out on native fragments of tenascin-C, the inherent heterogeneity of the constituting FnIII domains in native fragments makes it difficult to assign mechanical features observed on native tenascin fragments to specific FnIII domains, and makes it difficult to obtain a detailed molecular interpretation of the experimental results.

To overcome these difficulties, detailed studies of the mechanical unfolding dynamics of individual FnIII domains become necessary. Using polyproteins made of identical tandem repeats of the protein of interest has become the standard approach to study their mechanical properties using single molecule AFM(74, 92). Recently, mechanical φ value analysis has been carried out on the third FnIII domain of tenascin-C (TNfn3) via single molecule AFM and molecular dynamics simulations using an implicit water model(250). However, this study was only focused on probing the role of the hydrophobic core in the mechanical unfolding of TNfn3. The role of backbone hydrogen bonds and β sheet stability, two important factors for mechanical stability of proteins, were not probed. To address the importance of backbone hydrogen bonds as well as β
sheet stability in the mechanical unfolding of TNfn3, here we combine single molecule AFM, proline mutagenesis and steered molecular dynamics (SMD) simulations to investigate the mechanical unfolding dynamics and mechanical design of TNfn3 in detail. Our results revealed that the mechanical stability of TNfn3 is similar to that of other constituting FnIII domains of tenascin-C(148, 249), and the unfolding process of TNfn3 is an apparent two-state process. To probe the mechanical design of TNfn3, we employed proline mutagenesis to block the formation of backbone hydrogen bonds and introduce structural disruption in the β sheet in order to affect its mechanical stability and unfolding kinetics. Our results revealed that not only the hydrophobic core packing plays important roles in determining the mechanical stability of TNfn3, backbone hydrogen bonds in the β hairpins are also responsible for the overall mechanical stability of TNfn3. Furthermore, proline mutagenesis revealed the robust mechanical design of TNfn3, as the mechanical stability of TNfn3 is resistant to disruptive proline mutations in the β sheets of TNfn3. We also identified that residue Phe88 is a weak point in TNfn3 and a single substitution of Phe88 with a proline results in the unfolding of TNfn3 at forces that are lower than the detection limit in AFM. We also compared the AFM results with SMD simulations to understand the molecular details underlying the mechanical unfolding of TNfn3. A comparison between the mechanical features of TNfn3 with the tenth FnIII domain from fibronectin also revealed the significant differences in mechanical unfolding and design of these two structurally homologous FnIII domains. These results pave the way for systematically analyzing the mechanical architecture of other FnIII domains in tenascin-C and will help to gain a better understanding of some of the complex features observed for the stretching of native tenascin-C.
4.3 Results

4.3.1 The mechanical unfolding of TNfn3 is an apparent two-state process.

TNfn3 is an all-beta protein of 90aa residues. It has a typical immunoglobulin-like β-sandwich structure, in which the two β sheets of TNfn3 pack against each other (Fig. 4.1). The top sheet consists of strands A-B-E and the bottom sheet consists of C-C’-F-G. The two force-bearing terminal β strands A and G are parallel to each other and the N- and C-termini are pointing in opposite directions. Such an arrangement of the two terminal strands forms a shear topology upon stretching, which is a common feature among proteins that are mechanically stable (85, 92, 127, 129, 137, 198, 232, 252). It is of note that the two terminal force-bearing strands of TNfn3 are not directly bonded by backbone hydrogen bonds as other elastomeric proteins, such as I27 of the muscle protein titin (92).

Using protein engineering techniques, we engineered a polyprotein (TNfn3)_8, which consists of eight identical tandem repeats of TNfn3. We then used single molecule AFM to stretch polyprotein (TNfn3)_8 and characterize its mechanical unfolding behaviors. Stretching polyprotein (TNfn3)_8 resulted in force-extension relationships of the characteristic sawtooth pattern appearance, where the individual sawtooth peaks correspond to the mechanical unfolding of each individual TNfn3 domains in the polyprotein chain. The force-extension relationships of (TNfn3)_8 can be well described by the worm-like chain (WLC) model of polymer elasticity (236) and there is no apparent deviation from the WLC fits (Fig. 4.2). WLC fits to consecutive unfolding force peaks measure an average contour length increment (ΔLc) of 29.0±0.8 nm (average ± standard deviation). TNfn3 is 90 aa residues long and the distance of between its N-, and C- termini is 3.1 nm in the folded state. The contour length of the unfolded and fully stretched TNfn3
is 32.4 nm (90aa×0.36nm/aa). Hence, a complete unraveling of a TNfn3 domain should result in a contour length increment \( \Delta L_C \) of \( \sim 29.3 \) nm, which is in excellent agreement with the experimentally determined \( \Delta L_C \). This result suggests that the mechanical unfolding of TNfn3 corresponds to the complete unfolding of TNfn3 in an apparent two-state fashion and there is no visible intermediate state along its mechanical unfolding pathway.

The amplitude of the unfolding force peaks vary around \( \sim 120 \) pN. A histogram of the unfolding forces compiled from \( \sim 4000 \) unfolding events at a pulling speed of 400 nm/s measured an average unfolding force of \( 125 \pm 14 \) pN (\( n= 4198 \), Fig. 4.3A), which is in agreement with previous measurements on a similar TNfn3 polyprotein(250). The measured mechanical stability of TNfn3 is also similar to the average mechanical unfolding force of all the fifteen FnIII domains measured from a recombinant fragment of tenascin-C containing all the fifteen FnIII domains, consistent with the previous conclusion that all the FnIII domains of tenascin-C have similar mechanical stability (89, 148).

It is of note that it was shown that extending the C-terminus of TNfn3 by two residues can significantly increase the thermodynamic stability of TNfn3(253). However, we found that such an extension of TNfn3 does not affect the mechanical stability of TNfn3 in any way (data not shown), indicating that the appended two additional residues at the C-terminus of TNfn3 are already detached from the folded TNfn3 before TNfn3 reaches the mechanical unfolding transition state.
Figure 4.1 The three-dimensional structure of the third fibronectin type III domain of tenascin-C (TNfn3). A) TNfn3 has a typical $\beta$-sandwich structure. The two force-bearing $\beta$ strands are parallel to each other and are pointing in opposite directions. The backbone hydrogen bonds associated with the two force-bearing $\beta$ strands are indicated by black bars. B) The location of amino acids that are substituted by proline residues in this work.

Figure 4.2 Typical force-extension curves of polyprotein (TNfn3)$_8$. Stretching polyprotein (TNfn3)$_8$ results in force-extension curves of the characteristic saw-tooth pattern appearance. The equally spaced force peaks are resulted from the mechanical unfolding of the individual TNfn3 domains in the polyprotein chain. The last peak in force-extension curves corresponds to the detachment of the protein from either the AFM tip or substrate. WLC fits (thin lines) to the consecutive unfolding force peaks measure a contour length increment $\Delta L_C$ of $\sim 29.0 \pm 0.8$ nm.
4.3.2 The mechanical unfolding of TNfn3 is characterized by a long unfolding distance from the native state to the transition state.

To further characterize the mechanical unfolding of TNfn3 in detail, we measured the pulling speed dependence of the unfolding forces of TNfn3 by stretching (TNfn3)_8 at different pulling speeds. Similar to other elastomeric proteins, the mechanical unfolding of TNfn3 is a non-equilibrium process and its unfolding force depends on the pulling speed: the higher the pulling speed is, the bigger the unfolding force (Fig. 4.3B). It is of note that, as compared with Ig domains from the muscle protein titin(92, 102, 219), the pulling speed dependence of the unfolding force of TNfn3 is relatively weak: the unfolding force of TNfn3 increases from 109 pN at a pulling speed of 50 nm/s to 154 pN at a pulling speed of 2700 nm/s. To estimate the spontaneous unfolding rate constant at zero force ($\alpha_0$) and the unfolding distance ($\Delta x_u$) between the folded state and the transition state along the reaction coordinate, two important parameters characterizing the mechanical unfolding energy diagram, we carried out Monte Carlo simulations to reproduce the force-extension relationships of (TNfn3)_8. In the Monte Carlo simulation, we assumed that the unfolding of TNfn3 is a two-state process and the force-dependent unfolding rate constant follows the classical Bell model $\alpha(F) = \alpha_0 \exp\left(\frac{F\Delta x_u}{k_B T}\right)$, where $k_B$ is the Boltzmann constant and $T$ is temperature. We found that both the unfolding force histogram (Fig. 4.3A) and pulling speed dependency of unfolding forces (Fig. 4.3B) can be well described using a $\alpha_0$ of $1.5 \times 10^{-4}$ s$^{-1}$ and a $\Delta x_u$ of 0.42 nm. This result suggests that the mechanical resistance to unfolding is distributed over a distance of 0.42 nm. This unfolding distance is notably longer than that for other typical elastomeric proteins, such
as I27(92), ubiquitin(129) and GB1 domains(198), suggesting that the mechanical resistance of TNfn3 is distributed along a longer distance, which is in contrast with the highly localized mechanical resistance for other elastomeric proteins. The molecular basis for the observed long unfolding distance $\Delta x_u$ will be discussed in the Discussion section. The measured $\alpha_0$ and $\Delta x_u$ are comparable to those measured in previous studies on a tenascin-C fragment and the TNfn3 polyprotein(148, 250).

Figure 4.3 Unfolding force of TNfn3 and its dependence on the pulling speeds. A) Histogram of unfolding forces for TNfn3. The unfolding force histogram spans a range of ~100 pN (from ~60 pN to ~160 pN) with an average value of 125 ±14 pN (n= 4198). The red line corresponds to Monte Carlo simulation of the mechanical unfolding of TNfn3 using $\alpha_0$ of 1.5×10$^{-4}$ s$^{-1}$ and a $\Delta x_u$ of 0.42 nm. The pulling speed is 400 nm/s. B) The pulling speed dependence of the unfolding forces of TNfn3 (symbols). The pulling speed dependence of the unfolding forces of TNfn3 can be reproduced adequately by Monte Carlo simulations using $\alpha_0$ of 1.5×10$^{-4}$ s$^{-1}$ and $\Delta x_u$ of 0.42 nm (red line).

4.3.3 The SMD simulations of the mechanical unfolding of TNfn3

Single molecule AFM results suggest that the mechanical unfolding of TNfn3 is an apparent two-state process. In order to understand the molecular events leading to the mechanical unfolding of TNfn3, we carried out SMD simulations of the mechanical
unfolding of TNfn3. Different from previous molecular dynamics simulation work using an implicit water model (250), here we used the explicit water model TIP3P in our simulation to explicitly address the potential role of the solvent water molecules during the mechanical unfolding process of TNfn3. This strategy has been used extensively to simulate the mechanical unfolding of proteins, including FnIII domains from fibronectin and TNfn3 (150, 254).

In our SMD simulations, we simulated the mechanical unfolding of TNfn3 using both constant velocity and constant force protocols. Starting from 1 ns and 1.5 ns equilibrated conformations, TNfn3 was pulled at a constant force (500pN) or at a constant velocity (0.05 Å/ps). In total, we performed 16 constant force SMD and 17 constant velocity SMD simulations with a total simulation time of 57 ns. Both constant velocity and constant force trajectories revealed similar features of the mechanical unfolding of TNfn3.

In constant force SMD simulations, TNfn3 was stretched at a constant force of 500 pN from its N- and C-termini, and the distance between the two termini (RNC) was monitored as a function of time. Representative RNC-time curves from constant force SMD simulations were shown in Fig. 4.4A. The presence of multiple plateaus in RNC-time profiles is clearly visible, indicating the presence of stable intermediates populated along the unfolding trajectories. From the native state with RNC of ~34 Å, TNfn3 elongates by ~6 Å via straightening the disordered N-terminal end of the protein and enters into a stable intermediate state I1, in which the tertiary structure of TNfn3 remains largely intact. It is of note that the two backbone hydrogen bonds between residue Ser6 and Phe23 are relatively weak and break in state I1 during most of the trajectories. The rupture of the backbone
hydrogen bonds between Ser6 and Phe23 leads to the detachment of the first six residues of A strand from the folded structure of TNfn3. In the second stage, TNfn3 elongates further by ~10 Å to reach the intermediate state I2. During this process, the two β sheets, one containing A–B –E strands (colored in red) and the other one containing C’–C–F–G strands (colored in blue), rotate relative to each other and both align with the pulling force (Fig. 4.5), leading to the so-called aligned β sandwich intermediate I2. This alignment resulted in partial solvation of the periphery of the hydrophobic core. In intermediate I2, the remaining backbone hydrogen bonds in A-B strands between residues Glu9-Trp21, Lys11-Leu19, Asp12-Leu19, Thr14-Thr17 remained intact, so do the backbone hydrogen bonds in F-G hairpin.

Immediately following the transition from I1 to I2, TNfn3 elongated further following two distinct unfolding pathways (Fig. 4.5). The first pathway is characterized by the presence of the partially unfolded intermediate state I3 with $R_{N-C}$ of ~130 Å. Along this pathway (termed as “A-strand separates first” pathway), A-strand of TNfn3 separates first from the folded structure, followed by the subsequent unraveling of B and E strands. However, the overall structure of the β sheet containing C’-C-F-G strands remains largely intact. This unfolding pathway was observed in ~70% of the constant force SMD simulation trajectories (11 out of 16) and three representative trajectories of this type are shown in Fig. 4.4A (trajectories 1-3). The second pathway is characterized by simultaneous detachment of A and G strands from the folded structure and is termed as “A-G strands separate simultaneously”. After the simultaneous detachment of A and G strands, TNfn3 unfolds readily without any significant barrier and hence the intermediate state I3 is not detected in this pathway. This type of pathway was observed in ~30% of the
constant force SMD simulations (5 out of 16) and trajectory 4 in Fig. 4.4A is one of the examples.

Figure 4.4 Constant force and constant velocity SMD simulations of the mechanical unfolding of TNfn3. A) Representative $R_{\text{NC}}$ versus time profiles from constant force SMD simulations at a stretching force of 500 pN. The presence of three plateaus indicated that there exist kinetic intermediates, I1, I2 and I3. Trajectories 1-3 correspond to the “A-strand separates first” pathway, and trajectory 4 corresponds to the “A-G strands separate simultaneously” pathway. B) Representative Force- $R_{\text{NC}}$ curves from constant velocity SMD simulations. The pulling velocity is 0.05 Å/ps. The first force peak occurs at a $R_{\text{NC}}$ of ~40 Å, corresponding to the transition from I1 to I2; the second peak occurs at an $R_{\text{NC}}$ of ~50 Å and corresponds to the unraveling of intermediate I2.
Figure 4.5 Snapshots of TNfn3 during its simulated mechanical unfolding. For all the snapshots, the N terminus (red ball) was fixed and the C terminus (blue ball) was pulled during these simulations. From the native state, TNfn3 elongates by straightening the N-terminus and enters into the so-called twist intermediate state I1; then the two β sheets rotate relative to each other and align with the stretching force vector, leading to the so-called aligned intermediate state I2. After I2, TNfn3 unfolds via two distinct pathways: the first one is by separating the A-strand first from the folded structure, followed by the subsequent unfolding of the B and E strands, leading to a partially unfolded intermediate state I3; the second pathway is via simultaneous unfolding and detachment of A and G strands from the folded structure. After this event, TNfn3 unfolds readily and the existence of intermediate state I3 is not detected. For the snapshot of I3, the N terminus is artificially shortened to fit into the figure.

The dwell time of a given state along the unfolding trajectory is a measure of the stability of the given state. Constant force simulations revealed that the dwell time of intermediate I1 is on average longer than that of I2 and I3, as well as that of the native state, suggesting that intermediate state I1 seems to be the most stable one at the given force of 500 pN. This result suggests that under a stretching force of 500 pN, the native
state of TNfn3 is rapidly transformed into intermediate state I1 and the mechanical unfolding of TNfn3 is not directly from its native state. Thus, the intermediate I1 is the pseudo-native state for the mechanical unfolding of TNfn3. It is also important to point out that the dwell time of individual intermediate state does vary from trajectory to trajectory.

