Title
An investigation of the mechanical properties of the molten globule state of apomyoglobin

Author
Elms, Phillip James

Publication Date
2010

Peer reviewed|Thesis/dissertation

An investigation of the mechanical properties of the molten globule state of apomyoglobin

by

Phillip James Elms

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Susan Marqusee, Chair

Professor Carlos Bustamante

Professor Dan Fletcher

Professor Harvey Blanch

Fall 2010

Abstract

An investigation of the mechanical properties of the molten globule state of apomyoglobin

By Phillip James Elms

Doctorate of Philosophy in Biophysics

University of California, Berkeley

Professor Susan Marqusee, Chair

Single molecule force spectroscopy has provided important insights into the properties and mechanisms of protein folding. However, there are still many unanswered questions about how force affects the folding and unfolding of proteins and, in particular, the relationship between force and the rate-limiting transition state.

In this thesis, I developed two protein systems to address two specific questions. The first question arose form previous work on *E. coli* RNAse H, in which a molten globule-like intermediate was observed to have a large distance (5 ± 1 nm) to the transition state. This large distance was in sharp contrast to the smaller distances (< 2 nm) typically observed for natively folded proteins. This raised the question of whether this distance was a general property of the *E. coli* RNAse H intermediate or a more general property of a molten globule state. To this end, I investigated the equilibrium molten globule state of sperm whale apomyoglobin at pH 5 under force and demonstrated that this state had a large distance to the transition state of 6.1 ± 0.5 nm. Further, this state was shown to have a large distance to the transition state regardless of the axis of the applied force. This work suggests that a large distance to the transition state is a general property of the molten globule state.

The second system was developed using the SH3 domain from chicken c-Src in order to investigate if and how the structure of the transition state changes under force. I investigated the behavior of the protein under two different force axes observing significant differences in the mechanical unfolding of the protein. These experiments are ongoing but indicate that the change in behavior is because of a change in the structure of the transition state under force.

Finally, investigating the properties of the molten globule state revealed an error in previous methodology using constant force feedback experiments. In this thesis, I identify and explain the origin of this error. Further, work on the molten globule state required higher fidelity data and a more sophisticated approach for the analysis of the data. Working with John Chodera and colleagues, we implemented novel methods for the analysis of the data.

This thesis is dedicated to my parents, Joe and Suzanne Elms

Table of Contents

List of Figures and Tables

Acknowledgements

I have been very fortunate in my graduate career to fall into a project that involved working in two labs and being part of the Biophysics program at Berkeley. The great people from both labs and from my graduate program are too numerous to thank and everyone contributed to creating the great environment that is Berkeley.

I am deeply indebted to my advisor Susan Marqusee for her guidance and the great impact she has had on my growth as a scientist. Her support and patience has been invaluable. I was also lucky to have a second advisor in Carlos Bustamante, whose guidance and lab provided a great compliment to the Marqusee lab. Carlos is always striving to answer the big questions and his enthusiasm is infectious.

I have been blessed to have such great labmates and friends through graduate school and I am indebted to all the members, past and present, for creating a great environment. I joined the Marqusee lab in part because of the high quality of the older students in lab, in particular Beth Shank, Erik Miller, and Pete Wildes. Beth was an excellent mentor and has essential to helping me make the transition to my new lab and project. In the Marqusee lab, I would like to thank Katelyn Connell, Katie Hart, Geoff Horner, Jesse Dill, and Rachel Bernstein for their help and patience with my incessant questions. And I would like to give a special thanks to Katie Tripp with whom I had many passionate discussions. In the Bustamante Lab, I am particularly thankful for having worked with Jeff Moffitt, Yara Meija, Craig Hetherington, Lacramioara Bintu and Courtney Hodges. They provided me with guidance and helpful conversations whenever I found my self stuck on a problem. In the Biophysics program, I would like to thank Hari Shroff, Merek Sui, Derek Greenfield, Dave Richmond, Ailey Crow, and Dave Sivak for their advice throughout graduate school.

I would also like to give a special thanks to my collaborator John Chodera. He has been a great pleasure to work with and without his contributions my graduation would have been much delayed and my thesis would have been of lower quality.

My greatest discovery in graduate school has been Katherine Miller. She has challenged me, making me a better person, and supported me through my toughest challenges. She is a constant reminder of what is important in my life.

Lastly, I would not be here if it had not been for the continuing love and support of my parents, Joe and Suzanne Elms. I could not have asked for better parents.

Chapter 1 Introduction

1.1 Overview of thesis

In order to sustain life, many complicated process have to occur. To accomplish this, evolution has harnessed a variety of physical forces. Voltage is used to facilitate the communication between nerve cells, and chemical gradients generate the energy needed to synthesize ATP, the fuel of life. The movement of the largest animals that roam the world is a result of forces generated by single molecules working in concert. To understand how any one of these processes is accomplished, we have to study the driving force behind them. Our understanding of these processes has continued to progress with technological advances, which provide us with better ways to measure and manipulate the driving forces behind them. Due to advances in force spectroscopy, we can now probe the role of mechanical force at the level of single molecules.

During my graduate career, my research has addressed fundamental questions regarding the role of force in biology, focusing on how force affects the folding and unfolding behavior of proteins using the optical tweezers. In my introduction, I will review several topics of importance to my thesis. First, I will address the role force plays in biology and discuss force as a thermodynamic variable. Next, I will describe single molecule experiments and the instrument used in this thesis, followed by a summary of some of the issues to be considered when analyzing data from force spectroscopy experiments. Finally, I will briefly summarize some outstanding questions in the protein-folding field that I have addressed in my thesis work.

During my graduate work, I developed methods and evaluated the behavior of two protein systems under mechanical force. My thesis will therefore consist of three general sections. The first will address methodology and discuss my discovery of a serious flaw in previous constant force experiments and the analysis of the data. I will present the work I have done to identify and explain the origin of this error as well as explain a better approach to the methodology and analysis (Chapter 2).

The second section will include the work I have done on two different protein systems: sperm whale apomyopglobin and chicken srcSH3 (Chapter 3 and Chapter 4, respectively). Apomyoglobin was used to evaluate the mechanical properties of the molten globule state. The molten globule state represents a class of partially folded states that are thought to represent ubiquitous intermediates during folding to the native state of many proteins. My results suggest that the molten globule state unlike natively folded proteins has a large distance to the transition state. Also, as opposed to natively folded proteins, my work suggests that the molten globule state is more isotropic - a large mechanical compliance is observed regardless of the direction of the applied force.

Using the srcSH3 protein, I developed a model two-state system that has enabled an investigation of the effect that pulling axis has on the mechanical properties of the protein. I pulled on the protein along two different axes, observing large differences in the unfolding forces. Using the Crooks Fluctuation Theorem, I determined that the free

energy on the protein does not change with the different pulling axes that might result from the handle attachment, indicating that the different unfolding behavior must be due to differences in the transition state. Future work using this system will characterize the structure of these transition states using a mutational analysis (phi-value analysis). This will focus on a comparison between the different mechanisms of unfolding under force along different axes and in solution in the absence of force.

In the last section (Chapter 5 and 6), I will discuss some of the work I have done in collaboration with John Chodera and colleagues, who have developed two novel approaches for identifying states and their lifetimes from single molecule experiments. These methods provided a more robust analysis of the data and demonstrated the usefulness of these methods on experimental data.

1.2 The role of force in biology

Generating or resisting force plays a crucial role in many biological processes. Many single protein molecules have evolved to work together to generate force and resist force on a macroscopic scale. In fact, the most abundant protein in animals, collagen, which accounts for approximately a quarter of the total protein [1], is an important component in many tissues that need to resist deformation such as tendons and ligaments. Forces generated in muscles by the proteins myosin and actin work in concert to produce movement [2]. Other proteins play equally important roles such as titin which maintains sarcomeric structural integrity and acts as an entropic spring generating a passive force [3].

Force also plays an important role in many other cellular processes. On the scale of a single cell, actin networks generate forces allowing cellular mobility [4], and environmental forces can play a crucial role in determining the development of stem cells [5]. Within a cell, proteins such as kinesin generate forces to actively transport various cell cargo to different regions of the cell [6]. Evidence also suggests that proteins use force to facilitate transport across membranes, as is the case with mitochondrial import [7, 8]. Further, force is also thought to play a crucial role in protein degradation in the case of the protease ClpXP [9]. In these last two examples, force is thought to first unfold a target protein before import across the membrane or into a catalytic cavity for degradation.

1.2.1 Force as a thermodynamic variable

In order to gain an understanding of how force affects a biological reaction I will first discuss the role of force in the thermodynamics of a system. The first law of thermodynamics states that energy is extensive and therefore additive and conserved. Therefore any change in energy must result from either a flow of heat into or out of the system (dQ) or work done on or by the system (dW).

$$dE = dQ + dW = TdS + F \bullet dX \tag{1.1}$$

The work done on or by the system is defined by a driving force, F, multiplied by the extensive conjugate variable, represented by dX. The F·dX term represents any two variables whose produce has units of energy, such as a force and distance, pressure and volume, voltage and charge, chemical potential and the number of molecules. The work done on a system could be from a variety of forces all acting on the system at the same time in which case the notation would be $F_1 \cdot dX_1 + F_2 \cdot dX_2 + \cdots$.

To illustrate how a driving force changes the energy in a system, take the example of a two-state reaction at equilibrium under constant temperature. The equilibrium constant is a function of the free energy difference between the two states (denoted A and B) and is defined by the ratio of the populations in each state or equivalently by the ratio of the forward and reverse rate constants.

$$\Delta G_0 = -RT\ln K_{A\text{-}B} \tag{1.2}$$

$$K_{A\text{-}B} = [A]/[B] = k_{B\text{-}A}/k_{A\text{-}B} \tag{1.3}$$

Taking this as our reference state, we can then do work on the system with a driving force.

$$\Delta G(F) = \Delta G_0(F=0) + F\Delta X(F) \tag{1.4}$$

$$-RT \ln K_{eq}(F) = -RT \ln K_{eq}(F=0) - F\Delta X(F) \tag{1.5}$$

After the system has equilibrated, we can then measure the new equilibrium constant and determine the free energy difference between the two states.

The relationship between the free energy and the driving force need not be linear. Experimentally, this linear relationship between the free energy and force holds over small force ranges [10].

1.2.2 A review of rate theory

Before discussing the effect of force on the rate of a reaction, a review of rate theory will be helpful. Many chemical reactions can be described by the empirical relationship known as the Arrhenius equation:

$$k = Ae^{\frac{-E_a}{RT}} \tag{1.6}$$

While first developed by van't Hoff in 1884 [11], Arrhenius provided a physical interpretation of the relationship in 1889 [12]. Their insight was that the rate constant of a reaction is a product of a pre-exponential term multiplied by a Boltzmann weighted activation energy. In 1910, Marcelin and Kohnstamm, Acheffer and Bransnam separately determined that the activation energy can be interpreted as the free energy of the transition state [13, 14].

A physical interpretation of the pre-exponential term progressed with the development of the collision theory of reactions in the gas phase developed by Trautz in

1916 [15] and Lewis in 1918 [16] , which was based on the kinetic theory of gases. In this theory, reactions in the gas phase are explained by calculating the pre-exponential as a product of the frequency of the collisions and a steric or orientation factor, which accounts for reactions that require an anisotropic collision. This introduced the idea that the pre-exponential can be thought of as a frequency factor or an attempt rate.

Further progress was made in the 1930's with the development of transition state theory [17]. This theory states that the reaction rate can be described by the potential energy landscape of the reactants and the products. In order for a reaction to proceed, a high-energy activation state, or transition state, must form at the saddle point between the reactants and products. The pre-exponential is the fastest the rate could occur if there was no energy barrier between the reactants and product and therefore can be thought of as the speed limit of the reaction.

Originally, transition state theory was developed to describe simple chemical reactions such as the formation of H_2 ($H + H \rightarrow H_2$). As the hydrogen atoms are spherical and symmetric, the potential energy landscape of this simple system is defined by a single coordinate, the distance between the two atoms. In this case the reaction can be described by the Eyring equation [18]:

$$k = \frac{k_B T}{\hbar} e^{\frac{\Delta G^{\ddagger}}{k_B T}} \tag{1.7}$$

For more complex molecules and reactions, the potential energy landscape becomes more complicated as other coordinates become important, such as the relative orientation of reactants to one another.

In a seminal paper in 1940, Kramers presented a general treatment of more complicated reactions in a condensed phase [19]. In this theory, Kramers models the pre-exponential as a diffusion-limited process over a continuous coarse-grained potential energy landscape. As the process is in a condensed phase, the viscosity plays an important role in the diffusion along this potential energy landscape. The pre-exponential can then be thought of as the diffusion limit of the process or, again, as a speed limit to the reaction. In Kramers' treatment, the pre-exponential is a function of the average curvature of the potential well and the viscosity of the solvent. This theory has typically been used for interpreting reactions involving biomolecules such as protein folding.

In summary, an Arrhenius relationship is observed for many chemical reactions and two parameters can be extracted from the data, a pre-exponential and a Boltzmann weighted energy. The various theories introduced above provide a physical interpretation of the meaning of these two parameters. The next section will discuss the interpretation of rate constants as a function of force given that we can know monitor the reaction along a defined reaction coordinate, the end-to end extension of the molecule, in single molecule experiments.

1.2.3 The effect of force on reaction rate constants

Often, an energy landscape is illustrated in a single dimension, collapsing a highly multidimensional potential energy landscape onto a single reaction coordinate. In the case of a single molecule mechanical experiment, we are able to monitor a reaction of a single molecule along a single coordinate, the end-to-end extension of the system, and study the changes in rates as a function of force along this reaction coordinate. However, since one is restricted to observations along a single coordinate, this can result in neglecting important orthogonal coordinates that are important to the reaction. Therefore, it is necessary to test experimentally whether this simplified model explains all the observed behavior and thereby determine if the end-to-end extension is a good a reaction coordinate for the system.