To reveal a more detailed picture of the transition from mechanical unfolding of TNfn3, we monitored the breakage of backbone hydrogen bonds in A-B, F-G hairpins by calculating hydrogen bond energies along the SMD trajectories. Five backbone hydrogen bonds between A-B strands, Glu\(^9\)(H)-Thr\(^{21}\)(O), Glu\(^9\)(O)-Thr\(^{21}\)(H), Lys\(^{11}\)(H) - Leu\(^{19}\)(O), Asp\(^{12}\)(O) - Leu\(^{19}\)(H), Thr\(^{14}\)(H) - Thr\(^{17}\)(O) and another five backbone hydrogen bonds between F-G strand, Phe\(^{88}\)(H) – Tyr\(^{68}\)(O), Glu\(^{86}\)(O) – Val\(^{70}\)(H), Glu\(^{86}\)(H) – Val\(^{70}\)(O), Ala\(^{84}\)(O) – Leu\(^{72}\)(H), Ala\(^{84}\)(H) – Thr\(^{72}\)(O) were selected for the hydrogen bond energy calculation. Fig. 4.6 shows the hydrogen bond energy as a function of time for these ten backbone hydrogen bonds for the trajectory 2 and 4 shown in Fig. 4.4A. At the beginning of the unfolding trajectories, these backbone hydrogen bonds are stable and the hydrogen bond energies fluctuate between ~ -5 kcal/mol and ~ -3 kcal/mol with an average value of ~ -4.3 kcal/mol. Upon further stretching, β strands will begin to separate and the strength of its hydrogen bonds become weaker, accordingly, the hydrogen bond energies gradually increase till reaching zero, at which point the hydrogen bonds are already broken. In “A-strand separates first” pathway (Fig. 4.6A and 4.6B), after ~ 1.8ns, the hydrogen bond energy of Glu\(^9\)(H)-Thr\(^{21}\)(O), Glu\(^9\)(O)-Thr\(^{21}\)(H), Lys\(^{11}\)(H) - Leu\(^{19}\)(O), Asp\(^{12}\)(O) - Leu\(^{19}\)(H), Thr\(^{14}\)(H) - Thr\(^{17}\)(O) increase to zero fairly rapidly accompanying the transition from intermediate state I2 to I3, indicating that these five backbone hydrogen bonds in the A-B strands break concurrently during this process. In contrast,
the hydrogen bonds in the F-G strands (Phe$^{88}$(H) – Tyr$^{68}$(O), Glu$^{86}$(O) – Val$^{70}$(H), Glu$^{86}$(H) – Val$^{70}$(O), Ala$^{84}$(O) – Leu$^{72}$(H), Ala$^{84}$(H) – Thr$^{72}$(O)) remain steady during this process. After 2.6 ns, three hydrogen bonds, Phe$^{88}$(H) – Tyr$^{68}$(O), Glu$^{86}$(O) – Val$^{70}$(H) and Glu$^{86}$(H) – Val$^{70}$(O), start to break.

In the “A-G strands separate simultaneously” pathway, the hydrogen bonds energy for Glu$^{9}$(H)-Thr$^{21}$(O), Glu$^{9}$(O)-Thr$^{21}$(H), Lys$^{11}$(H) - Leu$^{19}$(O), Asp$^{12}$(O) - Leu$^{19}$(H), Thr$^{14}$(H) - Thr$^{17}$(O), Phe$^{88}$(H) – Tyr$^{68}$(O), Glu$^{86}$(O) – Val$^{70}$(H), Glu$^{86}$(H) – Val$^{70}$(O), Ala$^{84}$(O) – Leu$^{72}$(H), Ala$^{84}$(H) – Thr$^{72}$(O) increases simultaneously after 2.7 ns, indicating that these ten pairs of backbone hydrogen bonds break concurrently during the “A-G strands separate simultaneously” pathway (Fig. 4.6C and 4.6D).
Figure 4.6 Profiles of hydrogen bond energy of inter-strand hydrogen bond in A-B and F-G strands versus time in two representative SMD unfolding trajectories of TNfn3. A-B) The energy change of the hydrogen bonds as a function of time in the trajectory following the “A-strand separates first” pathway. The energy was calculated from Trajectory 2 in Fig. 4.4A. C-D) The energy change of the hydrogen bonds as a function of time in the trajectory following the “A-G strands separates simultaneously” pathway. The energy was calculated from Trajectory 4 in Fig. 4.4A. The sudden increase in hydrogen bond energy indicates the breaking of hydrogen bonds. In both unfolding pathways, the hydrogen bonds in A-B and F-G break during the unfolding of the intermediate state I2.
4.3.4 Constant velocity SMD simulations

We also carried out constant velocity SMD simulations of the mechanical unfolding of TNfn3 (17 trajectories in total). Fig. 4.4B shows representative Force-R_{NC} curves. It is evident that there exist multiple force peaks along the unfolding pathway: the first force peak occurs at a R_{NC} of ~40 Å, which corresponds to the transition from I1 to I2; the second peak occurs at an R_{NC} of ~50 Å and corresponds to the unraveling of intermediate state I2. In all the trajectories, the amplitudes of the first and second peak show slight differences: in 13 trajectories, the first unfolding force peak is higher than the second one (~1300 pN versus ~1050 pN), while in 4 trajectories, the second unfolding force peak is higher than the first one (~1200 pN versus 1100 pN). This observation is consistent with the observed variation in the dwell time of the intermediates.

4.3.5 Using site-directed mutagenesis to probe the nature of the unfolding transition state observed in single molecule AFM.

SMD simulations on TNfn3 revealed that the mechanical unfolding of TNfn3 proceeds via several intermediate states. However, single molecule AFM experiments indicate that the mechanical unfolding of TNfn3 is an apparent two-state process. The discrepancy between SMD simulations and AFM experiments suggests that some of the intermediate states observed in SMD simulations are not populated on the time scale of single molecule AFM experiments. Considering the large R_{NC} of the intermediate I3, we can easily rule out the possibility that the intermediate I3 is a stable intermediate state. Previous SMD simulations and mechanical φ-value analysis suggested that the transition from the twist intermediate I1 to the aligned intermediate I2 is most likely to be the rate
limiting step for the mechanical unfolding observed in AFM and that the energy barrier for the transition from intermediate I2 to I3 is too low to be experimentally observed (150, 250). Our SMD simulation results reveal that in the majority of the unfolding trajectories, the plateau of intermediate state I1 is the longest in constant force SMD and the unfolding force peak for the transition from I1 to I2 is the highest in constant velocity SMD, supporting the view that the transition from I1 to I2 is the rate limiting step, which corresponds to the mechanical unfolding force peak observed in single molecule AFM experiments.

If this view is correct, it means that the rupturing events of the backbone hydrogen bonds (Fig. 4.5) occur after the rate limiting step. Thus it is tempting to conclude that the hydrogen bonds are not critical for the mechanical stability of TNfn3. For example, a previous single molecule AFM study on TNfn3 singled out the importance of hydrophobic interactions to the mechanical unfolding of TNfn3 (250, 255). However, the backbone hydrogen bonds are important in protecting the hydrophobic core from the attack by water molecules, especially the backbone hydrogen bonds in the A-B and F-G hairpins. It can be imagined that destabilization of the A-B or F-G strands by deleting backbone hydrogen bonds and introduction of bulge would facilitate the attack of the hydrophobic core by water molecules and lead to reduced mechanical stability. To further explore such scenarios, here we use proline mutagenesis to selectively disrupt A-B and F-G stands to directly probe the role of backbone hydrogen bonds and \( \beta \) sheet stability on the mechanical stability of TNfn3.
4.3.6 The design of proline mutants of TNfn3

It is well known that proline substitution in β sheet region can block the formation of backbone hydrogen bonds, cause a bulge in the β strand and disrupt hydrophobic packing(256). These combined disruptive effects by proline substitution result in the discontinuity of the β strand and lead to the selective disruption of local β sheet structure. Such a relatively large structural perturbation is ideal for probing the mechanical unfolding pathway of proteins, as such a large perturbation to the protein structure generally cannot be easily compensated by the structural rearrangement of the protein. Therefore, the structural perturbation caused by proline substitution can be easily located and its effect on the mechanical unfolding pathway can be readily identified(135, 257).

Since the breaking of the hydrogen bond between S6 and F23 occurs in intermediate state I1 and is also the first event during the mechanical unfolding process of TNfn3, we engineered a S6P mutant to specifically probe the mechanical unfolding intermediate state I1. To probe the effect of backbone hydrogen bonds in the AB hairpin during the mechanical unfolding, we engineered mutants E9P, K11P and T14P. Similarly, we introduced proline mutations in the G strand. To investigate the importance of the C-terminus on the mechanical stability of TNfn3, we engineered mutant T90P. We also engineered proline mutants A84P, E86P and F88P to further probe the importance of the G strand in the mechanical unfolding of TNfn3. The locations of these proline probes in TNfn3 are highlighted in yellow in Fig. 4.1B.
4.3.7 Phenotypic effects of proline mutations on A-strand of TNfn3

To investigate the phenotypic effects of proline mutations in the region of A strand, we constructed four proline mutants S6P, E9P, K11P and T14P. In order to characterize the mechanical unfolding of proline mutants using single molecule AFM in an unambiguous way, we constructed the heteropolyprotein (GB1-ProlineMutant)$_n$, in which TNfn3 mutants alternate with GB1 domains (Fig. 4.7A). In the polyprotein chimera, the well-characterized GB1 domains serve as fingerprints for identifying single molecule stretching events and discerning the signatures of the mechanical unfolding of the TNfn3 proline mutants(122, 199, 227). The mechanical unfolding of GB1 is characterized by contour length increment $\Delta L_C$ of $\sim$18 nm and unfolding force of $\sim$180 pN at a pulling speed of $\sim$400 nm/s(198, 249). Typical force-extension curves of the four heteropolyproteins involving the proline mutations in the A-strand are shown in Fig. 4.7B. Since TNfn3 alternates with GB1 domains in the heteropolyprotein, if we observed $N$ unfolding events of GB1 in a given force-extension curve, we are certain that the force-extension curve must contain the signature of the stretching and unfolding of at least $N-1$ TNfn3 mutant domains. Indeed, in the force-extension curves shown in Fig. 4.7B, the GB1 unfolding events with $\Delta L_C$ of $\sim$18 nm (in red) are preceded by the low force unfolding events at $\sim$100-120 pN (colored in green). It is evident that these low force unfolding events correspond to the mechanical unfolding of the proline mutants (S6P, E9P, K11P and T14P) in their respective heteropolyproteins. Indeed, WLC fits to these unfolding events measure $\Delta L_C$ of $\sim$28 nm, corroborating that these events indeed correspond to the complete mechanical unfolding of the TNfn3 proline mutants. The average $\Delta L_C$ is 28.9 nm, 28.8 nm, 28.6 nm,
28.8 nm for S6P, E9P, K11P and T14P, respectively, which is identical to that of wt TNfn3 within the resolution of our experiments.

Figure 4.7 Mechanical unfolding of proline mutants of TNfn3. A) Schematic illustration of the polyprotein chimera (GB1-TNfn3*)₄. TNfn3* denotes the proline mutant of TNfn3. B, C) Typical force-extension curves of the polyprotein chimera (GB1-TNfn3*)₄ for each proline mutant. B) shows the force-extension curves of mutants in which residues in A strand was substituted with prolines; and C) shows the force-extension curves of mutants involving proline mutations in the G-strand. The mechanical unfolding events of the well-characterized GB1 domains (colored in red) occurred at ~180 pN with $\Delta L_C$ of ~18 nm and serve as fingerprints to identify single molecule stretching events. Except for F88P, all the proline mutants show clear mechanical unfolding events with $\Delta L_C$ of ~29 nm (colored in green), and their unfolding forces range from ~90 to ~130 pN, which are lower than that of wt TNfn3. Dotted lines are WLC fits to the experimental data. In contrast to other proline mutants, the unfolding of F88P does not result in clear mechanical unfolding events. Instead, long featureless spacer is typically observed prior to the mechanical unfolding events of GB1 domains (curve a), suggesting that F88P domains unfold at forces that are lower than our detection limit (~20 pN). Only a small fraction of F88P domains show clear unfolding events, as the one shown in curve b.
Figure 4.8 Typical force-extension curve of polyprotein chimera (GB1-E9P)$_4$ that shows missing unfolding force peaks of E9P. Four GB1 unfolding events, which occurred at ~180 pN with $\Delta L_C$ of ~18 nm, are present in the force-extension curve. Since GB1 and E9P alternate in the polyprotein chimera, there should be at least three unfolding events of E9P. However, only one E9P unfolding event, which occurred at ~100 pN with $\Delta L_C$ of 30 nm, can be clearly identified. The unfolding events of at least two E9P domains are “missing”, indicating that these E9P domains unfold at forces that are below our detection limit (~20 pN). Thin lines correspond to WLC fits to the experimental data.

It is important to note that in some force-extension curves, it appears that the unfolding events of some TNfn3 mutant domains are “missing”. For example, in the curve shown in Fig. 4.8, there are four GB1 unfolding events but only one E9P unfolding event. The missing of at least two additional E9P domains in this particular curve suggests that at least two additional E9P domains unfold at forces that are lower than 20 pN. This is a general feature for all the proline mutants investigated here, suggesting that a minute population of TNfn3 mutants are already “unfolded” prior to the stretching and there might be conformational heterogeneity in the native conformation of the proline mutants. Nonetheless, the origin for this observation will be investigated in detail elsewhere.
Figure 4.9 The unfolding force histograms of TNfn3 proline mutants as well as F88A. All the proline mutants, except F88P, show well-defined mechanical unfolding forces ranging from ~90 pN to ~130 pN. Most of F88P unfold at forces below 20 pN, and a small population of F88P show clear unfolding events at significant forces.
The unfolding force histogram of S6P, E9P, K11P and T14P are shown in Fig. 4.9A. It is noticeable that the average unfolding force of S6P is ~127 pN, almost identical to that of wt TNfn3, despite the significant local structure disruption. This result is in good agreement with the SMD simulation result. SMD simulations showed that the first six residues are detached from the rest of TNfn3 in the twist intermediate state I1. Since the first six residues are already disordered in the pseudo native state of TNfn3, disrupting the interactions in this region should not have any effect on the mechanical unfolding kinetics as well as mechanical unfolding force. Indeed, our results on S6P are consistent with this picture.

In contrast, mutations E9P, K11P and T14P reduce the average unfolding force of TNfn3 by an average of ~30 pN (see Table 4.1), reflecting the effect of structural disruption on the mechanical stability of TNfn3. As a control, we also engineered the polyprotein chimera (GB1-wtTNfn3)_4. The average unfolding force of wt TNfn3 domain in the control construct was found identical with that of the wt TNfn3 domain in the homo-polyprotein (TNfn3)_8. Compared with the destabilization effect of more subtle alanine mutations in the same region, for example L8A(250), proline mutations in the A strand cause a slightly bigger destabilization effect. Considering that E9P, K11P and T14P are mutations from polar or charged residues to proline, these results suggest that in addition to the hydrophobic interaction, the overall stability of the β hairpin also plays important role in determining the mechanical stability of TNfn3.
Table 4.1. Unfolding force and kinetic parameters for the mechanical unfolding of TNfn3 and its mutants.

<table>
<thead>
<tr>
<th>Strand</th>
<th>Mutant</th>
<th>Unfolding force (pN)*</th>
<th>ΔLC (nm)</th>
<th>α0 (s⁻¹)</th>
<th>Δxu (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wt</td>
<td></td>
<td>125±14 (n = 4198)</td>
<td>29.0±0.8</td>
<td>1.5×10⁻⁴</td>
<td>0.42</td>
</tr>
<tr>
<td>A</td>
<td>S6P</td>
<td>127±22 (n = 1053)</td>
<td>28.9±1.2</td>
<td>1.0×10⁻⁴</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>E9P</td>
<td>96±21 (n = 1176)</td>
<td>28.8±1.4</td>
<td>1.0×10⁻²</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>K11P</td>
<td>103±16 (n = 1499)</td>
<td>28.6±1.2</td>
<td>5×10⁻²</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>T14P</td>
<td>94±18 (n = 879)</td>
<td>28.8±1.4</td>
<td>8×10⁻³</td>
<td>0.42</td>
</tr>
<tr>
<td>G</td>
<td>A84P</td>
<td>118±20 (n = 959)</td>
<td>28.7±1.0</td>
<td>7×10⁻³</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>E86P</td>
<td>98±20 (n = 678)</td>
<td>28.0±1.2</td>
<td>4×10⁻²</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>F88P</td>
<td>Below 20 pN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F88A</td>
<td>115±23 (n = 1121)</td>
<td>28.9±1.1</td>
<td>1.5×10⁻³</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>T90P</td>
<td>110±17 (n = 1283)</td>
<td>28.0±1.4</td>
<td>7×10⁻⁴</td>
<td>0.44</td>
</tr>
</tbody>
</table>

*The data is represented as average±standard deviation, n indicates the number of observations. All of these unfolding forces were measured at the pulling speed 400 nm/s.

4.3.8 Phenotypic effects of proline mutations on the G-strand of TNfn3: F88 is the Achilles heel of TNfn3

Using similar strategies, we constructed four proline mutants in the G-strand region to weaken the FG hairpin to investigate their effects on the mechanical unfolding of TNfn3. Fig. 4.7C shows the typical force-extension curves of the heteropolyproteins containing proline mutants. The force-extension curves of mutant A84P, E86P and T90P show clear mechanical unfolding events with ΔLC of ~28 nm (colored in green), corresponding to the complete mechanical unfolding of these three proline mutants. The unfolding force histograms for these three proline mutants are shown in Fig. 4.9B. It is evident that the average unfolding force for these three proline mutants is slightly lower than that for wt TNfn3 by 10 to 20 pN, indicating that such a large perturbation to the β sheet structure of TNfn3 in the G strand has a very mild effect on the mechanical stability of TNfn3.
Compared with A84P, E86P and T90P, mutant F88P, however, exhibits the strongest phenotypic effect in its mechanical unfolding behaviors. In contrast to other proline mutants, the force-extension curves of (GB1-F88P)₄ are characterized by a long featureless spacer followed by the GB1 unfolding events (colored in red). Clear unfolding events of $\Delta L_C$ of ~28 nm were absent from the vast majority of the force-extension curves of (GB1-F88P)₄ (Fig. 4.7C, force-extension curve a). Since GB1 alternates with F88P in the heteropolyprotein, the force-extension curves should contain roughly the same number of the stretching and unfolding events of GB1 and F88P domains. Therefore, the long featureless spacer must correspond to the stretching and subsequent unfolding of F88P domains, suggesting that F88P domains unfold at forces below 20 pN. This result indicates that the mutation F88P causes significant destabilization on TNfn3 so that the mechanical unfolding of TNfn3 occurs at very low forces. Occasionally, we also observed that a small number of F88P domains can unfold at significant forces. For example, one F88P domain in the force-extension curve b shown in Fig. 4.7C unfolds at ~68 pN with $\Delta L_C$ of 27 nm, while the other three F88P domains in the same chimera polyprotein did not generate any unfolding events. The unfolding force histogram of such rare unfolding events for F88P (Fig. 4.9b) shows that the average unfolding force is of 104±49 pN (n=320). Such high unfolding forces are unlikely due to the fluctuations in unfolding forces from those occurred below 20 pN. Instead, the difference in unfolding forces suggests that there are two distinct populations of F88P that have different mechanical stability: the majority of F88P is mechanically weak and unfolds below 20 pN, and a small percentage of F88P can unfold at forces of ~100 pN. This observation is suggestive of heterogeneity of the native
state of F88P, which deserves further experimental work in the future to fully characterize the origin of such a conformational heterogeneity of the native states of F88P.

These results indicate that the phenotypic effect of proline substitution in the G strand is context dependent: at the N-terminal end of the G strand, proline substitution had little effect on the mechanical stability. Proline substitutions at the C-terminal end of G strand, with the exception of F88P, have mild effect on the mechanical stability of TNfn3. F88P has the strongest destabilization effect on the mechanical stability of TNfn3 and seems to be the Achilles heel for the mechanical unfolding of TNfn3. These observations suggest that the C-terminal end of TNfn3 can play important roles in the mechanical unfolding of TNfn3. It is of note that a similar context dependent phenotypic effect has also been observed in the FNfn10 domain(135). Comparison between the two FnIII domains will be discussed in detail in the Discussion section.

It is of note that the F88P substitution blocks the formation of backbone hydrogen bond between F88 and Y68 and also significantly disrupts the hydrophobic packing interactions of TNfn3 mediated by the hydrophobic residue F88. To distinguish the contribution of hydrogen bond from that of hydrophobic interactions to mechanical stability, we also constructed an alanine mutant F88A, which only affects the hydrophobic interactions but not the backbone hydrogen bond. Single molecule AFM experiments on (GB1-F88A)$_4$ shows that the average unfolding force of F88A is ~115pN, indicating that mutant F88A is much more stable than F88P. This result suggests that the backbone hydrogen bond between residues 88 and 68 is critical for the mechanical and thermodynamic stability of TNfn3. Proline mutation F88P will block the original backbone hydrogen bond and introduce a bulge at the C-terminal end of G-strand, which
is likely to open the flood gate for water molecules to enter and solvate the hydrophobic core of TNfn3, leading to the significant destabilization of TNfn3.

4.3.9 Proline substitutions do not affect the mechanical unfolding distance of TNfn3.

To investigate how the proline substitutions affect the mechanical unfolding distance, we carried out pulling experiments on TNfn3 proline mutants at different pulling speeds. Similar to that of wt TNfn3, the average unfolding force for TNfn3 proline mutants also exhibit weak dependence on the pulling speeds at which the polyprotein is being stretched and unraveled (Fig. 4.10). The slope of the pulling speed dependence of the unfolding force for TNfn3 proline mutants is similar to that of the wt FNfn3, suggesting that proline mutations in A and G strands do not have significant effect on the mechanical unfolding distance $\Delta x_u$ of TNfn3. Therefore, the phenotypic effects of proline mutants are a result of the lowering of the mechanical unfolding energy barrier. Using Monte Carlo simulations, we found that using the parameters tabulated in Table 1 for $\alpha_0$ and $\Delta x_u$ can adequately describe the unfolding force histogram and the pulling speed dependence of the unfolding forces for TNfn3 proline mutants. These results corroborate that the phenotypic effect observed on TNfn3 proline mutants are due to a reduced mechanical unfolding energy barrier, rather than a change in mechanical unfolding distance.
Figure 4.10 Pulling speed dependence of the mechanical unfolding force of TNfn3 mutants. For comparison, the data for wt TNfn3 is also shown (black symbols). It is evident that the slope of the pulling speed dependence of the mechanical unfolding force of TNfn3 mutants is similar to that for wt TNfn3, indicating that the unfolding distance $\Delta x_u$ of TNfn3 mutants is similar to that of wt TNfn3. Solid lines are Monte Carlo fits using the parameters shown in Table 1.
4.4 Discussion

4.4.1 Mechanical unfolding of TNfn3: an FnIII domain of a robust mechanical design

Our single molecule AFM experiments revealed that TNfn3 is a mechanically stable protein, which unfolds at an average force of ~130 pN at a pulling speed of 400 nm/s. The mechanical unfolding of TNfn3 is an apparent two-state process without any observable intermediate state populating along its mechanical unfolding pathway. Using proline mutagenesis, we demonstrated that disrupting the A-B or F-G β hairpin by proline substitution can lead to a reduction of the mechanical stability of TNfn3, and the amplitude of reducing mechanical stability depends on the location of the proline substitution. Recent work downplayed the role of backbone hydrogen bonds in the mechanical unfolding of FnIII domains and suggested that the hydrophobic packing is critical for the mechanical stability of TNfn3(250), and this hypothesis was the basis to swap the hydrophobic cores between different FnIII domains in order to increase the mechanical stability of engineered FnIII domains(258). Our results on proline mutants clearly demonstrate that hydrophobic packing is not the only important factor in determining the mechanical stability of TNfn3. Backbone hydrogen bonds as well as the structural integrity of force-bearing β hairpins are also important structural parameters in defining the mechanical stability of TNfn3. Hence, the mechanical stability of TNfn3 carries both global and local structural attributes, necessitating the synergetic consideration of both hydrophobic core packing and backbone hydrogen bonds in the force-bearing β hairpins when designing novel FnIII domains with tailored mechanical properties.

Although proline mutagenesis helped to reveal the importance of both hydrophobic core packing and backbone hydrogen bonds in β hairpins in determining the mechanical
stability of TNfn3, it is not possible to estimate the relative contributions of the two types of interactions to the mechanical stability of TNfn3. This is because a proline mutation not only deletes the backbone hydrogen bond, but also affects hydrophobic interactions and introduces local structural disruption. Amide to ester mutation would be an ideal approach to quantitatively evaluate the contribution of backbone hydrogen bonds to mechanical stability(259-261). Despite the current challenge in applying amide-to-ester mutagenesis techniques in protein mechanics, further developments of amide-to-ester mutagenesis techniques will make it feasible to answer this question in the future.

Our proline mutagenesis results also reveal the robustness of the mechanical design of TNfn3, which is often underestimated by conservative side chain deletion approaches in some AFM studies. Except for residue F88, introducing disruptive proline mutations into the β sheets of TNfn3 does not result in catastrophic effects on its mechanical stability. Instead, a mild reduction in mechanical stability by ~10-30 pN was observed for these proline mutants, an effect that is comparable to that from conservative side chain deletion mutations in the same region of TNfn3(250). Such mild effects are in contrast with the significant destabilization effect of proline substitution on the native state observed for the tenth FnIII domains of fibronectin(135). These results highlight the global nature of the mechanical resistance of TNfn3, which is in sharp contrast to the rather local attributes of the mechanical resistance observed for other well-studied elastomeric proteins, such as I27(92). Since the energy barrier for the transition from twist state I1 to aligned state I2 is likely to be the rate limiting step for the mechanical unfolding of TNfn3, proline substitutions in AB and FG hairpins in TNfn3 not only disrupts β hairpins, but also facilitates the rotation and alignment of the two β sheets.
However, the exact molecular mechanism underlying such mechanical weakening effects on TNfn3 still needs to be elucidated. Moreover, the observation of a single weak point in TNfn3, the Achellis heel F88, is also quite surprising. It seems that F88 is the key to protecting the hydrophobic core via a combination of backbone hydrogen bonds and hydrophobic interactions. Since the mechano-phenotype is context dependent, the dramatic phenotype observed for F88P cannot be directly extended to its neighboring residues 87 and 89. Since residues 87 and 89 do not involve the formation of backbone hydrogen bonds, it will be of interest to examine whether proline mutation on such residues would lead to a similar dramatic destabilization effect. Furthermore, since all the FnIII domains of tenascin-C share similar mechanical stability(89, 148, 249), it will be interesting to check whether the features observed here for the mechanical design of TNfn3 will also apply to other FnIII domains of tenascin-C.

4.4.2 Comparison of the mechanical unfolding of TNfn3 versus FNfn10: similar structure but different unfolding behaviors

The mechanical unfolding of the tenth FnIII domain of fibronectin (FNfn10) has been characterized in detail using single molecule AFM and SMD simulations(135, 150, 156, 254, 262). Despite the highly homologous structures between FNfn10 and TNfn3, their mechanical unfolding behaviors show interesting commonalities as well as differences.

SMD simulations show strikingly similar signatures in the mechanical unfolding of these two domains(150, 156, 254, 262). Both proteins follow very similar sequences of molecular events: first, the pre-detachment of the first a few residues of the N-terminal end of A strand leads to a slightly twist state I1; then the two β-sheets rotate and align with each
other to enter the aligned state I2; after that, the protein unfolds in two distinct pathways, e.g. A-strand separates first and A-G strands separate simultaneously, and reaches a partially unfolded intermediate state I3 in the “A-strand separates first” pathway; finally the domain completely unravels. The transition from the twist state I1 to the aligned state I2 has been indicated as the main unfolding barrier.

Despite the similarity in the unfolding sequence, a significant difference sets the two FnIII domains apart: the presence of stable intermediate state I3 predicted in SMD simulations for FNfn10 was observed experimentally in single molecule AFM experiments(135), while the similar intermediate state I3 was not observed for TNfn3 in the AFM experiments.

The similarity and differences between the two FnIII domains also resides in their response to proline mutation. The phenotypic effect of proline mutations in the A strand is catastrophic for FNfn10, as proline mutations in the A strand destabilize the protein so dramatically that the transition from the aligned state I2 to I3 is abolished(135). And the FNfn10 proline mutant unfolds directly from the intermediate state I3. In contrast, the phenotypic effect of proline mutations in the A strand of TNfn3 is much milder and the destabilization facilitates the transition from I1 to I2.

Compared with the difference in phenotypic effect in the A strand, proline mutations in the G strand have strikingly similar effects for both FNfn10 and TNfn3. Residue 88 seems to be the weak point for both proteins, as proline substitution at this residue leads to significant destabilization of both proteins: for FNfn10, destabilization leads to the elimination of the transition from I2 to I3 and FNfn10 unfolds directly from the intermediate state I3(135); for TNfn3, since the intermediate state I3 is much less stable,
destabilization caused by substitution of residue 88 with proline leads to the complete unraveling of TNfn3 at low forces. Proline substitution at the N-terminal end of G strand shows very similar minute effects on the mechanical unfolding of both proteins.

Understanding the relationship between sequence variation and mechanical unfolding features of the two homogenous proteins will be critical for future efforts of engineering FnIII domains with tailored mechanical properties. A recent study has made an encouraging first step towards this goal (258).

### 4.4.3 Single-molecule AFM versus SMD: similarities and discrepancies

Combining protein engineering, single molecule AFM and SMD simulation techniques, we have characterized the mechanical unfolding pathways of TNfn3 at two vastly different time scales. Comparing the single molecule AFM results versus SMD predictions, some of the SMD predictions are verified by the single molecule AFM results, but there are also some important discrepancies between the two and SMD simulation results cannot fully explain the experimental findings in single molecule AFM experiments.

Similar to previous SMD simulations (150, 250), our SMD simulations of the mechanical unfolding of TNfn3 using an explicit water model predicted that the first step of the mechanical unfolding process of TNfn3 is the disruption and detachment of the N-terminus (residues 1-6) followed by the alignment of the two β-sheets. Afterwards, the unfolding of the aligned β-sheets is initiated by the rupture of the AB β hairpin followed by a mechanical unfolding intermediate state I3 or by the simultaneous separation of A-G strands from TNfn3. The unraveling of the I3 state will lead to a fully unfolded TNfn3. These three on pathway unfolding intermediate states observed in the SMD simulations
have different stabilities. Analysis of both the unfolding force (from the constant velocity SMD) and dwell time (from the constant force SMD) of the three intermediate states showed that I1 is the most stable one and I2 is the least stable one.