Defining the reaction along a single coordinate, end-to-end extension, the simplest model of how a given applied force will affect the rate is a linear free energy relationship, such as the Bell relationship [20, 21],

$$k(F) = k_m k_0 \exp\left(\frac{Fx^{\ddagger}}{k_B T}\right) \tag{1.8}$$

where k_m represents the contribution of experimental parameters such as the bead size, trap stiffness, and other parts of the experimental system to the observed rates, k_0 is the intrinsic rate constant of the molecule in the absence of force, F is the force, $x^{\ddagger}$ is the distance to the transition state, k_B is the Boltzmann constant, and T is the temperature in Kelvin. This relation can be rewritten,

$$\ln k(F) = \ln k_m + \ln k_0 + \frac{Fx^{\ddagger}}{k_b T} \tag{1.9}$$

Alternatively, a constant potential can be applied to a system, in which case the above relation is modified to,

$$k(F) = k_m k_0 \exp\left(\frac{Fx^{\ddagger} + \frac{1}{2}\kappa x^2}{k_B T}\right) \tag{1.10}$$

where κ is the spring constant of the system [21]. Geometrically, this algebraic relationship can be thought of as tilting the energy landscape around some reference position (Figure 1.1) [10].

A positive spring constant results in an increased effective barrier height, increasing the average lifetime of the state. As depicted in Figure 1.1 b and c, this increased barrier height results for both positive and negative changes in the extension of the system (i.e. unfolding and folding events). Therefore, the average lifetime of a state

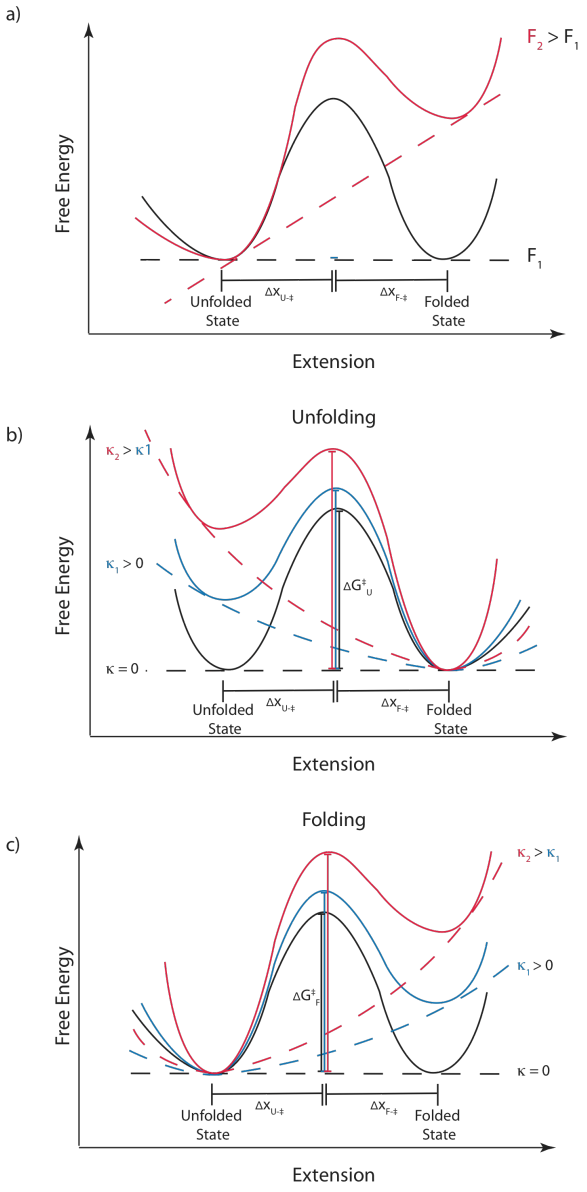

Figure 1.1 The affect of force on a potential energy landscape.

In **a**, the difference in the energy of two states is shown at two arbitrary forces with an effective spring constant of zero and with F_2 greater than F_1. The higher force changes the energetics such that the more extended unfolded state is the lower energy state. In **b** and **c**, the potential energy surface is shown for a positive and negative extension change (unfolding and folding, respectively) at various spring constants. The folded state in **b** and the unfolded state in **c** are depicted at the same average force. This illustrates the change in the effective barrier heights with a change in the effective spring constants of the system.

measured at an average force is dependent on the effective spring constant of the system. A greater spring constant will result in a longer average lifetime of the state.

Experimentally, we can measure the average lifetime ($\bar{\tau} = 1/k$) for a given state. Empirically, we can measure how the lifetimes change as a function of force and determine the distance to the transition state. For the systems studied in this thesis, the natural log of the rate constant varies linearly over small force ranges. A change in the effective spring constant, while affecting the lifetimes of the state, does not affect the determination of the distance to the transition state. Geometrically, this indicates that the curvature of the potential energy surface is sharp and such that the distances do not change over the force range studies.

1.2.4 Distinguishing between equilibrium and kinetic behavior

With any experiment, which causes a change in the energetics of a system, it is important to establish if the system is at equilibrium or in a kinetic regime, which is ultimately a question of timing. A system is at equilibrium if there is no hysteresis, meaning the macroscopic behavior of the system does not change with time. In other words, the distribution of the population molecules among different energetic states or the average time spent by a single molecule in the different states is a reflection of the Boltzmann distribution. Given a perturbation, such as a change in force, a system is driven out of equilibrium; the time it takes to re-equilibrate is defined by the relaxation time of the system. Characterizing the equilibrium or the kinetic behavior of a system is therefore determined by the timing of the experiment relative to the relaxation time of the system. Differentiating between equilibrium and kinetic behavior will be discussed later when discussing the particular types of force-control mode experiments.

1.2.5 Questions concerning force and transition states

The modified Bell's model provides a simple description of how force affects a single transition barrier along a single dimension defined by the reaction coordinate. However, the potential energy landscape of a biological molecule is multiple dimensional with potentially important orthogonal coordinates which are not accounted for in this model. In addition, as a system is perturbed, the rate limiting step or transition barrier could change as a function of force. Further, within the protein folding field in particular, there is significant debate over whether folding kinetics is accurately described by a single transition state or whether a protein folds through multiple parallel pathways [22].

Force spectroscopy can provide insight into resolving the nature of the transition state and how it changes as a function of force. This thesis discusses the transition state under force for protein folding and unfolding using two different two-state model systems, sperm whale apomyopglobin and chicken srcSH3. First, the transition state of the molten globule state is studied using apomyoglobin. Second, src SH3 is used to study the effect of the pulling axis on the transition state of a natively folded protein. Both systems have been extensively characterized in solution in the absence of force, enabling the comparison to the folding and unfolding behavior under force.

1.3 Single molecule experiments

Up until relatively recently, most experimental insights into chemistry and biology have been based on observations made on ensembles of molecules. An inherent characteristic of these experiments is the determination of an average property of the ensemble. Consequently, information can be lost about the distribution of the molecules or the diversity of behavior within the ensemble. And in some cases, the average of a measured value can be misleading. Often times in biology it is the distribution or fluctuations from the average that are of functional importance. Thankfully, we now have techniques that allow us to detect and measure the properties of individual molecules [23]. With these advances, we can now address questions about the properties and behavior of molecules previously inaccessible.

To illustrate the importance of single molecule experiments, consider the analogy of an ensemble of six-sided dice, with each die representing a molecule. The average value of the ensemble of dice is 3.5, but no individual die ever has a value of 3.5. The average property of the ensemble is a consequence of the behavior of the individual die; however, it does not reveal the range of behavior of the die. In order to fully understand the system, the dice could be inspected one at a time, measuring the value of each die. Alternatively, a single die could be rolled many times and the value of each roll could be measured. With either approach we still measure an average value of 3.5, but we gain additional information about the system such as the values are limited to integers between 1 and 6 and that each integer is equally probable.

The case of the dice illustrates two important concepts, the equipartition theorem and the ergodic hypothesis of thermodynamics. For a system at equilibrium, the statistical average of an ensemble is the same as the average of a molecule over time. For a die, each value represents a different microstate the die can populate. Each state is equally probable and, therefore, of equal energy. For an ensemble of dice, 1 out of 6 dice will have the value of 1, etc and the statistical average of this system is:

$$(1/6)\bullet 1+(1/6)\bullet 2+(1/6)\bullet 3+(1/6)\bullet 4+(1/6)\bullet 5+(1/6)\bullet 6 = 3.5$$

For the experiment using a single die, we determine the average value by observing the behavior of the die in time, measuring many independent rolls. In order to ensure that each measurement is independent, the time between each measurement should be longer than the relaxation time or correlation time of the system. In the case of the die, the correlation time of the system is the time it takes to re-roll the dice.

In addition to learning about the distribution and dynamics of the microstates of an ensemble, single molecule experiments can provide information about other microscopic properties of the system. Spatial information can be obtained providing precise localization of molecules or distance measurements. Experiments can exploit the high resolution to follow a reaction in time along a specified coordinate. The coordinate probes a section of the potential energy surface, or energy landscape, over which the reaction proceeds. Further, intermediates or transiently populated states can be observed, providing new insights into the mechanism of the reaction.

1.3.1 Single molecule force spectroscopy

Using an instrument such as an optical trap, an atomic force microscope (AFM), or a magnetic tweezer, the force and extension of a molecule can be measured [24]. In this method, the molecule is tethered between the probe and another surface. By measuring the force on the probe and the position of the probe, the conformation of the molecule or the progress of a reaction can be monitored in time. Applying force perturbs the energy of the system, influencing if and to what extent a reaction occur [10] (see section 1.1.3). For the purposes of this thesis, I will focus in this section on how force is applied and measured using an optical trap.

1.3.2 Optical tweezers

Optical traps, or tweezers, were first developed by Arthur Ashkin [25-27] when he demonstrated that light can apply a small force to an object and can be used to manipulate micron-sized or smaller objects, such as a polystyrene bead in water. Momentum from the light can be transferred to an object that is larger than the wavelength of the light by absorption, diffraction, refraction, or reflection. In the case of the polystyrene bead, light refracts as it passes through the bead because of the difference in the refractive index of the bead and the water. The momentum of the light changes resulting in an equal and opposite force according to Newton's third law. This phenomena is best explained with ray optics (Figure 1.2). In a Gaussian-distributed beam, more photons are in the center of the trap. If the bead is displaced from the center, there are more photons that refract creating a restoring force, which pushes the bead back towards the center. Light that reflects off the surface of the bead and imparts forward momentum, pushing the bead slightly out of the focus in the direction of the beam (Figure 1.3 a). In a given regime, the optical trap acts as a Hookean spring and the force is linearly proportional to the displacement from the center of the trap,

$$F = \kappa \Delta x \tag{1.11}$$

where F is the force in pN, κ is the spring constant in pN/nm, and the Δx is the displacement from the center of the trap in nm.

To create a laser trap that can measure force and extension, three general components are needed: a laser, optics, and a detector. The laser produces the light used to make the trap, the intensity of which determines the number of photons, and therefore the strength of the trap. The optical components focus the laser and allow for the manipulation and movement of the trap. A position-sensitive detector measures the change in the trapping light by measuring the voltage that corresponds to the position and intensity of the light thereby measuring the force on the bead.

There are two general instrument designs: a single beam optical trap and a dual beam anti-propagating optical trap (Figure 1.3). Both have advantages and disadvantages and allow for different approaches to measuring the force.

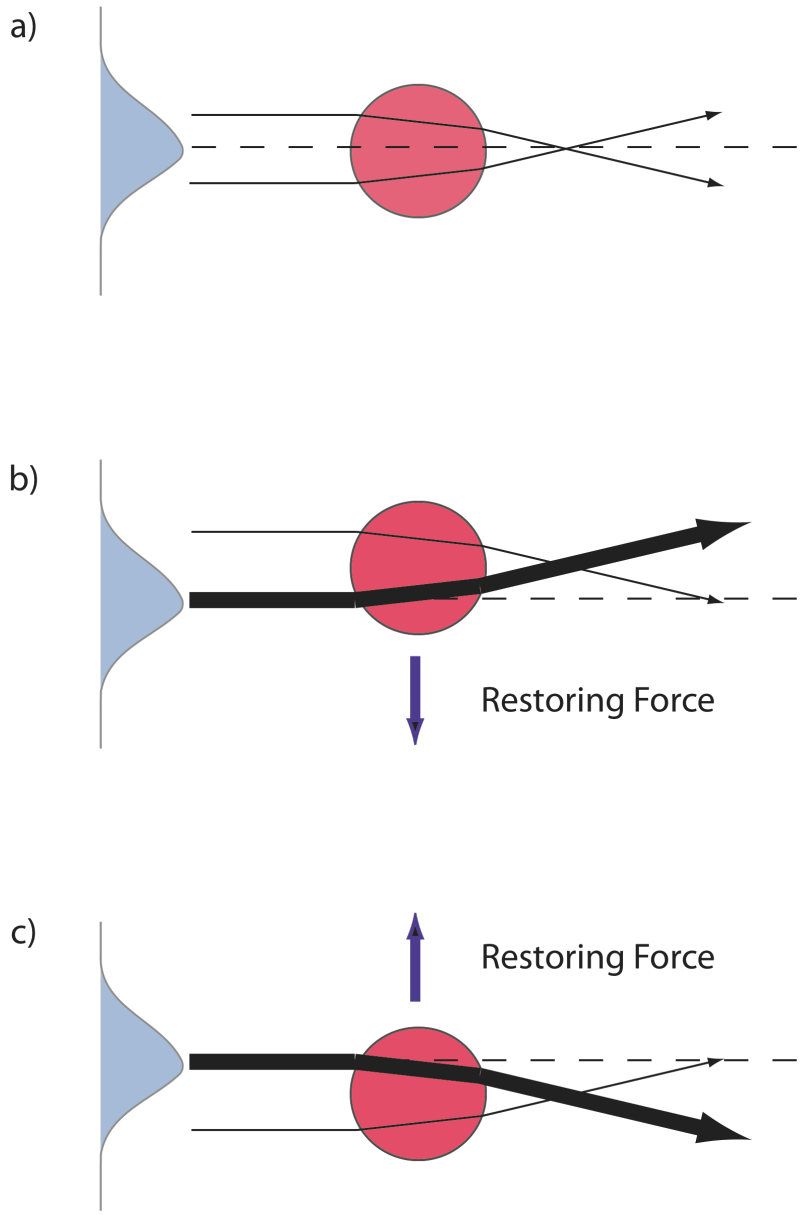

Figure 1.2 Trapping a bead with an optical trap.