Compared with the SMD simulation results, single molecule AFM results confirm that the disruption and detachment of the N-terminus of TNfn3 is likely to be the very first step in the mechanical unfolding of TNfn3, as S6P mutation does not have any effect on the mechanical unfolding of TNfn3. This agreement confirms that the mechanical unfolding of TNfn3 observed in single molecule AFM is not the unraveling of TNfn3 from its native state, but from a force-induced pseudo ground state, in which the N-terminal end of the A strand is already detached. In addition, SMD simulations predict that the transition from twist intermediate state I1 to the aligned state I2 is the rate-limiting step for the unfolding of TNfn3. This transition involves the rotation and alignment of the two β-sheets, and is accompanied by the extension of the distance between the N- and C-termini by ~10 Å. Therefore, TNfn3 can be deformed over a longer distance before it unfolds. This description provides a plausible molecule-level explanation for the observed longer unfolding distance $\Delta x_u$ for TNfn3. In contrast, unraveling I27-like elastomeric proteins requires simultaneous rupture of multiple hydrogen bonds holding the two force-bearing β strands together. Therefore, I27-like proteins can only be deformed over a shorter distance before they unfold, giving rise to shorter unfolding distances. As to the average mechanical unfolding force of I27, it is 200 pN at 400 nms$^{-1}$ which significant higher than that of TNfn3. This high mechanical unfolding force of I27 has been explained in terms of shearing 6 hydrogen bonds between strands A’ and G. While by examining the SMD simulation for the mechanical unfolding of TNfn3, the “A-G strands separate
simultaneously” pathway is found involving simultaneous rupture of 10 hydrogen bonds. The relatively lower unfolding force of TNfn3 is due to the fact that the unraveling of the protein domain is not simply determined by the number of involved hydrogen bonds. Many other factors are important to define the mechanical stability of a protein, such as the hydrophobic interaction and the β sheet stability. Therefore, the observation that I27 is mechanically more stable than TNfn3 is consistent with the SDM simulation results.

Despite the agreements between SMD simulation predictions with single molecule AFM experiments, some discrepancies between the two do exist. In contrast to the SMD prediction of the existence of intermediate I3, single molecule AFM results indicated that the mechanical unfolding of TNfn3 is an apparent two-state process, and no intermediate state is observed in single molecule AFM experiments. Although intermediate I3 was predicted to be mechanically stable in both constant force and constant velocity SMD simulations, intermediate state I3 does not populate on the time scale of single molecule AFM experiments. The origin of such a discrepancy remains to be explored. This situation is in sharp contrast with that for FNfn10, for which a very good agreement was reached for the similar intermediate state I3 between single molecule AFM experiments and SMD simulations(135, 150, 254, 262). Despite such limitations, it is clear that SMD has offered valuable insights into the molecular mechanism of mechanical unfolding of proteins. Improvements in SMD methodologies will continue to further our understanding of the mechanical unfolding and mechanical design of elastomeric proteins with unprecedented detail.

In summary, we have characterized the mechanical unfolding of TNfn3 using a combination of single molecule AFM, SMD simulations and proline mutagenesis
techniques. Our results revealed the robust mechanical design of TNfn3 that enable the protein to resist the disruptive mutations such as proline substitution. Moreover, both local and global structural features are important for determining the mechanical resistance of TNfn3. It has become evident that the mechanical unfolding of TNfn3 is much more complex than other typical elastomeric proteins, such as I27 (92, 263), due to the structural flexibility and deformability of TNfn3. Therefore, TNfn3 and FnIII domains in general present further challenges for protein engineers to enhance the mechanical stability of FnIII domain in a systematic and rational fashion.

4.5 Experimental section

4.5.1 Protein engineering

The DNA sequence coding TNfn3, flanked with a 5’ BamHI restriction site and 3’ BglII, HindIII restriction sites, was amplified using Polymerase Chain Reaction (PCR) from the plasmid TNfnALL encoding all the fifteen FnIII domains. The plasmid TNfnALL was a generous gift from Professor Harold Erickson (Duke University). Polyprotein (TNfn3)_8 was constructed using a consecutive DNA concatamerization method based on the identity of the sticky ends generated by the BamHI and BglII restriction enzymes.

To facilitate the identification of the mechanical unfolding signatures of the proline mutants of TNfn3 using AFM, we constructed heteropolyproteins consisting of alternating GB1 domains and proline mutant TNfn3, where the well-characterized GB1 domains serve as internal fingerprints for identifying single molecule stretching events. Since GB1 gene carries a 5’ BamHI restriction site and 3’ BglII, KpnI restriction sites, we
constructed a new version of TNfn3, which carries a 5’ BamHI restriction site and 3’ BglII, KpnI restriction sites, to facilitate the construction of the heteropolyprotein(92). Proline mutants were constructed using standard site directed mutagenesis method. Genes of heteropolyproteins were constructed using similar protocols as those used for constructing polyprotein (TNfn3)₈ based on the identity of the sticky ends generated by the BamHI and BglII restriction enzymes.

Polyproteins were overexpressed in DH5α strain and purified from supernatant using Ni²⁺-affinity chromatography. The polyproteins were kept at 4°C in PBS buffer at a concentration of ~200 μg/mL.

4.5.2 Single molecule AFM

Single-molecule AFM experiments were carried out on a custom built atomic force microscope, which was constructed as described(88). All the force-extension measurements were carried out in PBS buffer. In a typical experiment, the polyprotein sample (1 μL) was deposited onto a clean glass cover slip covered by PBS buffer (50 μL), resulting in a thin protein layer of the thickness of 10-20 nm. The thickness of this protein layer depends upon the amount of proteins being deposited onto the glass cover slip. The more protein is deposited, the thicker the layer. The thickness of the protein layer may contribute to the apparent contour length of the polyprotein, sometimes resulting in apparent contour length that is longer than the theoretical contour length of the polyprotein(122). The spring constant of each individual cantilever (Si₃N₄ cantilevers from Vecco, with a typical spring constant of 40 pN/nm) was calibrated in solution using the equipartition theorem before and after each experiment.
4.5.3 SMD simulation

Crystal structure of TNfn3 (PDB accession code 1TEN) was used as the starting conformation of TNfn3 for simulated equilibration. The protein was solvated in a water box (length 107 Å, width 62 Å, height 53 Å) with TIP3P water model(264). The whole protein–water system contains 33,194 atoms. SMD simulations of the mechanical unfolding of TNfn3 was carried out with the program of NAMD 2.6(265) and with CHARMM22 force field (266) following the protocol described previously(199, 267). The initial structure of TNfn3 was equilibrated for 1.5 ns at 300 K. Compared to the crystal structure, the equilibrated structures at 1ns and 1.5ns have a backbone RMSD of 0.93Å and 0.98 Å, respectively, and were used as the starting conformation for the constant force and constant velocity SMD simulations. During the SMD simulations, the N-terminal Cα atom was fixed and C-terminal Cα atom was pulled. A pulling speed of 0.05 Å/ps was used in constant velocity SMD simulations and a pulling force of 500 pN was used in constant force SMD simulations. The simulation time was 57 ns in total. System setup, structural analysis and calculation of hydrogen bond energies were performed using VMD1.86 (268).

4.5.4 Monte Carlo simulation

Monte Carlo simulations of the stretching and unfolding of the polyproteins were carried out according to published procedures(85). The Monte Carlo simulations were done as follows: The force applied on the polypeptide chain is calculated using the WLC model shown in equation 1.1, using the parameters from the Tnfn3 domain (a persistence length of 0.4 nm; unfolding of a TNfn3 domain always lead to a contour length increment
of 30 nm). The probability of unfolding can be calculated as $P = k_u(F) \times \Delta t$, where $\Delta t$ is the time interval and $k_u(F)$ is the rate constant for unfolding at given force $F$ (equation 1.2). The simulation was done by stretching the polypeptide chain for a small time interval $\Delta t$, computing the resulting force with the WLC model, calculating the unfolding probability for the domain under that force, and then comparing the unfolding probability with a random generated number in order to define its status. If the unfolding probability is larger than the random number, the domain is counted as unfolded; if the unfolding probability is smaller than the random number, the domain is counted as remaining folded. Then, the molecule will be stretched for one more time interval $\Delta t$. The new unfolding probability will be calculated and be compared with a newly generated random number. This process will be repeated until the TNfn3 domain is unfolded. The simulations gave force extension curves similar to the experimental data. By repeating the simulation trials, we can regenerate the unfolding force histogram of the TNfn3 domain at different pulling speed. By varying the value of $\alpha_0$ and $\Delta x_u$, we can find out the one set of $\alpha_0$ and $\Delta x_u$ which can best reproduce the experimental unfolding force histogram. Therefore, the unfolding rate constant at zero force $\alpha_0$ and the distance of the native state to the transition state $\Delta x_u$ along the reaction coordinate of the mechanical unfolding reaction were estimated using Monte Carlo simulations procedures in a trial-and-error fashion. Such a procedure is necessary due to the lack of analytical solutions to the unfolding force distribution obtained from force-extension measurement for polyproteins. The accuracy of the fitting parameters is exemplified in Fig. 4.11, which plots the simulated unfolding force histogram and the pulling speed dependence of the unfolding forces using different sets of unfolding rate constant $\alpha_0$ and unfolding distance.
Δx_u. Typically, α_0 is accurate within a factor of 3 and the unfolding distance Δx_u is accurate within 0.05 nm.

Figure 4.11 Using Monte Carlo simulation to estimate the unfolding rate constant α_0 and unfolding distance Δx_u. A) Monte Carlo simulation fits to the unfolding force histogram (grey) using the same Δx_u (0.42 nm) but different unfolding rate constant α_0 of 0.5×10^{-5} s^{-1}, 1.5×10^{-4} s^{-1} and 4.5×10^{-4} s^{-1}. B) Monte Carlo simulation fits to the pulling speed dependence of the unfolding forces of TNfn3 (black symbols) using the same α_0 (1.5×10^{-4} s^{-1}) and different Δx_u of 0.37 nm, 0.42 nm and 0.47 nm. The pulling speed dependence of the unfolding forces of TNfn3 is demonstrated as black symbols. In practice, α_0 is accurate within a factor of 3 and the unfolding distance Δx_u is accurate within 0.05 nm.
Chapter 5: Kinetic partitioning mechanism governs the folding of the third FnIII domain of tenascin-C: evidence at the single molecule level

5.1 Synopsis

Statistical mechanics and molecular dynamics simulations proposed that the folding of proteins can follow multiple parallel pathways on a rugged energy landscape from an unfolded state en route to their folded three-dimensional structures. The kinetic partitioning mechanism is one of the possible mechanisms underlying such complex folding dynamics. Here we use single molecule atomic force microscopy techniques to directly probe the multiplicity of the folding pathways of the third FnIII domain from an extracellular matrix protein tenascin-C (TNfn3). We mechanically manipulated single molecules of (TNfn3)₈ and forced TNfn3 domains to undergo mechanical unfolding and refolding cycles, allowing us to directly observe the folding pathways of TNfn3. We found that, after being mechanically unraveled and then relaxed to zero force, TNfn3 follows multiple parallel pathways to fold into its native state. The majority of TNfn3 fold into the native state in a simple two-state fashion, while a small percentage of TNfn3 were found to be trapped into stable folding intermediate states with well-defined three-dimensional structures. Furthermore, the folding of TNfn3 was also influenced by its neighboring TNfn3 domains. Complex misfolded states of TNfn3 were observed, possibly due to the formation of domain-swapped dimeric structures. Our studies

*A version of this chapter has been submitted for publication as “[Peng Q.], Fang J., Wang M., Li H. Kinetic partitioning mechanism governs the folding of the third FnIII domain of Tenascin-C: Evidence at the Single Molecule Level”.

114
revealed the ruggedness of the folding energy landscape of TNfn3, and provided direct experimental evidence that the folding dynamics of TNfn3 is governed by the kinetic partitioning mechanism. Our results demonstrated the unique capability of using single molecule AFM to probe the folding dynamics of proteins at the single molecule level.

5.2 Introduction

In Chapter 4, we have investigated the mechanical unfolding and the mechanical design of TNfn3 in detail using single molecule AFM. From the single-pulling experiments, the mechanical unfolding of TNfn3 exhibits an apparently simple two-state mechanism. Our results are consistent with previous studies on TNfn3 from other groups(250). However, the mechanical folding of TNfn3 has been left unexplored. As we all know, the robust and efficient folding into well-defined three-dimensional structures is an essential step for most proteins to acquire their biological functionality and thus plays an indispensable role in almost every aspect of life. Describing the folding mechanisms of globular proteins has been one of the challenges in the life Sciences. Significant progress in theory and experiments has revealed tremendous insights into the protein-folding mechanism and led to a conceptual framework of understanding(18, 21, 22, 59). It is generally accepted that, from an unfolded and random-coil-like state, proteins may follow multiple parallel folding pathways on a rugged high-dimensional energy landscape en route to their native states(18, 176), which is governed by the kinetic partitioning mechanism. In Chapter 2, we have characterized the multiple mechanical unfolding pathways of T4-lysozyme (T4L) which is direct experimental proof for the kinetic partitioning theory. However, T4L is a natural domain-insertion protein with
multiple subdomains. It remains challenging to experimentally test the theoretical predictions for kinetic partitioning on the simple small protein with only one domain and the related experimental evidence is limited (66, 139, 178, 227).

As demonstrated in Chapter 2, the development of single molecule techniques has provided tools that are uniquely suited to address such challenges (23, 24, 57), especially the single molecule AFM technique (23, 86, 88, 92, 269, 270). Single molecule AFM allows the mechanical manipulation of the same protein over an extended period of time, making it possible to acquire a large number of folding/unfolding trajectories on the same protein molecule and observe rare folding/unfolding events that are otherwise difficult to observe. Such a unique capability provides the possibilities to experimentally probe the multiplicity of folding/unfolding pathways of proteins. Here we used single molecule AFM to further probe the folding/unfolding pathway of TNfn3 in the refolding operation mode (unfolding & refolding the same polyprotein molecule repetitively for many times).

As mentioned in Chapter 4, the force-induced unfolding/refolding reactions of FnIII domains are likely an important part of tenascins dynamics in vivo (148). Here, by using single molecule AFM techniques to investigate the folding dynamics of TNfn3 in detail, we distinctly demonstrated that the folding of TNfn3 follows multiple parallel folding pathways. The majority of TNfn3 domains fold into their native states in a simple two-state fashion, while a small percentage of TNfn3 were observed to be trapped into multiple well-defined folding intermediate states. Our studies revealed the ruggedness of the folding energy landscape of TNfn3, and provide experimental evidence supporting the kinetic partitioning mechanism for protein folding.
5.3 Results

5.3.1 Mechanical folding dynamics of TNfn3

As we demonstrated in Chapter 4 (250, 271), stretching polyprotein (TNfn3)$_8$ results in force-extension curves of characteristic sawtooth-like appearance, where individual sawtooth peak corresponds to the mechanical unfolding of each individual TNfn3 domain in the polyprotein chain and the last peak generally corresponds the stretching and subsequent detachment of the unfolded polypeptide chain. Fitting the consecutive unfolding force peaks to the worm-like chain (WLC) model of polymer elasticity(236) allows for accurate determination of the contour length increment ($\Delta L_c$) upon domain unfolding. $\Delta L_c$ is an intrinsic parameter associated with the structure of the protein and can provide useful information about the folded/misfolded structure of proteins. These prior studies provided the foundation for single molecule AFM studies on the folding dynamics of TNfn3.