A Gaussian-distributed light beam is depicted in blue with the center of the trap marked by the dashed line. The black arrows depict the path of the refracted light and the thickness of the line is proportional to the number of photons. In **a**, the bead is in a position such that the force created from the refracting light produces a net zero force on the bead. In **b**, the bead has been displaced up and the light refracts to create a restoring force in the opposite direction, shown with the blue arrow. In **c**, the bead has been displaced down and the light refracts to create a restoring force in the opposite direction.

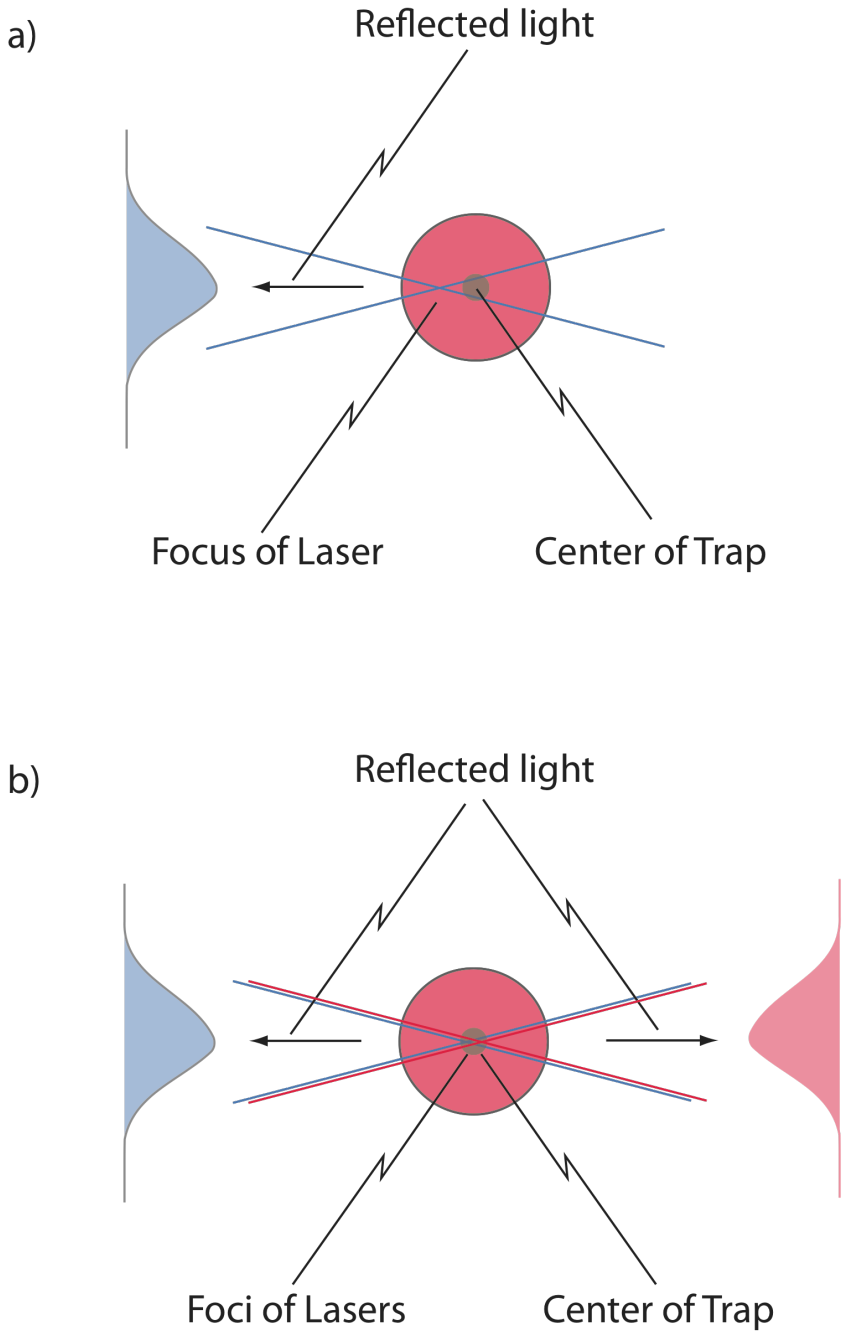

Figure 1.3 Single beam and dual beam anti-propagating optical traps.

In **a**, the center of the trap and the focus of the trapping light are offset in a single beam optical trap because of the reflected light. In a dual beam anti-propagating optical trap, if both traps are of equal strength, the reflected light produces an equal and opposite force. The center of the trap and the foci of the two are therefore aligned as shown in **b**.

1.3.3 Measuring force

There are two general approaches to measuring the force on a trapped object, each with its set of advantages and disadvantages. It should be noted that both methods require the calibration of the detector to determine the voltage conversion factor, either voltage-to-distance or voltage-to-force. The first method directly measures the total change in momentum of the light [28]. This requires the collection of nearly all of the photons used in the trap. To ensure this, the objectives are under filled with light to prevent lost and scattered photons, resulting in a lower power and weaker trap. However, using a dual anti-propagating geometry as described above can compensate for the weaken trap. This method has the advantage that the spring constant of the trap (which requires prior knowledge about the size of the trapped object and the viscosity of the solution) does not have to be determined. Without the spring constant, the position of the object in the trap is unknown. Other methods will be discussed later for extracting relative extension changes.

The second method requires the determination of the spring constant of the trap for each object. A common calibration strategy is to take a power spectrum of the free bead in the trap [24]. A power spectrum measures the noise as a function of the sampling frequency, which is a product of the thermal fluctuations or Brownian noise in the system. Using the equipartion theorem, the thermal noise in the system is related to the the average energy of our trap,

$$^1/_2 k_B T = {}^1/_2 \kappa \Delta x^2 \tag{1.12}$$

where k_B is the Boltzmann constant, T is the temperature in Kelvin, κ is the spring constant in pN/nm, and Δx is the displacement from the center of the trap in nm. From this relationship, the spatial (δx) and force (δf) resolution can be determined.

$$\delta x = \sqrt{k_B T/\kappa} \tag{1.13}$$

and

$$\delta f = \sqrt{\kappa k_B T} \tag{1.14}$$

Fitting the power spectrum, the corner frequency can be determined and the spring constant of the trap can be determined with the following equation,

$$f_c = \kappa/12\pi^2 \nu r \tag{1.15}$$

where f_c is the corner frequency in Hz, κ is the spring constant in pN/nm, ν is the viscosity of the solution in pN•sec/nm^2, and r is the radius of the trapped bead. This requires that the size of the object and the viscosity to be independently determined. With this method, the objectives are over filled and therefore more light is used to create a stronger trap, however, it is still weaker relative to a dual beam anti-propagating trap with similar laser intensities. Given the spring constant, the position of the bead in the trap is also known because of Hooke's law and therefore the extension of the molecule.

1.3.4 Instrument overview

Specifically, the instrument used for the work presented in this thesis was a dual beam anti-propagating optical trap (Figure 1.4) [29]. For this instrument, two independent traps of equal stiffness are created that approach the object from opposite directions. One advantage of this design is that a stronger trap is created with lower laser power. For a trapped object in a single trap, the trap is weakest along the direction of the propagating light. The trap is further weakened by light that is reflected off the object, pushing it out of the focal point of the trap. In a dual beam counter-propagating trap, the reflected light from each trap compensate for each other, placing the object in the focal point of the traps for maximum strength. This design also enables the force to be measured directly by measuring the rate of change in the momentum of the light, as described above.

1.3.5 Geometry of an experiment and sample preparation

In order to study a biological molecule with force spectroscopy, it must be attached to two different surfaces, one of which is the probe, either a bead in an optical trap or the cantilever tip of an AFM. By changing the relative position of the two surfaces, forces can be applied to the molecule. The behavior of a biological molecule is inferred by measuring the force and extension on the probe.

For the experiments discussed in this thesis, the sample was attached to two beads (Figure 1.5). One bead was held by suction on pipette tip and the other bead was in the optical trap. As the biological molecule is small relative to the attachment surface, a 'handle' consisting of functionalized dsDNA is used to tether the molecule to the beads. Handles provide separation between the surfaces to avoid non-specific interactions and provide specific attachment points to the molecule of interest.

In the experiments presented here, the beads are functionalized with covalently attached strepavadin or anti-digoxigen antibody. A PCR reaction using functionalized primers produces a biotin and thiol-labeled or a digoxigen and thiol-labeled 558 bp dsDNA handle. These handles are then attached to the molecule of interest by annealing complementary strands for a nucleic acid hairpin or through disulfide bonds for proteins [30, 31].

1.3.6 Force control modes

In force spectroscopy, there are different ways in which the force is applied to the sample. The three primary force control modes are force ramp, constant force, and constant-trap position. Below are brief summaries of each method.

In the force ramp experiment, the probe is moved relative to the other surface. Moving the surfaces away from each other pulls on the tether, resulting in an increase in the force, while moving the surfaces closer together, decreases the tension on the molecule. Take the example of a molecule that unfolds and folds in a two state manner. As the force increases, the more extended state of the molecule will become energetically

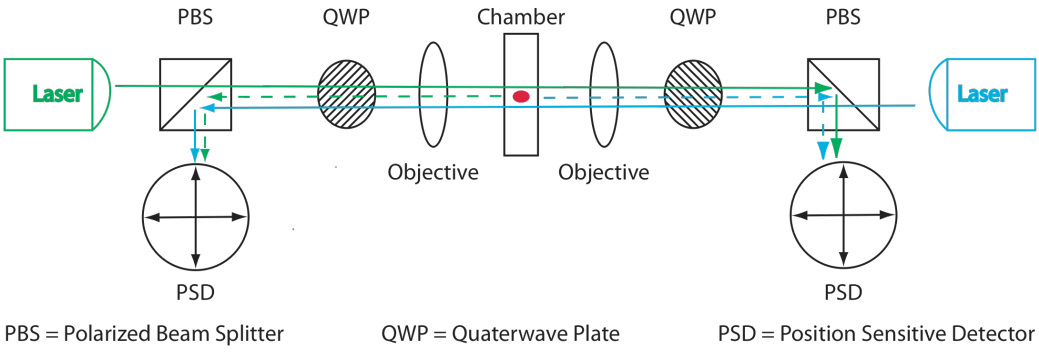

Figure 1.4 Schematic of a dual beam anti-propagating optical trap design.

This schematic depicts the primary components in a dual beam anti-propagating optical trap. The green and blue arrows depict the light path. The polarized beam splitter differential reflects and transmits linear polarized light. The quarter wave plate circularly polarizes linear polarized light. The two lasers are differentially linear polarized (horizontal and vertical) which results in the circular polarization of the light by the quarter wave plates in opposite directions (right and left). This reduces any interference between either light-beam and allows the separation of the beams to their respective position sensitive detectors. Light reflects backwards (shown in the dashed line) off the bead (red circle) in the chamber is reflected by the polarized beam splitter to the previous position sensitive detector. In this setup, most of the light (~98%) is captured allowing for an accurate determination of the force on the bead using a light momentum calculation.

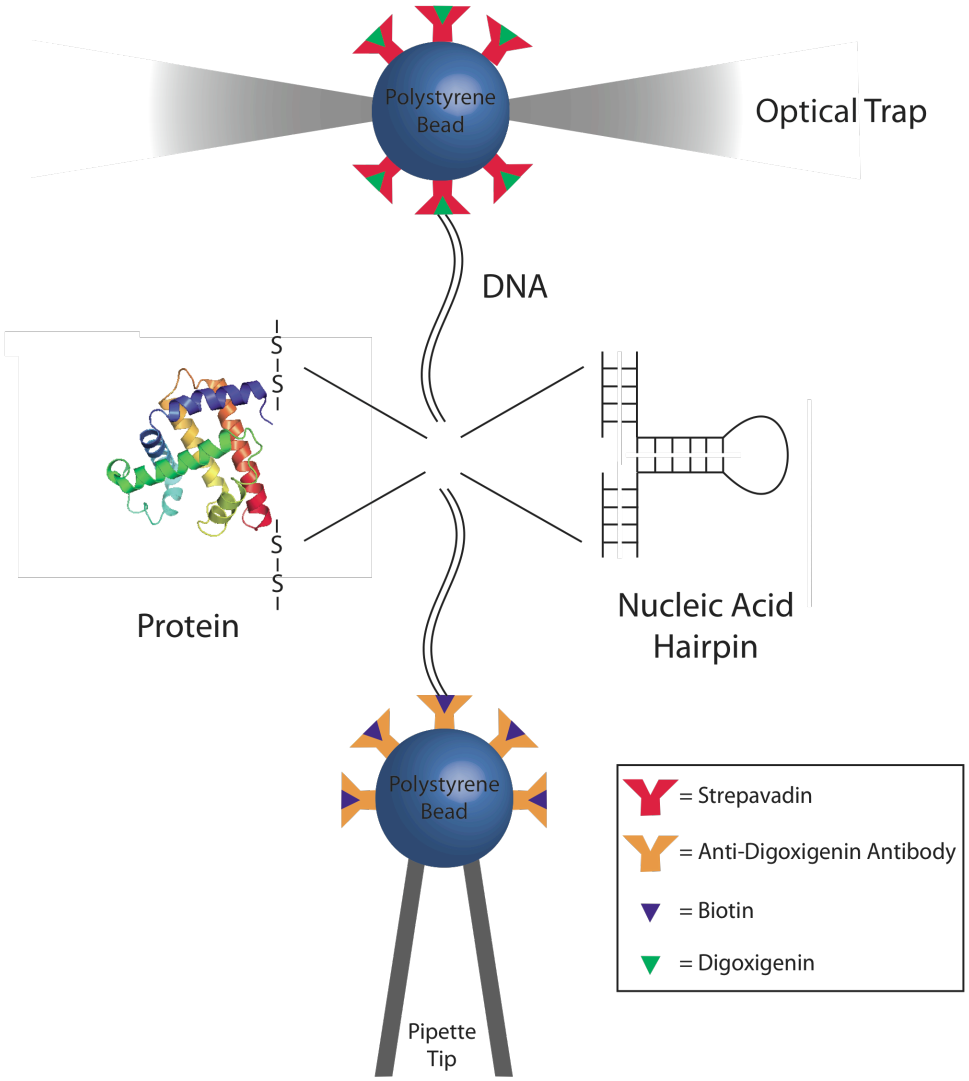

Figure 1.5 Geometry of an optical tweezer experiment.

The geometry of the optical trap experiments is depicted for both a nucleic acid hairpin and a protein system.

favorable and eventually unfold the molecule. As the force decreases, the more compact state will become the lowest energy state and the molecule will refold. Because unfolding and folding are thermally driven stochastic events, there will be a distribution of forces at which the molecule unfolds and folds, which reflect the underlying potential energy landscape of the system.By controlling the speed, the experiment can be at a constant pulling speed (i.e. constant change in trap position with time) or a constant loading rate (i.e. constant change in force with time). The distinction between these variations is important because of how the energy in the system is changing in time. As previously mentioned, the free energy of the system changes as a function of the force and affects the behavior of the molecule. In a force ramp experiment, the force is a product of the spring constant, the pulling speed, and the time.

$$F = \kappa_{eff}\delta x = \kappa_{eff}\nu t \tag{1.16}$$

where κ is the spring constant in pN/nm, ν is the pulling speed in nm/sec and t is the time in sec. Here, the spring constant is the effective spring constant of the entire system, including contributions from the trap, the handles, and the molecule of interest. This can be modeled as several springs in series:

$$1/\kappa_{eff} = 1/\kappa_{trap} + 1/\kappa_{handles} + 1/\kappa_{molecule} \tag{1.17}$$

with each spring constant a function of position. In the case of the handles and the molecule, the spring constants are not linear functions of position. This can result in nonlinear changes in the force on the system and therefore nonlinear changes in the energy of the system. If a constant pulling speed is used, the loading rate is not necessarily constant. A constant loading rate is a good approximation if the optical trap has a significantly softer spring constant relative to the other components of the system. Depending on the rate at which the system is perturbed (i.e. the loading rate of force per time) and the relaxation time of the system, either a kinetic or equilibrium behavior can be explored. The behavior regime is distinguished by the presence or absence of hysteresis. Examples of both types of behavior are shown in Figure 1.6.

If a constant force can be applied to the system, the potential on the system is constant during the experiment and the analysis is greatly simplified. There are two principal methods to sustain a constant force on a system. The first requires an active feedback that adjusts the position of the trap to maintain a constant force on the system at times greater than the timescale of the feedback. An alternative passive approach positions the bead in an optical trap where the potential of the trap is anharmonic and the force is constant over small displacements (± 5% over a range of 50 nm) [21]. With each method, the position of the trap or the bead is used to determine the state of the molecule.

There are two variations of a constant force experiment, either a force-jump experiment or a hopping experiment (Figure 1.7 a, b, and c), the application of which depends on if the system is in a kinetic or equilibrium regime during a force ramp experiment. Take the example of a molecule that unfolds in a kinetic regime. In this case, the relaxation time of the system is slow relative to the loading rate of the experiment. The system can be jumped to an intermediate force and a constant force can

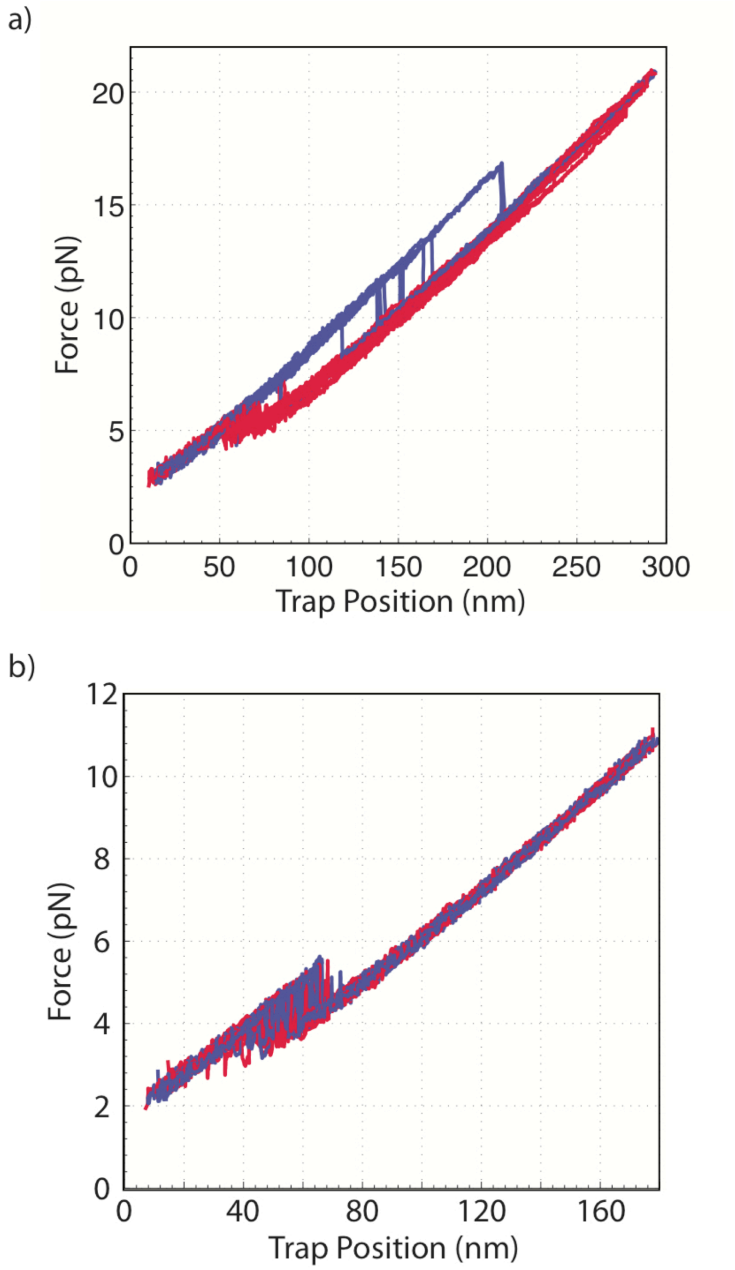

Figure 1.6 Force ramp sample traces.

Sample traces of force ramp experiments on apomyoglobin at pH 7.0 (**a**) and pH 5.0 (**b**). At pH 7, the behavior is in a kinetic regime as demonstrated by the hysteresis between the unfolding (blue) and refolding (red) events. At pH 5.0, there is no hysteresis between the unfolding and folding force distribution indicating that the molecule is in an equilibrium regime.

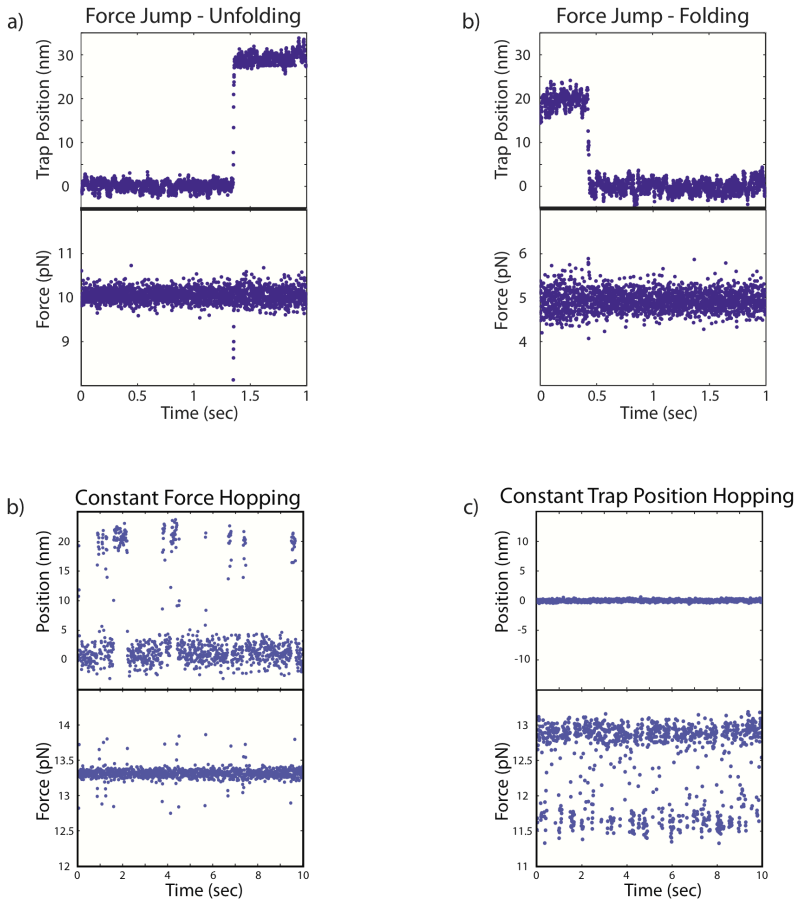

Figure 1.7 Sample traces of constant force jump experiments, constant force-feedback hopping experiments, and constant trap position experiments.

Sample traces depict the force and trap position as a function of time. The unfolding and folding of srcSH3 (**a** and **b**) is shown during a force jump experiment to illustrate the force spike (down for unfolding and up for folding) before the feedback is able to restore the original force by changing the trap position. Unfolding occurs at higher force and therefore the signal-to-noise ratio is better than for the low-force folding event. The force and trap position are also shown for equilibrium hopping experiments for a constant force-feedback experiment (**c**) and a constant trap position experiment (**d**). Data averaged down to 1000 Hz for all traces.

be maintained until the molecule unfolds or folds (Figure 1.7 a and b). From this, a lifetime can be measured at a constant force. By repeating this experiment and measuring the distribution of lifetimes, an average lifetime can be determined at the average force. As the average force increases, the average lifetimes will decrease and from this, information such as the distance to the transition state can be determined.

A molecule near equilibrium with little hysteresis in a force ramp experiment will fold and unfold many times during the experiment when held at a constant force. Hopping between the different conformations, many transitions are observed and the distribution of lifetimes can be measured. In principle, the lifetimes for each state as a function of force can be determined.

The last mode of force control is similar to a constant force experiment, in that the molecule hops between different conformational states. In this experiment, a trap position is held constant applying a constant potential to the system. When the molecule transitions to another conformational state with a different end-to-end extension, the average force changes (Figure 1.7 d). Lifetimes of each state are measured as a function of the average force and used to extract information about the energy landscape.

Regardless of the mode of force control, properly understanding how the force is changing and therefore how the energy of the system is evolving is critical to correctly interpret the data, i.e., the distribution of forces or lifetimes, and inferring information about system, such as the distance to the transition state. Part of this thesis will address the origin of a previously unreported discrepancy between the measured rate constants and distances to the transition state from constant force and constant trap experiments.

1.4 Considerations for the analysis of single molecule force spectroscopy data

In this section, I will focus on the factors that need to be considered for the analysis of single molecule data. This includes distinguishing the molecule of interest from artifacts and background in the system and characterizing the behavior of the molecule by correctly identifying the conformations and their lifetimes. Protein unfolding will be used as an example for the purposes of illustrating the factors that are important to the experiment.

The first factor is the choice of the pulling axis, or the reaction coordinate, which influences how force affects the reaction and therefore over what force range unfolding events occur. The combination of the length change of the molecule and the force at which the event occurs will determine the signal change associated with an event, hence the choice of the pulling axis dramatically affects the observable. A priori there is no way to predict over what force range these events will occur for a given protein and if the event will occur in single cooperative step or through intermediates, or a distribution of apparent cooperative and sequential events.

Another consideration is how to distinguish the relevant behavior from artifacts or background in the system. The first consideration is to determine that a tether is a single molecule and that the measured data do not arise from multiple tethers or non-specific interactions between the beads. By overstretching the DNA handles, a single tether can

be identified by the overstretching force (~ 67 pN) and measuring the length of the overstretching transition (Figure 1.8 a) [32].

The second is to assure that the behavior is due to a change in conformation of the protein of interested. The control for the experiments presented in this thesis is to tether beads together with dsDNA handles without a protein. These experiments allowed for the identification of an equilibrium-hopping artifact that occurs around 8 pN that is thought to arise from a conformation change in the attachment to the bead surface (Figure 1.8 b). This artifact occurred in approximately 5-10% of single tethers.