It was demonstrated that, through trial-and-error, it is possible to pick up a polyprotein and hold onto it for multiple stretch-relax cycles before it detaches from either the AFM tip or substrate. Using a well-established double-pulse protocol (Fig. 5.1B) (122, 148), we measured the folding kinetics of TNfn3 at zero force. During the first pulse, a fragment of polyprotein (TNfn3)$_8$ was stretched to unfold all the available TNfn3 domains (trace i). The number of unfolding force peaks in the force-extension curve measures the number of TNfn3 domains $N_{\text{total}}$ in the polyprotein fragment being stretched. Then the unfolded polyprotein chain was relaxed rapidly to zero extension. After being relaxed at zero extension for a time period $\Delta t$, the polyprotein was stretched again during the second pulse. The number of unfolding force peaks, $N_{\text{refolded}}$, observed in
the force-extension curve indicated the number of TNfn3 domains that managed to refold during the waiting time $\Delta t$. It is evident that the folding probability $\frac{N_{\text{refolded}}}{N_{\text{total}}}$ of TNfn3 depends exponentially on time $\Delta t$ (Fig. 5.1C). The folding kinetics can be well described by a first-order kinetics equation $\frac{N_{\text{refolded}}}{N_{\text{total}}}=1-\exp(-\beta_0 t)$ using a rate constant $\beta_0$ of 1.1 s$^{-1}$. This result indicates that the folding reaction of TNfn3 largely follows simple two-state kinetics.
Figure 5.1 The majority of TNfn3 folds in a two-state fashion. A) Three dimensional structure of TNfn3. TNfn3 is composed of seven β strands (labeled A to G) folded into a typical β-sandwich structure. B) The double-pulse protocol used to probe the folding kinetics of TNfn3. The polyprotein was first stretched to unfold all the TNfn3 domains in the chain and count the total number of domains in the polyprotein chain, $N_{\text{total}}$ (upper traces), and then the unfolded polyprotein was quickly relaxed back to its original length. After a relaxation time $\Delta t$, the protein was stretched again in the second pulse to count the number of domains that refolded within $\Delta t$, $N_{\text{refolded}}$ (lower traces). C) Plot of the refolding probability, $N_{\text{refolded}}/N_{\text{total}}$, versus $\Delta t$. Error bars represent standard deviation. Solid line corresponds to the fit of the first-order kinetic equation, $N_{\text{refolded}}/N_{\text{total}} (t) = 1 - \exp(-\beta_0 t)$, to the experimental data using $\beta_0 = 1.1 \text{ s}^{-1}$. 
5.3.2 Folding intermediate states are detected in the mechanical unfolding/refolding cycles of TNfn3

During repeated stretching-relaxation cycles, most of unfolding events of TNfn3 displayed contour length increment $\Delta L_C$ of 29 nm, indicating that these TNfn3 domains folded into their native state. However, a small percentage (~5%, 342 out of 6558 events) of unfolding events exhibit contour length increment that is significantly shorter than 29 nm. During repeated stretching-relaxation cycles, the same protein molecule was unfolded and refolded for an extended period of time without the interference of other molecules and non-specific adhesion. Therefore, we can ensure that the folding/unfolding trajectories are from the same protein molecule and the unfolding events with $\Delta L_C$ smaller than 29 nm genuinely resulted from the partial folding of TNfn3 domains in the polyprotein chain.

Five representative force-extension curves from the same Tnfn3 polyprotein molecule are shown in Fig. 5.2A. There are six TNfn3 domains in the polyprotein fragment being stretched, as evidenced by the six unfolding events with $\Delta L_C$ of ~29 nm (curve a). After the unfolded polyprotein was relaxed to zero extension and allowed to refold, we observed six unfolding events (curve c) in the subsequent force-extension curve. Out of the six unfolding events, only five of them showed $\Delta L_C$ of ~29 nm, and one unfolding event displayed $\Delta L_C$ of 7 nm (colored in blue), suggesting that five unfolded TNfn3 domains folded back to their native state, while the sixth domain only refolded partially. The other three force-extension curves showed that the unfolded TNfn3 domain can also partially fold into different structures, showing $\Delta L_C$ of 24 nm, 12 nm and 18 nm, respectively (colored in blue). It is of note that after unfolding, the partially folded TNfn3
could refold into its native state showing $\Delta L_C$ of 29 nm upon relaxation. In the curve b and e, only four normal unfolding force peaks for TNfn3 were found. It is possible that in those cases the measured abnormal unfolding events were actually the misfolding involving parts of two domains rather than folding intermediates of a single domain. However, the likelihood of this possibility is relatively low considering that in most cases there is no missing peak and only one normal peak replaced by a “strange” force peak with a shorter contour length increment.

Since the contour length increment $\Delta L_C$ is an intrinsic structural parameter determined by the number of amino acid residues involved in the folded structure of the protein, a shorter $\Delta L_C$ suggested that these unfolding events originated from the unfolding of a partially folded TNfn3, instead of a fully folded TNfn3. Such partially folded structures formed during the folding process and likely correspond to folding intermediate states of TNfn3. This is the first time that such partially folded intermediate states for TNfn3 domain were observed. This observation suggested that there are kinetic traps along the mechanical folding pathway of TNfn3 that may confine the TNfn3 domain into partially folded structures. The histogram of $\Delta L_C$ of the folding intermediate states showed four preferred peaks centered at $\sim 8 \pm 2.3$ nm, $\sim 12 \pm 1.7$ nm, $\sim 18 \pm 1.8$ nm and $\sim 24 \pm 1.6$ nm (Fig. 5.2B), suggesting that there are four well-defined folding intermediate states along the folding pathway. These results indicated that TNfn3 can sample multiple parallel folding pathways during its folding process: the statistically preferred folding pathway of TNfn3 is the simple two-state folding pathway, and a small percentage of TNfn3 can follow pathways involving partially folded intermediate states.
In addition, these folding intermediate states were found to be mechanically stable. The unfolding forces of the folding intermediate states showed a broad distribution with an average unfolding force of 94 ± 33 pN. However, these folding intermediate states are mechanically weaker than the correctly folded TNfn3 (average unfolding force of 133± 16 pN).
Figure 5.2 Mechanical folding experiments revealed that TNfn3 can fold into multiple distinct folding intermediate states. A) Multiple distinct folding intermediates were observed in the same (TNfn3)_k molecule. In trace a), all the unfolding events of TNfn3 showed normal a $\Delta L_C$ of $\sim 29$ nm, indicating that all the TNfn3 domains folded into their native state. In contrast, in traces b-e), there was one unfolding event showing significantly shorter $\Delta L_C$ in each curve. Such shorter $\Delta L_C$ in each curve is different from one another, suggesting that TNfn3 was trapped into distinct folding intermediate state. B) The histogram of $\Delta L_C$ of the folding intermediate states shows four preferred peaks centered around $\sim 8 \pm 2.3$ nm, $\sim 12 \pm 1.7$ nm, $\sim 18 \pm 1.8$ nm and $\sim 24 \pm 1.6$ nm, suggesting that there are four well-defined folding intermediate states along the folding pathway. Red lines correspond to the Gaussian fits to the experimental data. Inset: the histogram of $\Delta L_C$ of normal TNfn3 unfolding events show a narrow distribution with an average $\Delta L_C$ of $29.3 \pm 0.7$ nm. C) The histogram of unfolding force of folded folding intermediate states of TNfn3 shows a broad distribution with an average force of $94 \pm 33$ pN. The unfolding force of the folding intermediate states is significantly smaller than that of native state of TNfn3 ($133 \pm 16$ pN) as shown in the Inset.
5.3.3 Folding of TNfn3 is influenced by its neighboring TNfn3 domains: misfolded superfolds

In addition to the observation of partially folded intermediate states, we also observed that a small percentage (~3%, 200 out of 6558 events) of force-extension curves of (TNfn3)$_8$ showed missing force peaks, leading to the so called “skips”. For example, the first force-extension curve shown in Fig. 5.3A showed five equally spaced unfolding force peaks. However, after the unfolded polyprotein was relaxed at zero extension for 5s to allow for refolding, the subsequent stretching of the polyprotein resulted in force-extension curve with only four unfolding force peaks. The first peak showed $\Delta L_C$ that is significantly longer than the normal $\Delta L_C$, giving the appearance of a “skip”. After unfolding, the following force-extension curve showed five unfolding force peaks again, recovering its original number of folded TNfn3 domains, suggesting that the formation of the skip is reversible.

Fitting the WLC model to the data revealed that the skips displayed different $\Delta L_{C\text{-skip}}$ in different force-extension curves (Fig. 5.3B-D). The histogram for $\Delta L_{C\text{-skip}}$ shows a broad distribution ranging from 33 nm to 100 nm, with multiple preferred peaks centered around 36 nm, 48 nm, 54 nm and 66 nm, respectively (Fig. 5.4A).

It was observed that the number of unfolding events in force-extension curves containing a skip is always one fewer than the number of TNfn3 contained in the polyprotein fragment. Moreover, $\Delta L_{C\text{-skip}}$ is significantly larger than $\Delta L_C$ of a correctly folded TNfn3. These results suggested that the formation of a skip is due to the coalescence of two neighboring TNfn3 domains. Hence, these skips likely correspond to misfolded superfolds involving two or more TNfn3 domains.
Figure 5.3 The misfolding behaviors of TNfn3 involving neighboring TNfn3 domains. During repetitive stretching and relaxation experiments, most TNfn3 domain can fold into their native states. A small percentage of TNfn3 domains misfold into structures showing $\Delta L_{C\text{-}skip}$ that are significantly larger than the normal $\Delta L_C$ of TNfn3 (shown in A-D), leading to the observation of skips. The significantly larger $\Delta L_C$ suggested that two neighboring domains likely coalesced into a superfold. After unfolding of the misfolded superfold, TNfn3 domains can fold back again into the native state.

“Skip” events were first reported on polyprotein (I27)$_8$ and native tenascin-C and showed $\Delta L_C$ slightly bigger than $2 \times \Delta L_C$. They were attributed to the formation of a misfolded superfold (272) possibly via a domain-swapping mechanism. It is possible that a similar domain-swapping mechanism is also responsible for the observed misfolded superfold of TNfn3. However, the misfolding pattern for TNfn3 seems much more complicated than that of I27 and tenascin-C, as the $\Delta L_{C\text{-}skip}$ histogram showed much richer information than I27 and tenascin-C. The broad distribution of $\Delta L_{C\text{-}skip}$ (from 33 nm to
~100 nm) indicated that the misfolded superfolds for TNfn3 could involve up to four TNfn3 domains in the polyprotein (TNfn3)$_8$.

Figure 5.4 Mechanical properties of the misfolded superfold of TNfn3 domains. A) The histogram of $\Delta L_{C\text{-skip}}$ of misfolded superfolds. The histogram shows four preferred peaks centered around ~36 nm, ~48 nm, ~54 nm and ~66 nm, suggesting that there are four well-defined misfolded dimeric structures. B) The histogram of unfolding force of misfolded superfolds shows a broad distribution with the average force of 96 ± 33 pN.

Skips with $\Delta L_{C\text{-skip}}$ of 66 nm, which is greater than $2\times\Delta L_C$, likely resulted from domain-swapped misfolded dimeric states (272). The folding of such a domain-swapped misfolded state harbors the linker sequence between the two neighboring TNfn3 domains. If the distance between the N-C termini of the skip fold is the same as that of native TNfn3, the unfolding of such a skip should give rise to a $\Delta L_C$ of ~62.5 nm ($((2\times90+2)\times0.36\text{nm/aa-3nm})$, close to the experimentally measured $\Delta L_{C\text{-skip}}$. In contrast, the other three peaks showed $\Delta L_{C\text{-skip}}$ that were significantly greater than $\Delta L_C$ of TNfn3 but smaller than $2\times\Delta L_C$, suggesting that these misfolded structures ($\Delta L_{C\text{-skip}}$ of 36 nm, 48 nm, 54 nm) involve the coalesce and misfolding of only parts of two neighboring TNfn3 domains. Although these misfolding behaviors are complex, they could also be potentially
explained by a domain-swapping mechanism. It was shown that domain-swapping can involve only half of the domain (either the N terminus or the C terminus) (273-275) or both the N-/C- termini (276). Moreover, the same protein can swap into at least two different dimers and one trimer (276-279). These behaviors are reminiscent of those observed here. However, a detailed mechanism for misfolding of TNfn3 domains remains to be established.

It is also intriguing to note that there seemed to be a correlation between the skips and folding intermediate states of TNfn3. After offsetting the $\Delta L_C$-skip by 29 nm (which is $\Delta L_C$ of TNfn3), the location of the two preferred peaks of $\Delta L_C$-skip (indicated by *) coincide with the location of the folding intermediate states 3 and 4 (Fig. 5.4A, inset). This result suggested that the misfolded skips are likely formed by the fusion of a folded TNfn3 domain with one of the folding intermediate states (3 and 4). This conclusion was further supported by our experimental observations. The majority of the skips unfolded in a two-state fashion. However, in a few force-extension curves showing skip events with $\Delta L_C$-skip of 48 nm (Fig. 5.5), we observed that the unfolding event occurred in two steps with a relatively higher force unfolding event with $\Delta L_C$ of 29 nm immediately followed by a significantly lower force unfolding event with a $\Delta L_C$ of ~18nm, which is similar to that of the folding intermediate state #3 in Fig 5.2B. The sum of the two $\Delta L_C$ gives a $\Delta L_C$ of ~48nm. These results suggested that the skip with $\Delta L_C$-skip of ~48nm likely originated from the domain swapping involving a kinetically trapped folding intermediate state #3 and its neighboring TNfn3 domain. However, detailed structures of such misfolded skips remain to be illustrated.
Figure 5.5 Misfolded skips of TNfn3 domains can unfold in two steps. For a small number of misfolding skip events showing $\Delta L_{C\text{-skip}}$ of $\sim$48 nm, we observed a higher force unfolding event with a $\Delta L_C$ of 29 nm immediately followed by a significantly lower force unfolding event with a $\Delta L_C$ of $\sim$18 nm, while the sum of the two $\Delta L_C$ gives a total of 48 nm. This result suggested that the misfolded skip with a $\Delta L_{C\text{-skip}}$ of 48 nm could originate from the fusion of a folding intermediate state ($\Delta L_C$ of $\sim$18nm) with its neighboring folded TNfn3 domain.

5.4 Discussion

5.4.1 Single molecule AFM results provide supporting evidence for the kinetic partitioning mechanism of protein folding

Energy landscape theory has become a powerful tool to understand the folding dynamics of proteins. It predicts that the folding of a protein follows multiple folding pathways on a rugged energy landscape en route to its unique native state. Despite the limited experimental data(66, 139, 178, 227), this view has been generally accepted. It has been proposed that if the scale of roughness of the energy landscape is large enough, kinetic traps will form on the folding landscape and the folding can be described by the kinetic partitioning mechanism: a fraction of protein molecules can fold into their native
states rapidly, while the remaining fraction of proteins follow different folding pathways and are trapped in discrete intermediate states (21, 52, 177). Here we have used single molecule AFM to probe the multiplicity of folding pathways of a small protein TNfn3. By repeatedly unfolding and refolding the same polyprotein (TNfn3)₈ molecule, we obtained a large number of folding and unfolding events from the same molecule, allowing for the determination of the folding pathways taken by the very same protein molecule. We observed that the folding of TNfn3 molecules follows multiple distinct pathways: the majority of TNfn3 molecules fold in a simple two-state manner (which constitutes the statistically preferred folding pathway), and ~5% of TNfn3 molecules were found to be trapped into folding intermediate states that are of well-defined structures. These folding intermediate states are mechanically stable and showed an average unfolding force of ~90 pN, suggesting that these folding intermediate states correspond to deep kinetic traps along the energy landscape. These results suggest that the folding energy landscape of TNfn3 is largely smooth and the majority of TNfn3 fold efficiently into their native states in a two-state fashion. However, deep kinetic traps do exist and led to the formation of well-defined folding intermediate states. Hence, our results provided direct supporting evidence for the kinetic partitioning mechanism for protein folding.

5.4.2 Possible structure of the folding intermediate states

The folding intermediate states reported here are mechanically stable and have well-defined structure, as evidenced by their well-defined ΔL. Assuming that these folded intermediate structures are similar to the native state, we attempted to map the possible structure of these folding intermediate states. Taking into account the sequence
continuity of the polypeptide chain and the boundary of secondary structural elements in TNfn3, we calculated $\Delta L_C$ of different secondary structural elements and identified possible structures for different folding intermediate states (Fig. 5.6). We found that structural elements from the three-dimensional structure of TNfn3 could potentially explain the structure of folding intermediate states #1, #3 and #4 in Fig. 5.2B.

For folding intermediate state #1, three possibilities arose. The unfolding of hairpins C-C’ and F-G would result in $\Delta L_C$ of 7.3nm and 7.7nm, respectively, and the unfolding of C-C’-E strands would result in $\Delta L_C$ of 8.7 nm. It is well known that $\beta$-hairpins can form during the early stage of folding and sometimes can be stable in isolation (for example, the second GB1 hairpin(280)). It is possible that one of the $\beta$-hairpins (C-C’, F-G and C-C’-E) can form during the folding of TNfn3 and give rise to the observed folding intermediate state #1. However, it is intriguing that such $\beta$-hairpins are mechanically stable. We speculate that the interaction of the $\beta$-hairpin with other parts of TNfn3 provide further stabilization of the isolated hairpin.