The third is to determine that the behavior is reproducible. As with any experiment, the data should be reproducible, but for single molecule experiments this criteria is enforced. For any set of single molecule data, there are subsets of behaviors, which are rare and irreproducible given the finite amount of molecules analyzed. These events are assumed to arise from interactions outside of the molecule of interest or from a change in the system, such the formation of a non-specific interaction during an experiment. Because of this, rare events or a minority of the population of molecules that exhibits different behavior are ignored in the final analysis. Using these criteria increases the confidence in the conclusions drawn from the average reproducible behavior from the single molecule experiments but it does not mean that we have captured the complete range of behaviors of the molecule. Working with the subset of data that is reproducible, the next step is to extract kinetic and thermodynamic information from the data.

1.4.1 Precision and accuracy

Of primary concern for any set of data is the precision and accuracy of the measurements. The accuracy determines how close the measured value is to the true value and the precision determines the reproducibility of the value. In this section, I address the accuracy and precision of the time, trap position, and force for the dual beam counter-propagating optical trap.

The time resolution was determined by the sampling frequency and the hardware used in the design of the instrument. Data were collected at 4 kHz and averaged down to 1 kHz before being written to disk. Due to hardware constraints, approximately 40% of the data points at 4 kHz were dropped. Therefore, the data at 1 kHz varied from an average of one to four data points collected at 4 kHz and consequently approximately two percent of the data at 1 kHz was not reported. The precision of the time measurements was therefore approximately 1 kHz but with increased noise because of the dropped data and the resulting inconsistent averaging of the data. Bypassing the limiting hardware and recording the signal directly from the detector improved the sampling resolution to 100 kHz, the limit of the position sensitive detector. The signal could then be averaged down to the desired frequency, typically 1 kHz, giving both high accuracy and precision in the time measurements.

The precision and accuracy of measuring the extension of the molecule is determined by how well the position of the trap in known. As only the position of the

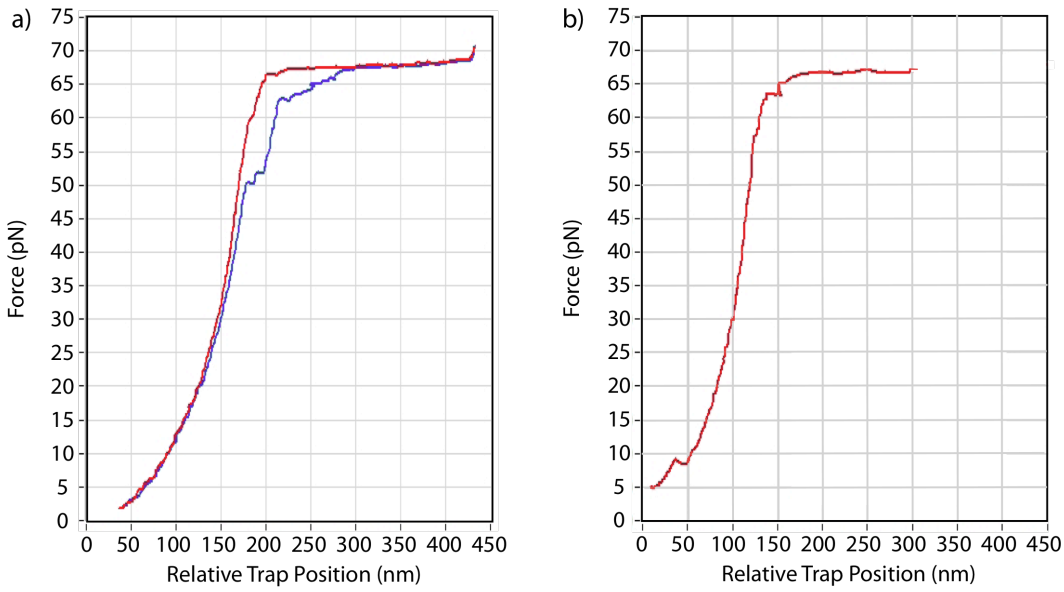

Figure 1.8 Overstretching of dsDNA handles.

The overstretching of 558 bp dsDNA handles attached via a disulfide bond. The dsDNA is shown without an artifact transition (a) and with the artifact transition at ~ 8pN.

trapped bead is monitored, the absolute extension of the molecule is unknown. Relative extension changes however can be determined and were inferred by measuring the relative changes in the position of the trap. In the experimental design, a part of the trapping laser was diverted from the origin to an independent detector. A piezo actuator moved the position of the origin of the laser thereby manipulating the position of the trap. The change in the position of the trap was then measured on the independent detector and provided precision in the position resolution to within 0.5 nm. As relative changes in extension were measured and not absolute distances, the accuracy and precision were the same.

The extension change of an unfolding event is measured by determining the trap position for both the folded and unfolded state at the same force. For example, during a force ramp experiment, the molecule unfolds at a given force and the force decreases because the length of the tether increases. The trap position continues to move and increase the force on the tether, pulling the molecule back to the original unfolding force. As the force is the same, the bead is the same distance from the center of the trap, and therefore, the difference in the trap position is the result of the extension change of the molecule at that force.

The force precision can be defined by the variability within a single tether or by the variability between tethers. The variability within a single tether can be quantified by measuring the reproducibility of a transition within a molecule, for example the overstretching transition in dsDNA or the folding and refolding rate of a molecule. In each of these cases, the standard error of the mean force is within 0.05 pN with a standard deviation of in 0.1 pN. However, between tethers there is much more variability. There is ~3 pN variability (95% confidence interval) in the overstretching transition around 67 pN between different fibers [32]. In other two-state systems, such as the p5ab RNA hairpin, there is a variability of ~0.5 pN (95% confidence interval) around 14 pN between equal rate constants (data will be shown later in this thesis). The precision determines how the data is processed, specifically, if tethers are analyzed separately or if the data is pooled from different tethers and then analyzed. Data can be pooled if the measurement spans a force range greater than the precision. As will be shown later, force ramp experiments demonstrating hysteresis typically have unfolding force distribution that spans 10 or more piconewtons and therefore the data can be pooled from multiple tethers. For equilibrium hopping behavior, the force range is approximately one piconewton and therefore pooling data from multiple tethers severely affects the results of the analysis.

The accuracy of the force measurement is calibrated by determining the voltage-to-force conversion factor. This calibration can then be checked through a variety of methods. For the experiments presented in this thesis, the calibration of the force trapping the bead was routinely checked using stokes law,

$$F = 6\pi\eta r\nu \tag{1.18}$$

where F is the force on the bead, η is the viscosity of the fluid, r is the radius of the bead, and ν is velocity of the bead. In this case, the instrument measures the force on the bead

as the chamber is moved with a known velocity. Given the viscosity of the fluid, the radius of the bead can be calculated and compared to the known value of the bead size. There is a variation in the size of the beads allowing for the force to be calibrated to within 10%.

1.4.2 Sampling frequency

As stated earlier, the behavior of the molecule of interest is inferred from the response of the probe. Despite the high sampling frequency of 100 kHz for position sensitive detectors, little information is gained by sampling at this rate as the response time of the bead is much slower. This response time of the bead is defined as the minimum time interval at which the data is uncorrelated from the previous data. This time interval can be determined with an autocorrelation function of the force or fitting the power spectrum of the force noise (Figure 1.9).

Some methods of analysis, such as a hidden Markov model approach, require that the each measurement is independent and uncorrelated from the previous time point. These methods require that the data is averaged or sub-sampled down to a time interval greater than the response time of the bead. One consequence of the limited response time of the probe is that events in the system that occur on a faster time scale will not be detected and any information at a higher frequency is averaged away.

Force feedback further limits the sampling frequency. The feedback works by changing the position of the trap in responding to changes in the force on the bead. To avoid oscillations in the force due to an overactive feedback, the feedback is over damped and is typically on a time scale of an order of magnitude slower than the response time of the probe. For typical experiments presented in this thesis, the corner frequencies range from 2.5 to 2 kHz and the feedback typically took from 200 to 100 Hz to return to the original force.

In this thesis, I will discuss a previously unreported complication that arises in constant-force feedback experiments. These experiments resulted in an underestimate of the measured rate constants of the system and an error in the determination of the distances to the transition state. Part of the reason this error went unnoticed was because limitations in the instrument and the data were neglected in the analysis of the data resulting in error.

1.5 Introduction to protein folding

Proteins perform a large variety of functions from structural roles to catalyzing reactions and are important in facilitating processes such as transcription, translation, signaling, and transport [33]. With 20 primary amino acids, the amount of possible sequences available to a protein is astronomically large, and perhaps, the large variety of functions performed by proteins should not be all that surprising. For example, given a small protein of 100 amino acids, there are 20^{100} different sequence combinations and this is not including post-translational modifications, which would further increase the sequence possibilities. However, given the potential sequences available, only a subset

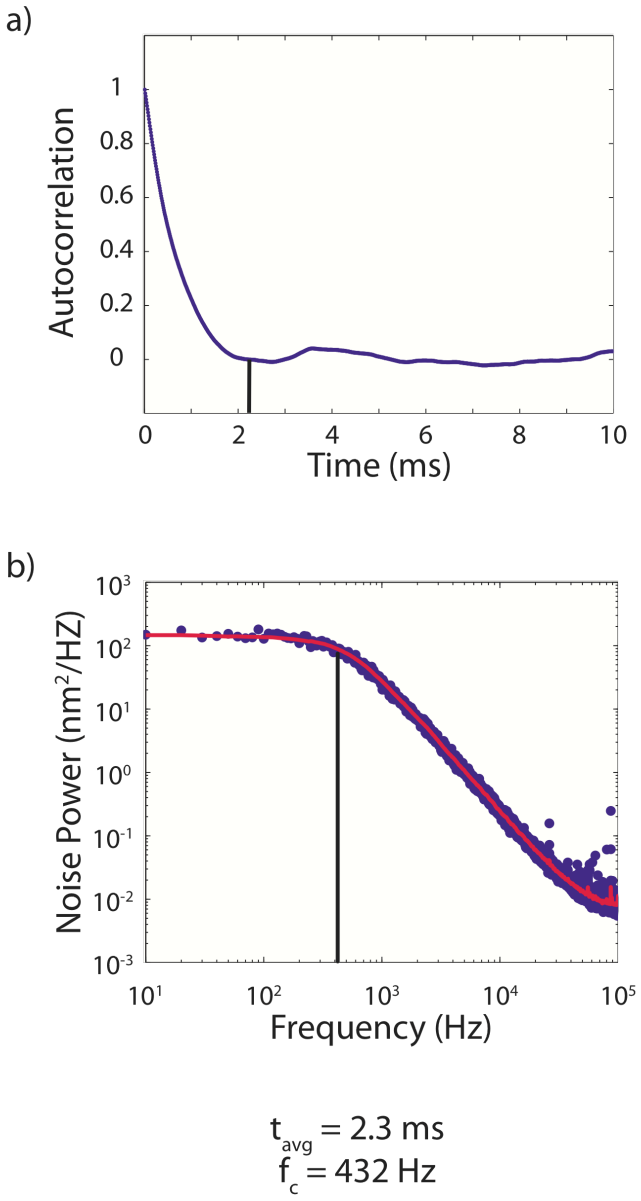

Figure 1.9 Power spectra and autocorrelation function of a free bead in an optical trap.

The autocorrelation function (a) and power spectra with fit of a free bead in an optical trap with an average correlation lifetime of 2.3 ms or 432 Hz. The black line marks the average lifetime and corner frequency in each figure.

has been observed in nature. Ultimately, the primary sequence defines the potential energy landscape of a polypeptide, which determines the conformations and interactions available to the polypeptide [34]. Natural selection works on this landscape to drive evolution.

Many proteins have different sequences yet still share a similar native state structure. Despite the structural similarities, these proteins can function very differently. One example that illustrates this point is homologous proteins from mesophiles and thermophiles, which have very different stabilities under the same conditions and function at very different temperatures despite sequence identities as high as 50% [35]. This exemplifies the complexities of determining the relationship between the sequence and the energetics of a protein. The link between sequence and structure was first elucidated in the 1960's, when Anfinsen and colleagues first demonstrated that many proteins are able to reversibly and spontaneously fold into their native structure, suggesting that for many proteins, the native, functional fold is the lowest energy state [36]. This work provided an understanding of the thermodynamics underlying the process but did not resolve anything about the kinetics of the folding process.

Regarding the kinetics of the folding process, Levinthal framed the dilemma in what came to be known as the Levinthal paradox [37]. The apparent paradox arose from the observation that single domain proteins typically fold to a relatively small native state ensemble on a millisecond to second time scale despite the huge number of possible conformations. To further illustrate the dilemma, take the example of a small protein with 100 amino acids. For this protein, there are 99 peptide bonds each with a phi and psi angle. If we assume that each angle can assume only one of three discrete conformations, then there are be $3^{(2*99)}$ or $\sim 10^{94}$ different possible conformations. Assuming the folding process was a random search of these conformations, even if the protein could explore a different conformation rapidly on the nanosecond to picosecond time scale, it would still take an astronomically long time ($\sim 10^{84}$ sec) to explore all the possible conformations and find the native state. The conclusion from this thought experiment is that there must be a biased search of the conformation space and therefore a folding pathway. It further suggests that the mechanism could be better understood if intermediates could be detected and characterized during the folding process.

1.5.1 The role of the molten globule state in protein folding

For many proteins, a burst-phase kinetic intermediate is detected early (less than 5 milliseconds) during folding to the native state [38, 39]. For several proteins, this state was shown to be similar to a partially folded state populated under equilibrium conditions at low pH. This equilibrium state, often referred to as a molten globule, appears to be a good experimental model for the early structures transiently populated during folding [39].

The molten globule state shares some common characteristics with natively folded proteins and other properties commonly associated with the unfolded state. Using x-ray scattering experiments, the molten globule state has shown to be compact and have a radius of gyration closer to that of the native state than the unfolded state [40]. Probing

the amount of secondary structure with circular dichroism has demonstrated that the molten globule state, like native proteins, contains secondary structure, although to usually a lesser extent than the native state. Unlike a natively folded protein, however, the molten globule state lacks fixed tertiary interactions and a well-packed hydrophobic core [41]. This has been demonstrated by a variety of methods. One assay for characterizing a molten globule is fluorescence due to 1-anilinonaphthalene-8-sulfonicacid (ANS). ANS binds to a hydrophobic pocket, resulting in the increase in fluorescence. For a molten globule state, ANS binding results in a higher fluorescence when compared with the native or unfolded state [42]. Heteronuclear single quantum coherence (HSQC) on a molten globule reveals poor dispersion, indicating poorly ordered side chains in the core of the protein [43].

An example of a well-characterized protein that folds through a molten globule-like intermediate is sperm whale apomyoglobin [44]. A structural model of the molten globule state of apomyoglobin has been generated by hydrogen exchange experiments [45] which were used to probe the protection factors of backbone amide hydrogen bonds in secondary structure elements. While the protection factors were significantly lower than those observed in the natively folded protein, the protected regions clustered in the residues that were part of the A, G, and H helices (Figure 1.10). The equilibrium molten globule state shares many properties with early kinetic intermediate, such as the amount of secondary structure (~35% of the native secondary structure) and the areas protected in hydrogen exchange experiments (helices A, G, and H) [44]. Another example of a protein that populates a molten globule-like intermediate is *E.coli* ribonuclease H (RNase H) [46-48]. Like apomyoglobin, *E. coli* RNAse H populates an early intermediate with properties of a molten globule. The role of this intermediate was investigated using the optical tweezers and following the folding trajectory of a single molecule, where the protein was observed to fold through an intermediate state that resembled the kinetic intermediate observed in ensemble experiments. Further, the intermediate appeared to be obligatory and on-pathway [49].

While significant progress has been made in understanding the structural nature and role of molten globules in the folding process, there are still many unresolved questions. Work using optical tweezers on *E.coli* RNase H indicated the intermediate has an unusually large distance to the transition state unlike natively folded proteins. This result raised the question as to if this property is a universal property of molten globule states or specific to *E.coli* RNase H or the pulling axis used in the experiment. Part of this thesis addresses this question, investigating the generality of this property using the equilibrium molten globule state of apo-myoglobin.

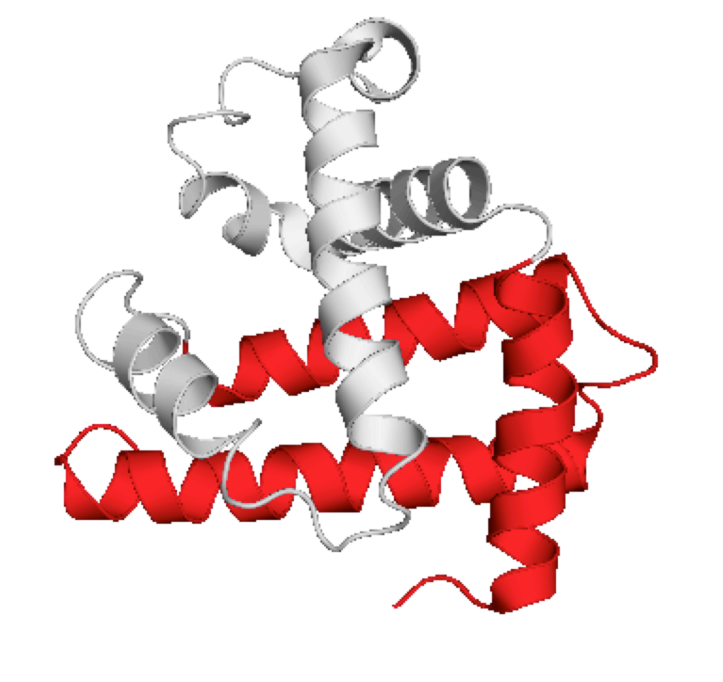

Figure 1.10 Model of the molten globule state of apomyoglobin.

The structure of holo myoglobin is shown with the A, G, and H helices highlighted in red. These helices correspond to the protected regions of the equilibrium molten globule and the kinetic intermediate during hydrogen exchange experiments.

1.6 References:

1. Di Lullo, G.A., et al., *Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen*, in *J Biol Chem*. 2002. p. 4223-31.
2. Cooper, G.M., *The Cell*. 2nd. ed. 2000.
3. Kellermayer, M.S., et al., *Folding-unfolding transitions in single titin molecules characterized with laser tweezers*, in *Science*. 1997. p. 1112-6.
4. Mitchison, T.J. and L.P. Cramer, *Actin-based cell motility and cell locomotion*, in *Cell*. 1996. p. 371-9.
5. Vogel, V. and M.P. Sheetz, *Cell fate regulation by coupling mechanical cycles to biochemical signaling pathways*, in *Curr Opin Cell Biol*. 2009. p. 38-46.
6. Bloom, G.E., S. , *Motor Proteins 1: Kinesin*. 1994, London: Academic.
7. Matouschek, A., et al., *Active unfolding of precursor proteins during mitochondrial protein import*, in *EMBO J*. 1997. p. 6727-36.
8. Sato, T., et al., *Comparison of the protein-unfolding pathways between mitochondrial protein import and atomic-force microscopy measurements*, in *Proc Natl Acad Sci USA*. 2005. p. 17999-8004.
9. Matouschek, A. and C. Bustamante, *Finding a protein's Achilles heel*, in *Nat Struct Biol*. 2003. p. 674-6.
10. Tinoco, I. and C. Bustamante, *The effect of force on thermodynamics and kinetics of single molecule reactions*, in *Biophys Chem*. 2002. p. 513-33.
11. Van't Hoff, J.H., in *Études de Dynamique Chimique*. 1884, Muller and Co.: Amsterdam. p. 114.
12. Arrhenius, S., Z. Phys. Chem., 1889. **4**: p. 226-248.
13. Scheffer, P.K.a.F.E.C., Proc. K. Ned. Akad. Wet., 1911. **13**: p. 789.
14. Brandsma, F.E.C.S.a.W.F., Recl. Trau.Chim.Pays- Bas, 1926. **45**: p. 522.
15. Trautz, M., *Z. Phys. Chem.* 1916. **96**: p. 1.
16. Lewis, W.C.M., J. Chem. Soc., 1918. **113**: p. 47.
17. Laidler, K.J. and M.C. Klng, *The Development of Transition-State Theory*, in *J. Phys. Chem.* 1983. p. 2657-2664.
18. Eyring, H., J. Chem. Phys., 1935. **3**: p. 107.
19. Hanggi, P., P. Talkner, and M. Borkovec, *Reaction-rate theory: fifty years after Kramers*, in *Reviews of Modern Physics*. 1990. p. 251-341.
20. Bell, G.I., *Models for the specific adhesion of cells to cells*, in *Science*. 1978. p. 618-27.
21. Greenleaf, W.J., et al., *Passive all-optical force clamp for high-resolution laser trapping*, in *Phys. Rev. Lett.* 2005. p. 208102.
22. Dill, K.A. and H.S. Chan, *From Levinthal to pathways to funnels*, in *Nat Struct Biol*. 1997. p. 10-9.
23. *Single-Molecule Techniques: A Laboratory Manual*, ed. P.R. Selvin and T. Ha. 2008. 1-507.
24. Neuman, K.C. and A. Nagy, *Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy*, in *Nat Methods*. 2008. p. 491-505.

25. Ashkin, A., *Acceleration and Trapping of Particles by Radiation Pressure*, in *Phys. Rev. Lett.* 1970. p. 156-159.
26. Ashkin, A. and J.M. Dziedzic, *PhysRevLett.38.1351*, in *Phys. Rev. Lett.* 1971. p. 1351-1354.
27. Neuman, K.C. and S.M. Block, *Optical trapping*, in *Rev Sci Instrum*. 2004. p. 2787-809.
28. Smith, S.B., Y. Cui, and C. Bustamante, *Optical-trap force transducer that operates by direct measurement of light momentum*, in *Meth Enzymol*. 2003. p. 134-62.
29. Bustamante, C.J. and S.B. Smith. *Light-force sensor and method for measuring axial optical-trap forces from changes in light momentum along an optic axis*. 2006 US 7133132 B2].
30. Cecconi, C., et al., *Protein-DNA chimeras for single molecule mechanical folding studies with the optical tweezers*, in *Eur Biophys J*. 2008. p. 729-38.
31. Bustamante, C., et al., *Single-molecule studies of DNA mechanics*, in *Current Opinion in Structural Biology*. 2000. p. 279-85.
32. Bustamante, C., et al., *Entropic elasticity of lambda-phage DNA*, in *Science*. 1994. p. 1599-600.
33. Berg, J.M., Tymoczko, J.L., Stryer L., *Biochemistry*. 2007.
34. Baldwin, R.L., *Energetics of protein folding*, in *Journal of Molecular Biology*. 2007. p. 283-301.
35. Jaenicke, R. and G. Böhm, *The stability of proteins in extreme environments*, in *Current Opinion in Structural Biology*. 1998. p. 738-48.
36. Anfinsen, C.B., *Principles that govern the folding of protein chains*, in *Science*. 1973. p. 223-30.
37. Levinthal, C., *Are there pathways for protein folding?*, in *Journal de Chimie Physique et de Physico-Chimie Biologique*. 1968. p. 44-45.
38. Kuwajima, K., *The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure*, in *Proteins*. 1989. p. 87-103.
39. Ptitsyn, O.B., et al., *Evidence for a molten globule state as a general intermediate in protein folding*, in *FEBS Letters*. 1990. p. 20-4.
40. Kataoka, M., et al., *Structural characterization of the molten globule and native states of apomyoglobin by solution X-ray scattering*, in *Journal of Molecular Biology*. 1995. p. 215-28.
41. Ptitsyn, O.B., *Structures of folding intermediates*, in *Current Opinion in Structural Biology*. 1995. p. 74-8.
42. Semisotnov, G.V., et al., *Study of the "Molten Globule" Intermediate State in Protein Folding by a Hydrophobic Fluorescent Probe*, in *Biopolymers*. 1991. p. 119-128.
43. Dobson, C.M., *Protein folding. Solid evidence for molten globules*, in *Curr Biol*. 1994. p. 636-40.
44. Jennings, P.A. and P.E. Wright, *Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin*, in *Science*. 1993. p. 892-6.
45. Hughson, F.M., P.E. Wright, and R.L. Baldwin, *Structural characterization of a partly folded apomyoglobin intermediate*, in *Science*. 1990. p. 1544-8.
46. Dabora, J.M. and S. Marqusee, *Equilibrium unfolding of Escherichia coli ribonuclease H: characterization of a partially folded state*, in *Protein Sci*. 1994. p. 1401-8.

47. Chamberlain, A.K., T.M. Handel, and S. Marqusee, *Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH*, in *Nat Struct Biol*. 1996. p. 782-7.
48. Raschke, T.M. and S. Marqusee, *The kinetic folding intermediate of ribonuclease H resembles the acid molten globule and partially unfolded molecules detected under native conditions*, in *Nat Struct Biol*. 1997. p. 298-304.
49. Cecconi, C., et al., *Direct observation of the three-state folding of a single protein molecule*, in *Science*. 2005. p. 2057-60.

Chapter 2 Equilibrium force spectroscopy experiments on macromolecules: The problem with force feedback experiment

2.1 Introduction

Single molecule force spectroscopy has provided important insights into the properties and mechanisms of biological molecules and systems. Here I focus on the frequently used approach of evaluating the force dependence of conformational changes at equilibrium. I evaluate the experimental method and analysis of equilibrium folding and unfolding studies of macromolecules under conditions of constant force or constant trap position. Using three different model systems (DNA, RNA, and protein), I demonstrate a previously unreported complication that arises from missed folding and unfolding events that are especially prominent in constant-force feedback experiments. These missed transitions lead to errors in calculated parameters such as the rate constants for the conformational transitions and the distances to the transitions state. I elucidate the cause of this problem and recommend a more robust strategy for collecting and analyzing such equilibrium data.

In recent years, the application of single molecule force spectroscopy to the study of biological molecules has provided insights unobtainable by previous methods. These studies have improved our understanding of the properties and mechanisms of a variety of systems under force such as DNA structure [1], RNA folding [2], protein folding [3, 4], and various molecular motors, from the initial work on kinesin [5] and myosin [6], to polymerases [7] and more complicated ring ATPases [8].

Several inherent advantages of force spectroscopy are exploited to allow for these novel insights. A typical experiment follows the trajectory of a single molecule in time, allowing for transient or rare events to be observed that would otherwise be masked when observing the average properties of a group of molecules in ensemble experiments. Applying force to a molecule perturbs the energetics of the system, influencing if and to what extent a conformational change occurs [9]. Finally, the reaction is measured along a defined reaction coordinate, the end-to-end distance of the molecule. Landmarks along this reaction coordinate, such as the distance to the transition state, can be inferred by measuring the lifetimes of given states as a function of force, allowing for a detailed mapping of the energy landscape.

Mapping this energy profile of a reaction requires the accurate identification and measurement of the lifetime of a given state at a given force. The accurate analysis of such single molecule time trajectories has been a long standing problem and several strategies have been developed for identifying the states and subsequently determining the associated lifetimes or rate constants [10-12]. In force spectroscopy, the force-dependence of these rate constants are typically fit using a modified Bell relationship [13, 14],

$$k(F) = k_m k_0 \exp\left(F x^{\ddagger} + \frac{1}{2}\kappa x^{\ddagger 2} \right) / k_B T \tag{2.1}$$

where k_m represents the contribution of experimental parameters such as the bead size, trap stiffness, and handle length to the observed rates, k_0 is the intrinsic rate constant of the molecule in the absence of force, F is the force, $x^{\ddagger}$ is the distance to the transition state, κ is the spring constant of the system, k_B is the Boltzmann constant, and T is the temperature in Kelvin. From this simple relationship, the distances to the transition state barriers ($x^{\ddagger}$) are determined.