For the folding intermediate state #3 with $\Delta L_C$ of 18nm, three possible structures emerged: the unfolding of A-B-C-C’-E, C-C’-E-F-G and B-C-C’-E-F would lead to $\Delta L_C$ of $\sim17.4$ nm, 18.6 nm and 18.4 nm, respectively. It is very interesting to note two of the possible folding intermediate states are the very unfolding intermediate states predicted in steered molecular dynamics simulations for the unfolding of TNfn3 domain. SMD simulations on the mechanical unfolding of TNfn3 showed that the stretching of TNfn3 leads to the formation of unfolding intermediate state by unraveling the A-B $\beta$ hairpin, or by peeling off the A and G strands from TNfn3(250, 271). However, the unfolding intermediate state C-C’-E-F-G was observed to be much more stable(271). Such an
unfolding intermediate state was also observed in single molecule AFM experiments, but their occurrence was much rarer than in SMD simulations (Fig. 5.7). Considering the stability of C-C’-E-F-G and B-C-C’-E-F predicted from SMD simulations, we think that the folding intermediate state #3 more likely correspond to C-C’E-F-G. Moreover, in the tenth FnIII domain of fibronectin (FNfn10), mechanical unfolding intermediate C-C’-E-F-G was observed in both SMD simulations and single molecule AFM experiments as an obligatory unfolding intermediate state(135, 254, 262). A naturally occurring domain anastellin from fibronectin is a truncated FnIII domain and its three dimensional structure was shown to be the structure of C-C’-E-F-G(281). These results raised the interesting possibility that the intermediate state formed by the $\beta$-strands C-C’-E-F-G is a general folded structure among FnIII domains.

For the folding intermediate #4 with $\Delta L_C$ of 24nm, two possible structures are A-B-C-C’-E-F and B-C-C’-E-F-G, with predicted $\Delta L_C$ of ~24 nm and ~25nm, respectively. In both cases, one $\beta$ strand is peeled off TNfn3 leading to the possible intermediate state. In SMD simulations of TNfn3, an unfolding intermediate state was observed by peeling off strands A and G(271). Hence, it is possible that a folding intermediate state can exist with one strand detached from the rest of structure. However, it is not possible to determine which of these two strands is more likely to detach first.

Although the contour length increment analysis yielded possible structures for the observed three folding intermediate states, the above analysis did not yield a structurally sensible solution for folding intermediate state #2 (Fig. 5.2B). One possible reason is that the folding intermediate state #2 with $\Delta L_C$ of 12 nm is a mis-folded state containing
non-native interactions and structures. Further experimental and simulation efforts are needed to determine the structures of this folding intermediate state.

Figure 5.6 Possible structures of the observed folding intermediate states. Taking into account of the sequence continuity of polypeptide and the boundary of secondary structural elements, we calculated $\Delta L_C$ of different secondary structural elements in the three dimensional structure of TNfn3. Possible structural elements, which would give rise to the observed folding intermediate states, were thus identified and shown in the figure.
Some TNfn3 can unfold in two steps, through an unfolding intermediate state. The four force-extension curves presented show five three-state unfolding events of TNfn3. B) Histogram of the $\Delta L_{C1}$ (blue) and $\Delta L_{C2}$ (red) of the three-state unfoldor. Typically, the first step is shorter than the second step. C) Histogram of the sum of ($\Delta L_{C1} + \Delta L_{C2}$). The sum of ($\Delta L_{C1} + \Delta L_{C2}$) shows a narrow distribution with the average of 29.7±1.1 nm. This narrow distribution indicate that the relatively broader distribution of the histogram for $\Delta L_{C1}$ and $\Delta L_{C2}$ reflects the intrinsic properties of the unfolding intermediates, which implies multiple three-state unfolding pathways.
In previous SMD simulations, there are two possible unfolding intermediates along the unfolding pathways of TNfn3. For the first one, the A-B β-hairpin is mechanically stretched leaving the remaining part folded. In this pathway, the $\Delta L_{C1}$, corresponding to the transition from the native to the intermediate state, is about 11nm and the $\Delta L_{C2}$, corresponding to the transition from the intermediate to the unfolded state, is about 18nm. For the second one, the A- and G-strand are unstructured simultaneously. The SMD simulation also suggested that the first several residues of the A-strand is weakly attached and will lose its structure at the very beginning of trajectory. If the first eight residues of the A-strand and the whole G-strand are pared off from the main body of TNfn3 in the first step of unfolding, the corresponding $\Delta L_{C1}$ will be about 8 nm and the left-over structure will give a $\Delta L_{C2}$ of about 21 nm. Therefore, the two observed mechanical pathways are consistent with the SMD simulation results.

5.4.3 Misfolding behavior of neighboring TNfn3 domains

Due to the complexity of protein folding, it was proposed that misfolding is an inevitable phenomenon during the folding process(142). *In vivo*, the cell has a very sophisticated system to prevent the polypeptide chain from misfolding or to degrade the misfolded ones. If this system fails, the misfolded proteins will accumulate in the cell and form the amyloid fibrils which has often been related to a novel concept of “conformational disease”(142, 282-284), which includes some most perplexing medical problems, such as Alzheimer’s disease (AD) and Huntington’s disease. For protein aggregation and amyloid formation, interactions between different protein domains are important(285). Domain-domain interactions will inevitably influence and shape the
folding energy landscapes of individual domains and potentially lead to misfolding. Our results on the TNfn3 polyprotein provided new cases for such scenarios.

In our single molecule AFM experiments on polyprotein (TNfn3)$_8$, most TNfn3 domains fold independently of each other in a homopolyprotein setting. However, the folding of a small percentage (~3%) of TNfn3 was influenced by neighboring TNfn3 and formed misfolded superfolks (skips). It seems that the formation of the misfolded structures is correlated with the folding intermediates of TNfn3. The misfolding behavior of TNfn3 could be the result of the folding intermediates which act as “seeds” to induce the neighboring domains to swap with themselves. This behavior is similar to the mechanism whereby the prion protein infects the normal protein by inducing the remodeling (misfolding) of the normal protein(286, 287).

The formation of such skips indicated that there are weak domain-domain interactions between TNfn3 domains, leading to the roughness of the folding energy landscape and misfolding of neighboring domains. Identifying key mutations that will promote the misfolding of TNfn3 will be important future tasks for understanding the rich misfolding patterns of TNfn3 domains. For these efforts, combining protein engineering, molecular dynamics simulations and single molecule AFM techniques will be critical.

Although homo-polyproteins do not occur abundantly in nature, tandem modular proteins are very common. Thus, the insights we obtained here should also have important implications for tandem modular proteins, as weak domain-domain interactions could have a similar effect on their folding energy landscape. For example, about 4% of the FnIII domains of tenascin-C were found to fold into misfolded superfolks during single-molecule AFM experiments(272). Moreover, it was proposed that anestellin,
whose three dimensional structure is very similar to that of folding intermediate state #3, can serve as an amyloid precursor through domain swapping (281). There results highlight the importance of the weak domain-domain interactions in the folding of tandem modular proteins. Similar methods can be used to understand the misfolding mechanism of some common prion or amyloid-forming proteins(288).

5.5 Experimental section

5.5.1 Protein engineering

The gene encoding the polyprotein (TNfn3)₈ was constructed as described (272). (TNfn3)₈ was overexpressed in DH5α strain and purified from supernatant using Co²⁺-affinity chromatography. The polyprotein was kept at 4°C in PBS at a concentration of ~200 μg/mL.

5.5.2 Single molecule AFM experiments

Single-molecule AFM experiments were carried out on a custom built atomic force microscope(88). All the force-extension measurements were carried out in PBS buffer. In a typical experiment, the polyprotein sample (1 μL) was deposited onto a clean glass cover slip covered by PBS buffer (50 μL). The spring constant of each individual cantilever (Si₃N₄ cantilevers from Vecco, with a typical spring constant of 40 pN/nm) was calibrated in solution using the equipartition theorem before and after each experiment.
Chapter 6: Direct observation of the tug-of-war during the folding of a mutually exclusive protein*

A version of Chapter 6 has been published as “[Peng Q.], Li H. (2009) Direct observation of tug-of-war during the folding of a mutually exclusive protein. Journal of the American Chemical Society, 131(37):13347-54”. DOI: 10.1021/ja903480j

According to the copyright policy of American Chemical Society (ACS), the republishing of ACS full articles in a thesis/dissertation on a website is not permissable. Therefore, Chapter 6 has been removed when the thesis is published on-line.

As an alternative, the link to the article's DOI is provided as following:

http://pubs.acs.org/doi/full/10.1021/ja903480j

---

*A version of this chapter has been published as “[Peng Q.], Li H. (2009) Direct observation of tug-of-war during the folding of a mutually exclusive protein. Journal of the American Chemical Society, 131(37):13347-54”.

137
Chapter 7: Summary and prospects

7.1 Summary

Protein folding/unfolding has been studied for more than fifty years (308, 309), yet many aspects remain an open question. As more and more novel powerful techniques are being invented (7, 24, 32, 35, 39, 44, 58, 67-78, 310, 311), researchers are getting closer to revealing the mystery of protein folding. Single molecule AFM is one of the techniques which have empowered the scientists to extract more and new insights besides the current view of protein folding/unfolding (84, 85). Our studies on protein folding/unfolding using single molecule AFM, especially on the aspect of protein mechanical folding/unfolding, have pushed the general understanding on protein folding enigma one big step further towards its final answer.

Here in my Ph.D thesis, I have accomplished three main projects to date:

1) For the first time, we have studied the mechanical folding/unfolding kinetics and pathways of two domain-insertion proteins using single molecule AFM: one natural domain-insertion protein T4-lysozyme (T4L)(227) and one artificially designed domain-insertion protein GL5/T4L, where GL5 is a loop insertion mutant of protein GB1(312).

For T4L, a non-mechanical protein, it was found mechanically labile with an unfolding force of 50 pN, which is not surprising and consistent with a previous study(93). However, by pulling the T4L from its N-/C- termini, which is different from the earlier study, we identified that the mechanical stability of T4L arises from a unique structural arrangement which has never before been reported. The interaction between the
helix A and the remainder of the C-terminal subdomain controls the unfolding cooperativity of the two sub-domains of T4L and provides T4L the mechanical resistance to external force. Our results revealed the importance of the topological organization. In the case of T4L, helix A is selected during the evolution to be the N-terminus of the sequence while it is part of the C-terminal subdomain in structure. In this way, helix A couples the two sub-domains and underlies the communication between the two sub-domains. This conclusion is further supported by the results from a circular permutant of T4L – PERM1(201), in which the helix A was relocated to the C-terminus of the whole sequence. Therefore, the communication between the two sub-domains is severed and the unfolding cooperativity is lost. Recently, another group published a parallel paper with similar conclusions (313). Besides these, we also detected the existence of multiple unfolding pathways between the folded and unfolded states of T4L. The mechanical unfolding of T4L, when pulled from its N-/C- termini, does not follow a well defined pathway. Instead, there are abundant routes underlying the mechanical unfolding of T4L. Our findings are the first direct experimental proofs supporting the kinetic partitioning assumption (multiple folding pathways) for protein folding at the single molecule level.

GL5/T4L is an artificially designed domain-insertion protein in which the mechanically labile T4L is spliced into a flexible loop of the mechanically stronger host domain GL5 without major disruption to the structures of either protein. The mechanical properties of the hybrid protein GL5/T4L were determined using single molecule AFM and the folding/unfolding dynamics of GL5/T4L were monitored at the single molecule level. Our results demonstrated that we have successfully reversed the mechanical
unfolding hierarchy of GL5 and T4L through domain insertion. In GL5/T4L, the mechanically labile T4L domain unfolds only after the mechanically stronger host domain GL5 has unfolded, even though the reverse would usually be expected. The host GL5 can effectively protect the guest T4L from “seeing” the mechanical force and prolong the lifetime of T4L by >1000 times. The protein can also be thought of as a mechanically controlled enzyme switch that could be switched off by applying force. Therefore, we proved that domain insertion is an effective strategy to protect mechanically labile domains, which opens the possibility to incorporate labile proteins into elastomeric proteins for engineering novel multifunctional elastomeric proteins for novel applications in nanomechanics/nanobiotechnology. This is also a new concept to program the mechanical unfolding pathway of multi-domain proteins.

2) Tenascin-C is a typical extracellular matrix protein which experiences repetitive force-induced unfolding/refolding reactions in vivo to fulfill its physiological functions. To obtain a detailed molecular interpretation of the mechanical folding/unfolding of Tenascin-C is therefore critical for understanding its mechanical design and physiological functions. As the only fibronectin type III domain of tenascin-C with its three dimensional structure solved, the third fibronectin type III domain of tenascin-C (TNfn3) has been chosen as our model. Combining single molecule AFM with SMD simulations and protein engineering, the mechanical folding/unfolding of TNfn3 has been studied in detail(123, 271).

Through the mechanical unfolding studies on TNfn3, the backbone H-bonds of TNfn3 were found to be critical for its mechanical stability and the mechanical design of
TNfn3 was found to be robust. Proline mutagenesis revealed that the mechanical stability of TNfn3 is very resistant to structural disruptions caused by proline substitutions in β sheets. Proline mutation at the site F88 is one exception, which led to a labile mutant that unfolds below 20 pN. The residue Phe88 is identified as the weakest point of the mechanical resistance for TNfn3. Combining SMD simulations together, the molecular details underlying the mechanical unfolding of TNfn3 have been revealed. The mechanical unfolding and design of TNfn3 were found different from its structural homologue, the tenth FnIII domain from fibronectin(135, 156). Our results demonstrated the sophisticated mechanical design of FnIII domains and established the basis for further study on other FnIII domains in tenascins.

On the other hand, we also characterized the mechanical folding pathway(s) of TNfn3 systematically at the single molecule level. Multiple parallel folding/unfolding pathways of a protein have been proposed by theorists in the field of protein folding. Yet, the supporting experimental proofs are still few. In Chapter 2, we have already presented the first experimental evidence that a same T4L molecule can be mechanically unfolded \textit{via} multiple unfolding pathways at the single molecule level. Furthermore, our single molecule AFM investigation on the folding process of TNfn3 revealed that an individual TNfn3 molecule can sample multiple parallel folding pathways to fold back into its native structure after being mechanically unravelled. It has also been shown that the folding of TNfn3 can be influenced by its neighboring TNfn3 domains when multi-copy of TNfn3 domains are polymerized into the tandem arrangement. In such cases, the adjacent TNfn3 domains can misfold into some domain-swapped dimeric structures. Our studies revealed the ruggedness of the folding energy landscape of TNfn3, and provided
direct experimental evidence that the folding dynamics of TNfn3 is governed by the kinetic partitioning mechanism.

3) In order to mimic the natural domain-insertion proteins, which are typically difficult to study directly, and hence study their folding/unfolding behaviours, we designed a novel mutually exclusive protein GL5/I27w34f (I27w34f: a tryptophan removed mutant of 27th immunoglobulin domain from the protein titin)(314). The host protein is designed to be kinetically more favourable but less stable in thermodynamic terms. However, the guest protein is designed to be thermodynamically more stable but a slower folder. Therefore, in principle, upon triggering the folding reaction, the host protein will fold first. The guest protein starts to fold soon after and in doing so mechanically unfolds the host protein. Relying on the stopped-flow technique and fluorescence measurements, for the first time, we captured direct evidence of this folding tug-of-war. Using single molecule AFM, we also verified that the system reaches equilibrium in the end and both conformers—folded guest with unfolded host, and unfolded guest with folded host—exist. The balance of this tug-of-war can be affected by other factors such as chemical denaturants.

If protein folds properly, it works properly. Problems of protein folding underlie conditions such as Alzheimer’s disease and transmissible spongiform encephalopathies (TSEs)(315-317), which are always untreatable and fatal diseases in medicine today. Therefore, the molecular mechanism behind protein folding drives much research into it. The general ways to study protein folding is often using denaturants, such as chemical denaturing agents, temperature, pressure, and pH, to regulate the folding process.
However, the denaturants can produce undesired changes in the protein. Using mutually exclusive proteins is one of the alternative ways to control protein folding without utilizing potentially problematic solution conditions. Our study not only provided the primary proof of the folding tug-of-war for the mutually exclusive protein, but also testified, again for the first time, the long-time hypothesis that protein folding can generate sufficient mechanical strain to unravel a host protein. Mutually exclusive proteins provide us with a new system for manipulating protein folding.

In summary, our results demonstrated the unique capability of using single molecule AFM to probe the folding/unfolding dynamics of proteins at the single molecule level. Our results directly support the assumption that protein folding/unfolding is governed by kinetic partitioning (52). When combined with SMD simulations, protein engineering, stopped-flow fluorescence measurements or other techniques, single molecule AFM can be a very powerful technique for protein folding studies and will provide more insights on this problem in future.

7.2 Prospects

7.2.1 Revealing the molecule mechanism of protein misfolding

As reported in the Chapter 5, during the folding process of a protein, there is a probability, typically low but non-negligible, that a protein molecule can misfold(142, 272). During misfolding, the protein molecule is organized into a structure different from its native structure. Protein misfolding is a fundamental process that remains a puzzle despite being a process critical to life(315). If proteins fold improperly, they work improperly. Protein misfolding is the underlying cause of conditions such as Alzheimer’s
disease and transmissible spongiform encephalopathies (TSEs) (315-317), which are always untreatable and fatal diseases characterized by neurodegeneration, dementia, and paralysis wasting. Therefore, the underlying question behind why proteins misfold drives much research into the molecular mechanism of protein misfolding. However, our understanding on the mechanism of protein misfolding is still limited.

One of the reasons for our poor understanding on protein misfolding is that the misfolding process of a protein is typically inaccessible for most of the current experimental methods, especially for the ensemble measurements(318). The key nodus for tracing the protein misfolding process is that the misfolding species only exist transiently in a minor population during the misfolding at any given time. Since protein misfolding is a low probability event, it is impossible to synchronize the whole population for misfolding. To study the misfolding dynamics in real time, the capability to manipulate protein molecules individually is desirable. Single molecule AFM has been proven to be a powerful tool to directly study the protein misfolding at the single molecule level(272). By first mechanically unfolding an individual protein molecule and then quenching the stretching force applied onto the molecule, in real time, the end-to-end distance of the molecule can be monitored by using single molecule AFM and the force signal generated during the refolding will also be recorded(88). In this way, the folding/refolding cycles of the same molecule can be measured for many times or the folding/refolding cycle can be measured on many individual molecules, one molecule at a time. Given the misfolding happens stochastically, when the size of the recorded data pool is big enough, the misfolding process of a protein will be detected as long as it occurs. Typically, the misfolded protein exhibits different characteristics from their
properly folded counterpart, either showing a distinct mechanical stability or involving different numbers of amino acid residues into the folded structure. In both cases, the trajectories resulting from misfolded protein molecules can be easily identified by using single molecule AFM. From such recordings, a lot of insightful information can be extracted. This information will be important for elucidating the molecular mechanism of protein misfolding, a fundamental process of life. The studies on protein misfolding using single molecule AFM will also be a clear contribution to the understanding of protein folding and may open up a new regime of single molecule force microscopy research. In the long run, if the molecular mechanism of protein misfolding can be clarified, it will have a profound impact on disease and health studies. Therefore, the study of protein misfolding will continuously be one of the hot spots for single molecule AFM study.

7.2.2 Revealing the folding mechanism of the transmembrane proteins

Membrane proteins, especially transmembrane proteins, are a very special type of protein (319, 320). The polypeptide chain of transmembrane proteins can cross the phospholipid membrane one or more times when they exist in their native states. There are two dominant types of transmembrane proteins: the cross membrane channels, such as Na\(^+\)/K\(^+\) channel and H\(_2\)O channel, and the transmembrane receptors. Transmembrane receptors are specialized integral proteins that locate themselves onto the membrane and are involved in the communication between the cell and the extracellular environment(321). If the membrane receptors are misfolded or deficient under some conditions, the cellular communication will be hindered and lead to diseases(322, 323). Furthermore, the transmembrane receptors can also be the targets of many viruses(324,
To elucidate those important processes, knowledge of the structure and the corresponding folding mechanism of transmembrane proteins is necessary.

Hundreds of different transmembrane proteins have been discovered and many more are yet to be identified. However, except for a few examples, such as the G-protein coupled receptors, the structure of transmembrane proteins is unknown in any detail for most of them. One of the major difficulties for the transmembrane proteins studies is that those proteins or protein assemblies are typically insoluble or unstable when they exist themselves alone. To form their stable and functional structure, they need to be imbedded into the phospholipid membrane. Presently, the major techniques for protein structure study, such as X-ray crystallography and NMR spectroscopy, are mainly targeting the pre-purified water soluble proteins. Since most pure transmembrane receptors are insoluble or (and) unstable, they are often not accessible for either X-ray crystallography or NMR spectroscopy. Although about 30 percent of all genes in human encode transmembrane proteins, our understanding of transmembrane proteins is extremely limited comparing with their importance. Techniques that allow us to investigate membrane proteins in a native membrane or a reconstructed lipid bilayer are thus of interest.

Single molecule AFM provides a new tool for studying the transmembrane proteins in their native state within their physiological-like environment. In a single molecule AFM experiment, what you see is what it is. In each time, only one molecule is picked up and manipulated. Therefore, in principle, there is no strict constraint on the sample purity, since the observation on contaminations can be easily singled out due to their distinct responses. In this case, single molecule AFM can be used to study the
transmembrane proteins without extracting them out of the lipid bilayer with detergents, which means the transmembrane proteins can even be studied at the surface of a living cell. Several pioneer groups have done some investigations on the membrane protein bacteriorhodopsin using single molecule AFM. F. Oesterhelt’s group, H. E. Gaub’s group and D. J. Muller’s group have collaborated together to study the unfolding pathways of individual bacteriorhodopsins (BR) embedded in the native purple membrane patches from *halobacterium salinarum* (329). Individual BR molecules localized into the membrane were extracted one by one using single molecule AFM. When extracted, the seven transmembrane $\alpha$-helices of BR can be pulled out of the membrane and unfolded. For different transmembrane $\alpha$-helices, the respective extracting force varies. The anchoring forces between BR and the membrane were found ranging from 100 to 200 pN(329). The resulted force-extension curves corresponding to the extraction of BR from the membrane disclosed the unfolding pathways of individual BR molecules. It was found there are two pairs of helices which always unfold in pair: helices G/F and helices E/D. Occasionally, helices B and C were observed to unfold one after the other. By employing a loop cleaved mutant of BR, the origin of the diagnostic unfolding pathway of BR is identified as the stabilization effect of neighboring helices on helix B.

Besides the mechanical unfolding BR, H. E. Gaub group and D. J. Muller groups have further studied the folding of BR into the membrane against an external force(330). An individual BR molecule was first partially unfolded and stretched out of the purple membrane from it C-terminus. Then the refolding process of BR can be traced by reducing the external force. The refolding into the membrane of BR can generate force as big as several tens of pN. The upper limit for $\Delta G$ (the change of Gibbs free energy) of BR
folding was estimated based on the mechanical work done by BR during the folding. These studies demonstrate that single molecule AFM can precisely handle individual membrane proteins and control their conformations at the level of secondary structure elements.

Lots of new drugs are targeting membrane proteins(331). Knowledge on the structure of membrane proteins is in imperative need. Exploring the membrane protein with sophisticated single molecule AFM seems very promising. The study of membrane proteins will be one of future focuses for single molecule AFM research.

7.2.3 Designing novel molecular indicators/probes

The mechanical folding/unfolding of many proteins has been studied(88, 91, 92, 122, 126-128, 145, 146). Besides contributing to our understanding on the protein folding/unfolding, the information extracted from those studies can also imply some applications. It has been proved by several groups independently that a protein can generate significant, at least detectable, force (> several tens of pN) during folding(124, 125, 144-146, 330). In our own experiments (chapter 6), we also demonstrated that the force generated by a protein during its folding is big enough to trigger the unfolding of another protein(314). Therefore, the mechanical strains could be ruled as a driving force to activate a protein based probe/indicator. One of our future objectives is to design some mutually exclusive protein based molecular probes. For the two subdomains of the mutually exclusive protein, they will be artificially designed so that they will be stabilized under distinct conditions, respectively. For example, say, the host protein with some special properties is relatively more stable than the guest protein under condition A,
while the relative stability of the host and guest domains will be reversed under condition B. Therefore, the mutually exclusive protein will have its host domain folded under condition A and leave the guest domain unfolded. When the environment changes from condition A into condition B, the stimulation will trigger the folding of the guest domain which will lead to the unfolding of the host domain. Accompanying the unfolding the host domain, the probe will also lose the characteristic properties belonging to the host domain. Therefore, by observing the emergence or the disappearance of the stamp of the host domain, we can detect the change of the probe and hence the alteration of the environment where the probe exist.

To design this kind of molecular probes, the first step is to fully understand and rationally tune the mechanical properties and structural information of the host domain and the guest domain, respectively. Single molecule AFM is the ideal tool to fulfill this task. Computer design or directed evolution can also be adopted to optimize the properties of the enrolled proteins. Our previous experience on the engineering and characterization of the mutually exclusive proteins will be a solid ground for the further study on the construction of the novel molecular probes with potential applications. The study of the controllable molecular probes will be another focus of our future endeavours.
References


of America 97, 6521-6526.
America 97, 1525-1529.
Appendix A: Polyprotein engineering

A1. Sequences of proteins and the encoding cDNAs

A1.1 pseudo wild type T4-lysozyme (T4L*)
Protein:
MNIFEMLRIDEGLRLKIDTEGYYTIGIGHLLTAKSPSLNAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGLRNAKLKPVYDLSDAVRRAALINMVQMGETGVAGFTNSLRLMQKRWDEAAVNLAKSRWYNQTPNRAKRVTITFTRTGTWDAYKNL

cDNA:
ATGAATATATTGAAATGTTACGTAATAGTAAGGTCTTAGACTTAAAATCATAAAAGACACAGAGGCTATTACAACACTATTGGAATCGGATTTGCTTACAAAAAGTCCATCACTTAATGCTGCTAAATCTGAATAGATAAAGCTATTGGGCGTAATACTAATGGTGTAATTACAAAAGATGAGCTGAAAAACTCTTTAATCAGGATGTTGATGCTGCTGTTCGCGGAATTCTGAGAAATGCTAAATTAAAACCGGTTTATGATTCTCTTGATGCGGTTCGTCGCGC

A1.2 PERM1
Protein:
MLRLKIYKDTEGYYTIGIGHLLTAKSPSLNAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGLRNAKLKPVYDLSDAVRRAALINMVQMGETGVAGFTNSLRLMQKRWDEAAVNLAKSRWYNQTPNRAKRVTITFTRTGTWDAYKNLSGGA

A1.3 wild type GB1
Protein:
MDTYKLILNGKTLKGETTTEAVDAATAEKFQYANDNGVDGEWTYDDATKFTTVTE
cDNA:  
ATGGACACCTACAAACTGATCCTGAACGGTAAAAACCCTGAAAGGTGAAC  
CACCACACAAAAAGCTGAGACGCTGCTACTGCAAAGATTTTCAAACAGTA  
CGCTAACGAACACGGTGTCGACGTGTAATGGACCTACGACGACGCTACCA  
AAACCTTCACGGTTACCGAA

A1.4 GB1-L5 (GL5)  
Protein:  
MDTYKLILNGKTLGETTTEAVDAATAEKFQYANDNGVGGGLGDGEWTYDDATKTFVTVE

cDNA:  
ATGGACACCTACAAACTGATCCTGAACGGTAAAAACCCTGAAAGGTGAAC  
CACCACACAAAAAGCTGAGACGCTGCTACTGCAAAGATTTTCAAACAGTA  
CGCTAACGAACACGGTGTCGACGTGTAATGGACCTACGACGACGCTACCA  
AAACCTTCACGGTTACCGAA

A1.5 GL5/T4L  
Protein:  
MDTYKLILNGKTLGETTTEAVDAATAEKFQYANDNGVGGGLGDGEWTYDDATKTFVTVE

where the sequence in italic is from the host domain GL5, and the sequence in bold is from the guest domain T4L. The junction between GL5 and T4L is Leu-Gly resulted from *Av*(a)l site.

cDNA:  
ATGGACACCTACAAACTGATCCTGAACGGTAAAAACCCTGAAAGGTGAAC  
CACCACACAAAAAGCTGAGACGCTGCTACTGCAAAGATTTTCAAACAGTA  
CGCTAACGAACACGGTGTCGACGTGTAATGGACCTACGACGACGCTACCA  
AAACCTTCACGGTTACCGAA
A1.6 wild type TNfn3
Protein:
RLDAPSQIEVKDVTDTTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLTEDEN
QYSIGNLKPDEYEVLISRRGDMSSNPAKETFTT
cDNA:
CGCTTGGATGCCCCCGCAGCCAGATCGAGGTGAAGAGATGTACAGACAGACACCAC
GCCTTGATCACTGGTCTAAGCCCTCTGGGCATGATCGATGGCATTAGCTG
ACCTACGCGCATCAAGAGGTGGAGGACCGTACCACCTAGCTACATCA
GAGGAACGAGAACAGTACTCCATCGGGAAACCTGAAGCCTGACACTGA
CGAGGTTCCTACTCTCCGGCAAGGTGACATGTCAAGCAACCCAGCCAA
AGAGACCTTCACAACAA

A1.7 S6P(TNfn3)
Protein:
RLDAPPQIEVKDVTDTTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLTEDEN
QYSIGNLKPDEYEVLISRRGDMSSNPAKETFTT
cDNA:
CGCTTGGATGCCCCCGCAGATCGAGGTGAAGAGATGTACAGACAGACACCAC
GCCTTGATCACTGGTCTAAGCCCTCTGGGCATGATCGATGGCATTAGCTG
ACCTACGCGCATCAAGAGGTGGAGGACCGTACCACCTAGCTACATCA
GAGGAACGAGAACAGTACTCCATCGGGAAACCTGAAGCCTGACACTGA
CGAGGTTCCTACTCTCCGGCAAGGTGACATGTCAAGCAACCCAGCCAA
AGAGACCTTCACAACAA

A1.8 E9P(TNfn3)
Protein:
RLDAPSQIEVKDVTDTTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLTEDEN
QYSIGNLKPDEYEVLISRRGDMSSNPAKETFTT
cDNA:
CGCTTGGATGCCCCCGCAGATCGAGGTGAAGAGATGTACAGACAGACACCAC
GCCTTGATCACTGGTCTAAGCCCTCTGGGCATGATCGATGGCATTAGCTG
ACCTACGCGCATCAAGAGGTGGAGGACCGTACCACCTAGCTACATCA
GAGGAACGAGAACAGTACTCCATCGGGAAACCTGAAGCCTGACACTGA
CGAGGTTCCTACTCTCCGGCAAGGTGACATGTCAAGCAACCCAGCCAA
AGAGACCTTCACAACAA

A1.9 K11P(TNfn3)
Protein:
RLDAPSQIEVPDVDTDTTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLTEDEN
QYSIGNLKPDEYEVLISRRGDMSSNPAKETFTT
cDNA:
CGCTTGGATGCCCGAGCCAGATCCCGGATGGAAGAGATGTACAGACAGACACCAC
GCCTTGATCACTGGTCTAAGCCCTCTGGGCATGATCGATGGCATTAGCTG
ACCTACGCGCATCAAGAGGTGGAGGACCGTACCACCTAGCTACATCA
GAGGAACGAGAACAGTACTCCATCGGGAAACCTGAAGCCTGACACTGA
CGAGGTTCCTACTCTCCGGCAAGGTGACATGTCAAGCAACCCAGCCAA
AGAGACCTTCACAACAA
ACCTACGGCATCAAAGACGTGCCAGGAGACCGTACCACCATCGATCTCACA
GAGGACGAGAACCAGTACTCCATCGGGAAACCTGAAGGCTGACACTGAGTA
CGAGGTGTCTCCCTCATCTCCCGCAGAGGTGACATGTCAAGCAACCCAGCAA
AGAGACCTTCACAACA