All of this depends on an accurate determination of the rates and the effect of force, both of which depend critically on the type of experiment and its effect on the behavior of the molecule of interest. Of primary concern here is the mode of force control: constant force, constant trap position, or constant loading rate. Other parameters that also need to be considered are the trap stiffness, bead size, tether length, viscosity, and the sampling frequency. Previous work has attempted to determine the effect of some of these experimental parameters on the measured kinetics of a system [15, 16]; our evaluation of these studies, however, reveals a significant discrepancy between the results obtained with constant force and constant trap position experiments. This discrepancy was not noted and therefore unexplained.

Here, we revisit the effect experimental parameters have on the measurement of conformational lifetimes. Using data from DNA, RNA, and protein systems in which the molecule folds and unfolds during either a constant force or constant trap position experiment, we now identify a previously unreported complication arising from constant-force feedback experiments, and demonstrate that many of the reported properties, such as the rate constants as a function of force and the distance to the transition state, do not reflect the true behavior of the molecule of interest. Given an understanding of this problem and the role of the experimental parameters, we now suggest a more robust strategy for collecting and analyzing single molecule force spectroscopy data.

2.2 Materials and methods

2.2.1 Materials

The p5ab RNA hairpin from *Tetrahymena thermophilia* was provided by Jin Der Wen and was prepared as previously published [15]. The DNA hairpin data was provided by F. Ritort [17]. The wild-type myoglobin gene was provided by D. Barrick and the H36Q mutant of apomyoglobin was prepared as previously published [18].

2.2.2 Instrumentation

The instrument used in this experiment was a dual beam counter propagating optical trap. A piezo actuator controlled the position of the trap and allowed position resolution to within 0.5 nm with drift of less than 1 nm per minute [19]. The feedback controlled the position of the trap, and therefore the force on the bead and molecule, with a frequency of 2 kHz and a step size proportional to 10% of the force difference between the two states. An average force could be maintained within 0.01 pN of the set value with a standard deviation of 0.1 pN at 100 Hz. Data was collected at 4 kHz and averaged down to 1 kHz before being written to disk. Due to hardware constraints, approximately 40% of the data points at 4 kHz were dropped. Therefore, the data at 1 kHz varied from

an average of one to four data points collected at 4 kHz and consequently approximately two percent of the data at 1 kHz was not reported. All constant force experiments and the constant trap position experiments for the DNA hairpin were collected at 1 kHz. For the constant trap position experiments on the RNA and protein systems, higher frequency data was recorded at 50 kHz by bypassing the limiting hardware and recording the voltage corresponding to the force on the tether directly from the position sensitive detectors.

2.2.3 Methods of analysis

For all the constant force experiments, the data were averaged down to 100 Hz for analysis. For the constant trap experiments, the data was sub-sampled down to 1000 Hz for analysis. Because of the limited force precision between each single molecule (+/- 0.5 pN), they were analyzed separately.

2.2.4 Partition methods

Using a histogram of the data with a bin size of 0.5 nm, a partition was set to the minimum in the signal between the two states. A transition was detected when the signal crossed this partition, defining the beginning or end of the lifetime of the state. At a given average force, the rate constant for a given state was calculated by subtracting the sampling resolution (10 ms) off all lifetimes, and taking the inverse of the average of the lifetimes.

2.2.5 Bayesian hidden Markov model

The Bayesian hidden Markov model (BHMM) approach [12] employed automatically samples over likely assignments of the force measurements to the states, producing estimates of average forces and lifetimes characterizing each state, as well as confidence intervals that characterize the uncertainty in these values due to finite-sample statistics. After sub-sampling the force data to produce Markovian statistics (verified by examination of force autocorrelation functions; data not shown), the method first fits a maximum-likelihood HMM using standard procedures and then samples models consistent with the data using a Gibbs sampling strategy that assumes the force measurements of each state (including measurement error) are normally distributed about the average force for that state [12]. Here, the number of states was fixed to two after verifying the two-state nature of the data by inspection of the force traces. The first 50 HMM samples after starting from the maximum likelihood estimate were discarded to avoid any initial bias, and 1000 samples were generated to collect statistics on average forces and lifetimes, as well as generate the 95% confidence intervals reported here.

2.2.6 Determination of the distance to the transition state and the coincidental rate constants

For a given state, a liner fit of the natural log of the rate constants at each average force determined the distance to the transition state using the modified Bell model. The crossing point between the two fits determined the coincidental rate constant. All reported fits had R^2 values greater than 0.9. The values reported were the average of at

least five different tethers each analyzed separately and with data collected from five to twenty-five different average forces.

2.2.7 Simulation

The simple simulation of the constant force experiments modeled the molecule behavior under the constant force feedback. The simulation was run at 10 kHz and it assumed that the bead responded instantaneously to a change in the force. At each time point, the probability of a transition at the instantaneous force was calculated using the kinetic parameters measured from the constant trap position experiments. A random number generator determined an event and modeled the stochastic nature of folding and unfolding events. The feedback controlled the position of the trap, and therefore the force on the bead and molecule, with a frequency of 2 kHz and a step size proportional to 10% of the force difference between the two states. For each set of conditions, five simulated experiments were averaged down to 100 Hz and analyzed using the partition method and fit with the modified Bell model. For the nucleic acid system, the initial effective spring constant was set to 0.1 pN/nm and for the protein system the initial effective spring constant was set to 0.05 pN/nm, similar to the experimental conditions. The kinetic parameters of each system (i.e. the distance to the transition state and the intrinsic rate constants as a function of force) were representative of the results from the constant trap position experiments. The initial rate constants and the effective spring constants were varied between simulations in order to test the effect of the parameters on the experiment. As the purpose of this simulation was to probe the role of missed transitions, the simulation neglected any changes in the signal-to-noise ratio and the intrinsic rate constants as a function of the effective spring constant.

2.3 Results and discussion:

Single molecule force spectroscopy experiments were carried out using an optical trap. In this set-up, a single molecule is tethered between two polystyrene beads; a pipette tip holds one bead (2.1 um diameter) in place by suction and a dual counter-propagating beam optical trap manipulates the second bead (3.2 um diameter) (see Figure 2.1). By monitoring the bead in the trap, both the force and the relative extension of the tether can be determined [20]. The molecule of interest (either DNA, RNA, or protein) is attached to the beads through functionalized dsDNA (referred to as ‘handles’). These dsDNA handles provide space between both the bead surfaces and the molecule preventing any non-specific interactions with or between the beads from influencing the behavior of the molecule. The DNA handles are attached to the target molecule at specific sites thereby determining the axis along which the force is applied [2, 17, 21].

We carried out experiments on three types of macromolecules, all previously studied in the optical trap: a DNA hairpin [17], the p5ab RNA hairpin [2, 15], and the protein sperm whale apo-myoglobin at pH5 (manuscript in preparation). All three display two-state folding and unfolding transitions in both constant force and constant trap position experiments. In order to determine the lifetimes (and corresponding rate constants) at various forces, we followed many folding and unfolding events for a single tether in each experimental set up and repeated the experiment on multiple tethers (Figure 2.2).

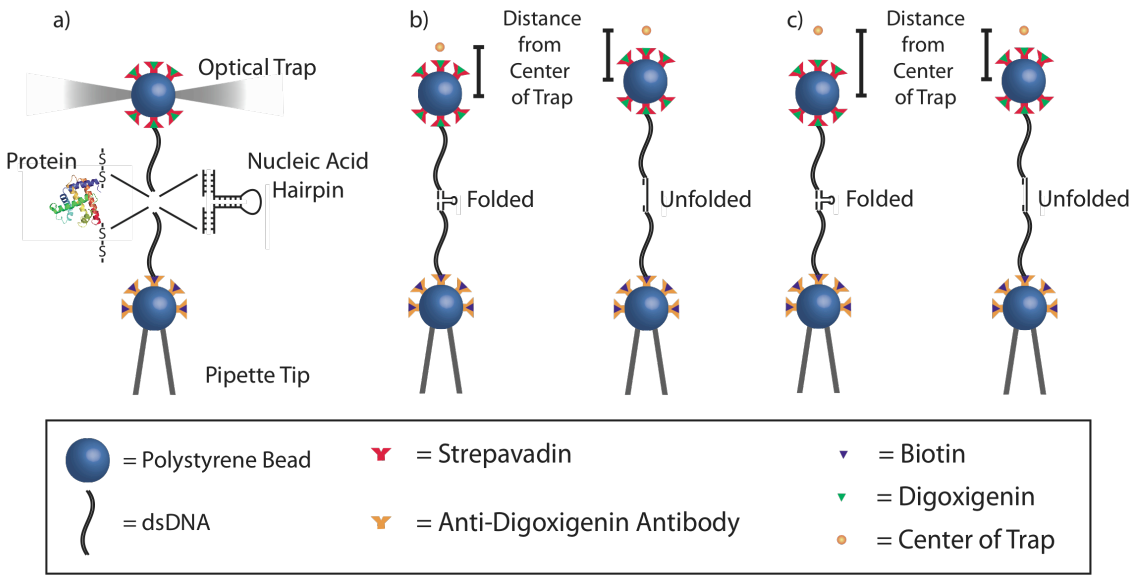

Figure 2.1 Optical trap experimental design.

a, Geometry of experiments depicting bead attachment via handles to the macromolecule of interest. **b**, In a constant force experiment, the bead distance from the trap center is maintained constant by the force-feedback controlling the position of the optical trap. **c**, In a constant trap position experiment, the trap position is constant and as the molecule folds or unfolds, the bead distance from the trap center changes, resulting in a change in the force on the bead.

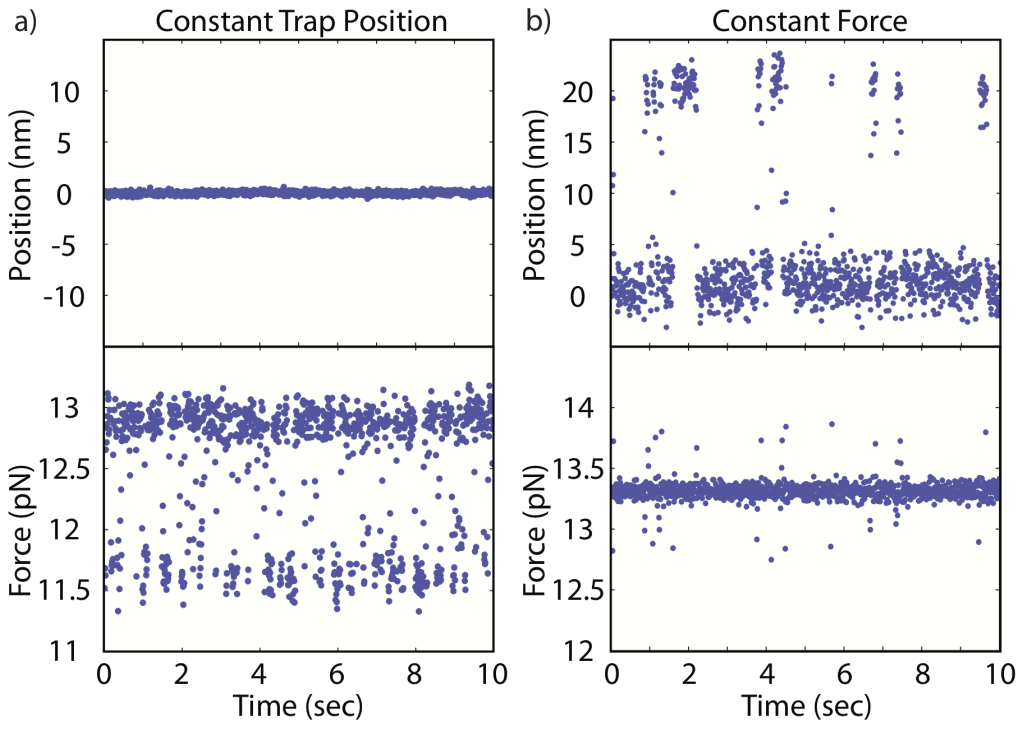

Figure 2.2 Constant trap position and constant force experimental data.

a, Trap position and force verse time for a constant trap position experiment. **b**, Trap position and force verse time for a constant force experiment. Data average down to 100 Hz.

Each experimental setup (constant force or constant trap position) requires a different signal to follow the trajectory of the molecule. In a constant force feedback experiment, the tether becomes shorter and the force increases when the molecule folds and the opposite occurs during unfolding. In order to maintain a constant force, the trap position moves and therefore, the extension, or trap position, reflects the conformation of the molecule. In this set up, the time scale of the force feedback determines the sampling time resolution. In the constant trap position experiments, when the molecule undergoes a folding (or unfolding) transition, the force increases (or decreases), and therefore, the force measurements are used to infer the state of the molecule. In this set up, the sampling time resolution is limited by the response time of the bead. This response time was determined experimentally by measuring the power spectrum of the force on the bead and determining the corner frequency of the system (which ranged between 1 kHz to 2.5 kHz).