A1.10 T14P(TNfn3)
Protein: RLDAPSQIEVKDPDTTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLTEDEN QYSIGNLKPDTYEYEVSLISRGGDMSSNPaketFTT
cDNA: CGCTTGGATGCCGCCGCCAGCATCGAGGTGAAGATGTCGGCAGTCATCGAGAGCTGAGCTG
ACCTACGGCATCAAAGACGTGCCAGGAGACCGTACCACCATCGATCTCACA
GAGGACGAGAACCAGTACTCCATCGGGAAACCTGAAGGCTGACACTGAGTA
CGAGGTGTCTCCCTCATCTCCCGCAGAGGTGACATGTCAAGCAACCCAGCAA
AGAGACCTTCACAACA

A1.11 A84P(TNfn3)
Protein: RLDAPSQIEVKDVTDRTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLTEDEN QYSIGNLKPDTYEYEVSLISRGGDMSSNPKETFTT
cDNA: CGCTTGGATGCCGCCGCCAGCATCGAGGTGAAGATGTCGGCAGTCATCGAGAGCTGAGCTG
ACCTACGGCATCAAAGACGTGCCAGTACCACCAGGCGAAGGCTGACACTGAGTA
CGAGGTGTCTCCCTCATCTCCCGCAGAGGTGACATGTCAAGCAACCCAGCAA
AGAGACCTTCACAACA

A1.12 E86P(TNfn3)
Protein: RLDAPSQIEVKDVTDRTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLTEDEN QYSIGNLKPDTYEYEVSLISRGGDMSSNPaketFTT
cDNA: CGCTTGGATGCCGCCGCCAGCATCGAGGTGAAGATGTCGGCAGTCATCGAGAGCTGAGCTG
ACCTACGGCATCAAAGACGTGCCAGTACCACCAGGCGAAGGCTGACACTGAGTA
CGAGGTGTCTCCCTCATCTCCCGCAGAGGTGACATGTCAAGCAACCCAGCAA
AGAGACCTTCACAACA
A1.13 F88P(TNfn3)
Protein:
RLDAPSQIEVKDVTDTTALITWFKPLAEIDGIELTYGIKDVGDRRTTIDLTEDENQYSIGNLKPDEYTEVELISRRGDMSSNPAKEFTT

A1.14 T90P(TNfn3)
Protein:
RLDAPSQIEVKDVTDTTALITWFKPLAEIDGIELTYGIKDVGDRRTTIDLTEDENQYSIGNLKPDEYTEVELISRRGDMSSNPAKEFTPT

A1.15 F88A(TNfn3)
Protein:
RLDAPSQIEVKDVTDTTALITWFKPLAEIDGIELTYGIKDVGDRRTTIDLITEDENQYSIGNLKPDEYTEVELISRRGDMSSNPAKEFTATT

A1.16 I27w34f
Protein:
LIEVEKPLYGVEVFVGETAHFEIESEPDVHGFKLKGQPALAESPDCEIIEDGKKHILILHNCQLGMTGEVSFAANTKSAANLKVKEL

cDNA:
CTAATAGAAGTGGAAAAGCCTCTGTACGGAGTAGAGGTGTTTGTTGGTGAAACAGCCCACTTTGAAATTGAACTTTCTGAACCTGATGTTCACGGCCAGTTTA
AGCTGAAAGGACAGCTTTGCGACTTCCCCTGACTGTGAATCATCATTGAGG
ATGGAAAGAGCATATTCTGATCTTTCAATGACGCTGAGTTGATGACAG
GAGAGCTTCTCGAGGCTGCTAAATACCAAATCTGAGCGGAAATCATTGAAAG
TGAAAGAATTG

A1.17 GL5/I27w34f
Protein: 
MDTYKLILNGKTLKGETTTEAVDAATAEKFQYANDNGVGGGLLIEVEKPLYG
VEYFVGGETAHFEIELSEPVDHGQFKLKGQPLAASPDCEIEDGKKHLILHN
CQLGMCSTEVFQAANTKSAANLKVKELlLGEDWYDDATKTFVTE  
where the sequence in italic is from the host domain GL5, and the sequence in bold is from
the guest domain I27w34f. The junction between GL5 and I27w34f is Leu-Gly resulted
from Avf site.

cDNA:
ATGGACACCTACAAACTGATCCTGAACGTTAAAAACCTGAAAGGTGAAAAC
CACCACCAAGCTCTGAGCGCTCTACGTCAAGAATACAGTAAAAAGTTTTCAAAACAGTA
CGCTAACGACACGCGGCTGCTGAGCGACTCGGCTAGTAAACAGTGGAAA
AGCTCTGTACGGAGTGAGGTTGGGTAAAAAGGTTTCAACACTTGAAA
TTGAACTTCTGAACTGATGTTCCACGACCTTTTTTGTTGATGAGGAGGAGCTA
CTTGGGTCGACCTTCCCCTGACTGTGAAAATCATTTAGGATGGAAGAAGCATATA
TTCTGATCCCTACATGACTTGAGGTATGACAGAGAGGTTTTCTGCCCA
GGCTGCTATACAAATCTGACGCCAATCTGAAAGAATGAGATTTCTCGGG
GGACGGTGAATGACCTACGACGACGCTACAAAAACCTTACGGTTACCGG
AAGACCTTGTTGCTAATAQ

A2. Engineering the polyprotein from the gene level

To avoid the undesired influence on the protein structures induced by chemical
modification, we adopted the well-established multiple-cloning strategy to engineer the
polyproteins at the gene level. To facilitate the protein gene engineering, two sets of three
restriction sites are chosen for the cloning: a) KpnI, BamHI and BglII; b) HindIII, BamHI
and BglII (Fig. A1). These restriction sites are introduced at the beginning and ending of
the gene of the protein of interest by polymerase chain reaction (PCR).
Figure A1. The restriction sites we used for protein engineering. After digestion, the sticky end generated by BamHI and BglII are identical.

After screening the available restriction enzymes, we found that, after digesting their target restriction site, the sticky ends generated by enzymes BamHI and BglII are identical (GATC). Therefore, the species generated by the digestion of these two enzymes could be cross-ligated into a perfectly matched double-strand DNA (GGATCT or AGATCC) which cannot be digested by neither BamHI nor BglII. This is the reason that we choose BamHI and BglII as the working restriction enzymes (Fig. A1). Referring to the well-established multiple clone strategy, we can construct the gene of polyproteins through multiple cloning steps. Figure A2 demonstrated the idea of how to build the dimmer (TNfn3)_2 based on the gene of monomer TNfn3. First, the plasmid containing the sequence of TNfn3 is digested by restriction enzymes BamHI and KpnI which will yield the insert; Second, another portion of plasmid containing the sequence of TNfn3 is digested by restriction enzymes BglII and KpnI which will yield the vector; Third, the insert and the vector are mixed in a ratio of 3:1. T4-ligase is used to ligate the insert and vector into a circular plasmid; Fourth, the resulted circular plasmid DNA molecules are transformed into the competent cells of E. coli (XL1/Blue strain); Fifth, the clones
containing the correct dimmer (TNfn3)₂ sequence will be screened out and incubated to amplified the gene of (TNfn3)₂; Finally, the plasmid containing the sequence of (TNfn3)₂ is purified from the E coli cells. By repeating these steps, we can easily further construct the sequences of tetramer (TNfn3)₄ and octamer (TNfn3)₈.

Typically, when we are building the gene sequence of the polyproteins, we work on the vector puC19. After we synthesized the gene sequence of the final construct, we will clone the final sequence into the expression vector pQE80L and transform the resulted plasmid into the expression strain of E coli. (DH5α). Within the pQE80L vector, there are sequences encoding a T5 promoter and a 6-histidine tag at the N-terminal of the polyprotein for the purpose of affinity chromatography purification.

Besides constructing the homo-polyprotein which contains multiple copies of the identical protein domain of interest, we can also build the hetero-polyprotein. One way is utilizing the well identified tetramer protein (GB1)₄ as the handle to flank the protein of interest. Therefore, the final construct can be built more efficiently and the tetramer protein (GB1)₄ can both increasing the protein expression level and facilitate the single molecule AFM experiment. Figure A3 demonstrats the gene of the final construct of (GB1)₄/T4-lysozyme/(GB1)₄ in the expression vector pQE80L. This schematic also shows how to use tetramer protein (GB1)₄ to flank the target protein domain.
Figure A2. General procedure for engineering the gene of polyprotein. Here we demonstrated how to build the gene of the dimmer \((\text{TNfn3})_2\) from the gene of the monomer TNfn3. Similarly, by repeating this procedure, the gene of the tetramer \((\text{TNfn3})_4\) and the octamer \((\text{TNfn3})_8\) can be constructed sequentially.
Figure A3. The schematic of the gene of the constructed polyprotein chimera (GB1)₄/T4-lysozyme/(GB1)₄. Slightly different from the construction of the homo-polyprotein (TNfn3)₈, there is only one copy of the gene for the protein of interest in the construct. The well identified GB1 tetramer (GB1)₄ was used as handles to flank the T4-lysozyme from both ends. The gene of the whole construct ((GB1)₄/T4-lysozyme/(GB1)₄) was inserted into the expression plasmid vector pQE80L.
Figure A4 The photos of DNA electrophoresis in the agarose gel for the genes encoding the (A) monomer, (B)dimer, (C) tetramer and (D) octamer of the protein TNfn3. The agarose gel contains ethidium bromide (EB) at a concentration of $1.5 \times 10^{-5}$ mg/mL. The image is taken by AlphaImager from Alpha Innotech (San Leandro, CA) under 365 nm UV light. The vector cloning monomer, dimer and tetramer is pUC19. The vector cloning octamer is pQE80L. The 2-log DNA ladder from New England Biolabs is used as the DNA size indicator and the DNA size of those critical marker band are labelled. For the DNA electrophoresis, the gene corresponding to monomer, dimer, tetramer or octamer of TNfn3 is released from the plasmid by restriction digestion ($\text{BamHI}$ and $\text{KpnI}$ restriction enzymes from New England Biolabs). The arrows indicate the correct size of different genes. In some cases, besides the released insert (highlighted by arrows), there are several bands above the insert band, which are resulted from the incomplete enzymatic digestion. The highest band should be the lineated plasmid (vector+insert) which is only digested by one of the two restriction enzymes. The second band should be the empty vector with the insert completely released by digestion. The lowest one should be the undigested plasmid which is circular and forming the super coiled conformation.

A3. Expression of the protein

We express the protein in the E. coli expression system and then purify the protein from the bacteria. First, the DH5α bacteria cells containing target polyprotein gene are grown in 3 mL 2.5% Luria-Bertani broth (LB) containing 10 mg/L ampicillin over night at 37 °C and 225 rpm. The seeds are from the pre-prepared bacteria stock. The harvested cell culture after the overnight incubation culture is again used as seed and inoculated into 300 L of LB medium (1/100 dilution) containing 10 mg/L ampicillin. The
new inoculation fraction is grown at 37 °C and 225 rpm while its optical density (OD) is kept tracking. When the OD of the cell culture falls into the region between 0.6 and 1 (log phase), which usually takes about 3 hours, the expression of the target polyprotein will be induced. Isopropyl-1-β-D-thiogalactoside (IPTG) is added into the cell culture with a final concentration of 1 mM. The induced cell culture is then further incubated at 37 °C and 225 rpm for ~ 3.5 hours. After the expression finished, the cells are harvested by centrifugation at 15,000g for 15min. The supernatant is discarded and the cells are lysised by adding lysozyme (final concentration 1mg/mL) and incubating on ice for 30 min. The DNAs and RNAs which are co-released from the cells are digested by Dnase I and Rnase A (final concentration 0.005mg/mL). The soluble fraction is passed through an affinity chromatographic column packed with Co²⁺ resin. The column bound with the polyprotein is washed by 200 mL washing buffer (10mM phosphate buffer with 300mM NaCl and 7mM imidazole). The polyprotein is then eluted with elution buffer (10 mM phosphate buffer saline (PBS) with 250mM imidazole). The yield of polyproteins varies for different proteins and can fluctuate from 10mg to 80mg per liter culture. By estimating from the SDS-PAGE using AlphaEaseFC software (Version 4.0.0, Alpha Innotech, San Leandro, CA 94577), the purity of the yielded proteins is > 90%. A typical SDS-PAGE gel has been shown in the figure 3.9.
Appendix B: The mechanical unfolding of T4-lysozyme using single molecule AFM

B1. Multiple unfolding pathways of T4-lysozyme are not originated from the heterogeneity due to different attachment sites of the cantilever.

Since T4-lysozyme is flanked by (GB1)$_4$ on both ends, if five or more unfolding events of GB1 are observed in a given force-extension curve, we are certain that the unfolding event prior to the unfolding events of GB1 domains must correspond to the stretching and unraveling of T4-lysozyme. Therefore, in these molecules, the cantilever must have been attached to one of the GB1 domains (or purification tags) for such force-extension recordings containing five or more GB1 unfolding events. Because the cantilever is not directly attached to T4-lysozyme, the possibility that T4-lysozyme is being stretched with different pulling geometries/directions can be safely excluded. Another possibility is the heterogeneity in the pulling angles for different molecules. As described by Carrion-Vazquez et al. (138), it is possible that some polyprotein molecules are not stretched by the cantilever along the direction perpendicular to the surface of the solid substrate, but at a particular angle (Fig. B1). However, the result of pulling angles is that only a component of the stretching force $F \sin \theta$ is applied to effectively unfold the protein. Pulling angles will only affect the measured unfolding forces (leading to a slightly broader distribution of unfolding forces) and contour length increment, but will not affect the native state of the protein in any way. Therefore, pulling angle cannot explain our results.
Figure B1. Pulling angle may affect the measured unfolding force and contour length increment of T4-lysozyme, but does not affect the native conformation of T4-lysozyme in any way.

**B2. Two-state unfolding of T4-lysozyme may involve unfolding intermediate state that unfold at forces below 20 pN.**

It is possible that the two-state unfolding routes for T4-lysozyme that do not show any detectable unfolding intermediates may involve intermediates that unfold at forces below 20 pN. However, this scenario does not invalidate our conclusion of parallel unfolding pathways in any way. On the contrary, this scenario presents additional support for the conclusion of parallel unfolding pathways: it is evident that the unfolding forces for the detectable intermediate states center on ~50 pN, which is significantly higher than 20 pN. If there is another population of the intermediate states that unfold below 20 pN, the unfolding force histogram would have a bimodal distribution, which would indicate two populations of intermediates with distinct mechanical stability, leading to distinct conformations.
B3. Classification of unfolding events of T4-lysozyme.

The unfolding events of T4-lysozymes were identified and classified according to the following procedures and criteria: the force-extension curves were smoothed using standard median filter procedure in Igor Pro software (Wavemetrics) and the peaks were detected based on a threshold force of 20 pN. After confirming that the force-extension curve contains at least five unfolding events of GB1, we can ensure that the low unfolding force event preceding the unfolding of GB1 corresponds to the unfolding of T4-lysozyme. Additional constrain is the contour length increment of the unfolding event. For example, in rare cases, we observed putative unfolding events of T4-lysozyme with $\Delta Lc$ of more than 90 nm, which is unreasonably larger than the contour length of T4-lysozyme. These recordings were not included in our data analysis. We then classify the unfolding traces according to the number of peaks which show up for the unfolding event of T4-lysozyme: if the force-extension curve of T4-lysozyme shows only one unfolding peak, this event would be classified as a two-state unfolding event; if the force-extension curve of T4-lysozyme shows two unfolding force peaks and the sum of $\Delta Lc1$ and $\Delta Lc2$ is $\sim$60 nm, this event would be classified as a three-state unfolding event (Fig. 2.2).