Following these signals, identification of each state and their respective lifetimes were determined using two different approaches. The first is a simple partition method similar to those used in previous studies [2, 3, 15], where a histogram of the conformational signal over time shows a bimodal distribution, and a partition is set at the saddle point between the two peaks. A transition is defined whenever the signal crosses this defined partition, signaling the beginning or ending of a lifetime. Once defined, these lifetimes are used to calculate the average lifetime and corresponding rate constant. While simple and direct, this method requires clear resolution between the signals of each state; poor separation between the two states results in an overestimate of the number of transitions and a corresponding underestimate of the average lifetime of the state. For data with a lower signal-to-noise ratio, a second, more sophisticated approach is needed. This approach employs a Bayesian hidden Markov model [12] to identify both the states and determine the corresponding lifetimes with estimates of the error. For the constant-force data, we used the simple partition model - since a comparative study on a small set of data showed equivalent results with each method (data not shown). Constant trap position experiments, which have a lower signal-to-noise, required the BHMM method.

The resulting rate constants as a function of force were fit using the modified Bell model (Figure 2.3). Because the force range over which data were collected was small (1 to 2 pN), there is no detectable change in the distances as a function of the force (i.e. no change in the slope of ln k versus force), and therefore other models that account for a change in the distances to the transition state were not needed [22]. The resulting rate constants, distances to the transition state, and the rate constant at which the forward and reverse rate constants are equal (the coincident rate constant (k_c)) are shown in Table 2.1.

Table 2.1 demonstrates that the resulting kinetic parameters are dependent on the experimental set up. For each system, the coincident rate constants are lower when determined from the constant force experiments compared to the constant trap position experiments. The magnitude of this discrepancy varies depending on the molecule. For the DNA and RNA hairpin, the difference is 1.8 s^{-1} and 2 s^{-1}, respectively, while for the protein apomyoglobin, the difference is 9 s^{-1}. The sum of the distances to the transition state should equal the total distance between the two states as measured directly by the

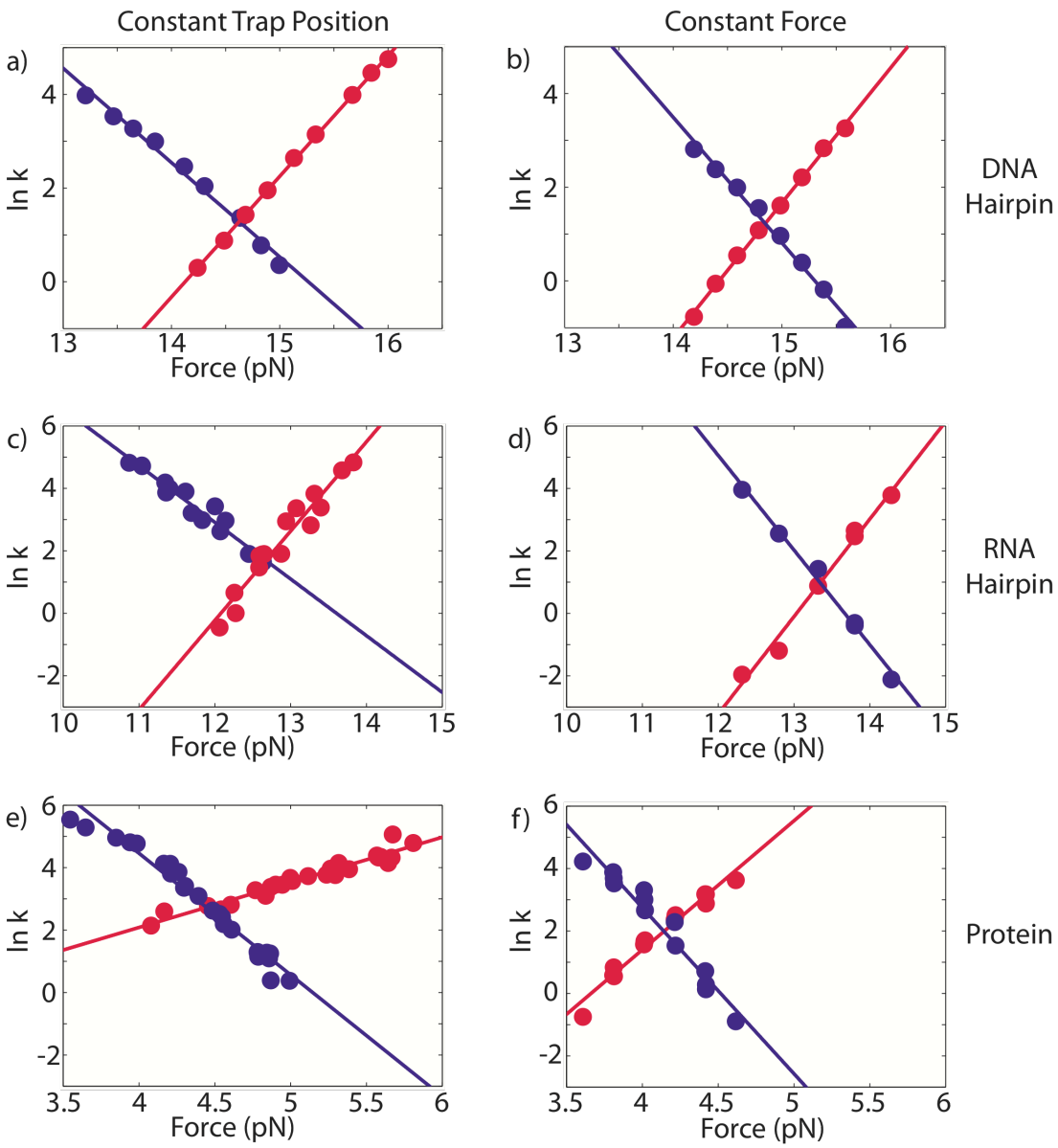

Figure 2.3 Linear fits of the ln *k* verse force.

a and **b**, Fits of the constant trap position and constant force data for the DNA hairpin. **c** and **d**, Fits of the constant trap position and constant force data for the RNA hairpin. **e** and **f**, Fits of the constant trap position and constant force data for the protein system.

Molecule	$\Delta x^{\ddagger}_{Unfolding}$ (nm)[1]	$\Delta x^{\ddagger}_{Folding}$ (nm)[1]	Δx_{Total} (Sum)(nm)[1]	Δx_{Total}(Measured) (nm)[2]	Ratio of Δx_{Total}[3]	$\ln(k_{Coincident})$
Constant Force						
DNA Hairpin						
Fiber 1	11.1 ± 1.2	11.9 ± 0.5	23.0 ±1.3			1.2
Fiber 2	11.8 ± 1.5	12.8 ± 0.8	24.6 ± 1.7			1.6
Fiber 3	11.1 ± 1.5	11.8 ± 0.4	22.9 ± 1.5			1.1
Fiber 4	11.3 ± 1.0	12.6 ±1.6	24.0 ± 1.9			1.3
Fiber 5	11.3 ± 1.6	11.5 ±2.8	22.8 ± 3.2			1.3
Average	11.3 ± 0.6	12.1 ± 1.1	23.5 ± 1.6	17.7 ± 0.4	1.33	1.3 ± 0.4
RNA Hairpin						
Fiber 1	12.3 ± 1.6	12.8 ± 1.8	25.1 ± 2.4			1.0
Fiber 2	12.6 ± 1.9	13.4 ± 8.4	26.0 ± 8.6			0.7
Fiber 3	14.7 ± 4.7	13.6 ± 6.1	28.3 ± 7.7			0.6
Fiber 4	11.4 ± 3.6	16.2 ± 3.5	27.6 ± 5.0			1.0
Fiber 5	11.3 ± 1.7	15.0 ± 2.9	26.3 ± 3.4			1.1
Fiber 6	11.7 ± 1.6	13.3 ± 0.8	25.0 ± 1.8			1.1
Fiber 7	11.0 ± 1.4	12.7 ± 0.9	23.7 ± 1.7			1.0
Average	12.1 ± 2.5	13.9 ± 2.6	26 ± 3.2	19.2 ± 0.4	1.35	0.9 ± 0.4
Protein						
Fiber 1	16.8 ± 2.1	21.8 ± 3.4	38.6 ± 4.0			1.9
Fiber 2	10.4 ± 4.3	28.3 ± 5.3	38.7 ± 6.8			2.1
Fiber 3	19.3 ± 1.3	19.8 ± 7.9	39.1 ± 8.0			2.4
Fiber 4	9.5 ± 1.9	23.4 ± 5.3	32.9 ± 5.6			2.1
Fiber 5	17.9 ± 15.6	21.5 ± 7.0	39.4 ± 17.1			1.5
Fiber 6	14.6 ± 3.1	19.7 ± 2.1	34.3 ± 3.7			1.7
Average	14.8 ± 8.1	22.4 ± 6.4	37.2 ± 5.6	18.9 ± 0.4	1.97	2.0 ± 0.6
Constant Trap Position						
DNA Hairpin						
Fiber 1	7.8 ± 1.1	10.3 ± 2.0	18.1 ± 2.3			1.9
Fiber 2	7.4 ± 0.9	10.2 ± 0.7	17.6 ± 1.1			1.5
Fiber 3	8.2 ± 0.6	10.1 ± 1.1	18.3 ± 1.3			2.0
Fiber 4	7.5 ± 0.9	9.3 ± 4.3	16.8 ± 4.4			1.7
Fiber 5	8.3 ± 0.8	10.6 ± 0.3	18.9 ± 0.9			1.3
Fiber 6	7.7 ± 0.5	10.3 ± 0.8	18.3 ± 1.0			1.7
Fiber 7	7.8 ± 0.5	10.2 ± 1.4	18.0 ± 0.9			1.6
Average	7.8 ± 0.7	10.2 ± 0.9	18.0 ± 1.3	17.7 ± 0.4	1.02	1.7 ± 0.5
RNA Hairpin						
Fiber 1	8.7 ± 0.8	13.3 ± 0.6	22.0 ± 1.0			1.4
Fiber 2	7.4 ± 1.1	11.7 ± 1.8	19.1 ± 2.1			1.7
Fiber 3	9.2 ± 0.5	10.6 ± 0.5	19.8 ± 0.7			1.4
Fiber 4	7.3 ± 0.7	11.5 ± 0.6	18.8 ± 0.9			1.0
Fiber 5	7.0 ± 0.5	10.9 ± 0.8	17.9 ± 0.9			1.8
Average	7.9 ± 1.9	11.6 ± 2.1	19.5 ± 3.1	19.2 ± 0.4	1.02	1.5 ± 0.7
Protein						
Fiber 1	6.1 ± 1.1	12.6 ± 2.3	18.7 ± 2.5			2.6
Fiber 2	6.5 ± 1.0	12.6 ± 2.3	19.1 ± 2.5			3.4
Fiber 3	6.7 ± 1.2	14.0 ± 1.9	20.7 ± 2.2			2.6
Fiber 4	5.3 ± 0.8	16.9 ± 1.3	22.2 ± 1.0			2.9
Fiber 5	5.9 ± 0.6	16.0 ± 1.3	21.9 ± 1.4			2.7
Average	6.1 ± 1.1	14.4 ± 3.9	20.5 ± 3.2	18.9 ± 0.4	1.09	2.8 ± 0.7

Table 2.1 Results from the linear fits of the constant force and constant trap position experiments for each individual molecule.

[1] Average values reported with a 95% confidence interval.

[2] Distance determined from fitting a histogram of the trap position from a constant force experiment with 2 Gaussian distribution and determining thedifference between the two centroids with a 95% confidence interval.

[3] Ratio of the calculated sum of the distances to the transition state to the experimentally measured distance between the two states.

change in extension upon undergoing a transition and consistent with a worm-like chain model of the macromolecule [23, 24]. Only the constant trap position experiments yielded distances consistent with the measured extension change (Table 2.1). However, for all three systems, the constant-force experiments produced a sum that was significant larger than the measured value (Table 2.1). The constant force experiments overestimated the total distance between the two states by an average of 34% for the nucleic acid hairpins and 97% for the protein.

For all three experimental systems studied, DNA, RNA and protein, the constant-force set up yields kinetic parameters that differ from the constant trap position experiments. The severity of the deviations between the values obtained in these two experimental set-ups depends on the specific molecule measured. Only the constant trap position experiments yield parameters consistent with the independently determined distances.

Why do the two types of experiments yield different results? The discrepancy must result from the effect of the force feedback, as all other experimental variables (sample, trap stiffness, bead size, handle length, etc.) were the same for each set-up. The discrepancy is not a result of the analysis method, as both the partition and BHMM method yield similar results when used to analyze the same constant force data set. The ability to maintain a constant force depends on the force feedback. Due to the finite limitations of the force feedback, the system can only maintain an average constant force on a time scale greater than 10 milliseconds. Force fluctuations that occur at smaller time scales could significantly alter the measured behavior of the molecule. If such fluctuations affected the behavior of the molecule, the constant force analysis would lead to inaccurate results. To illustrate this, consider a molecule in the unfolded state. When the molecule folds, the tether becomes shorter and the bead in the trap moves, resulting in a transient increase in the force. This higher force would then lead to a transient increase in the unfolding rate constant (Equation 2.1) which could in turn result in the molecule unfolding before the feedback has a chance to alter the position of the trap, resulting in a missed transition (Figure 2.4a).

Such missed transitions will affect the measured kinetic parameters, resulting in an underestimate of the rate constants and an overestimate of the distances to the transition state. Therefore, during force-feedback experiments transitions are missed when the molecule populates a given state for a short period relative to the time scale of the feedback (Figure 2.5). A missed set of transitions to and from the short-lived state results in an overestimate of the lifetime of the long-lived state. While these effects contribute to an overestimate of the average lifetime of both states, they have a more severe effect on the longer-lived state, resulting in a corresponding longer underestimate of its rate constant at the measured average force.

The difference between the actual rate constant and the measured rate constant as a function of force is proportional to the number of missed transitions. This difference becomes larger at more extreme forces when more transitions are missed to and from the short-lived state. As a consequence, the measured change in the rate constants (i.e. the distance to the transition state) is a function of not only the average

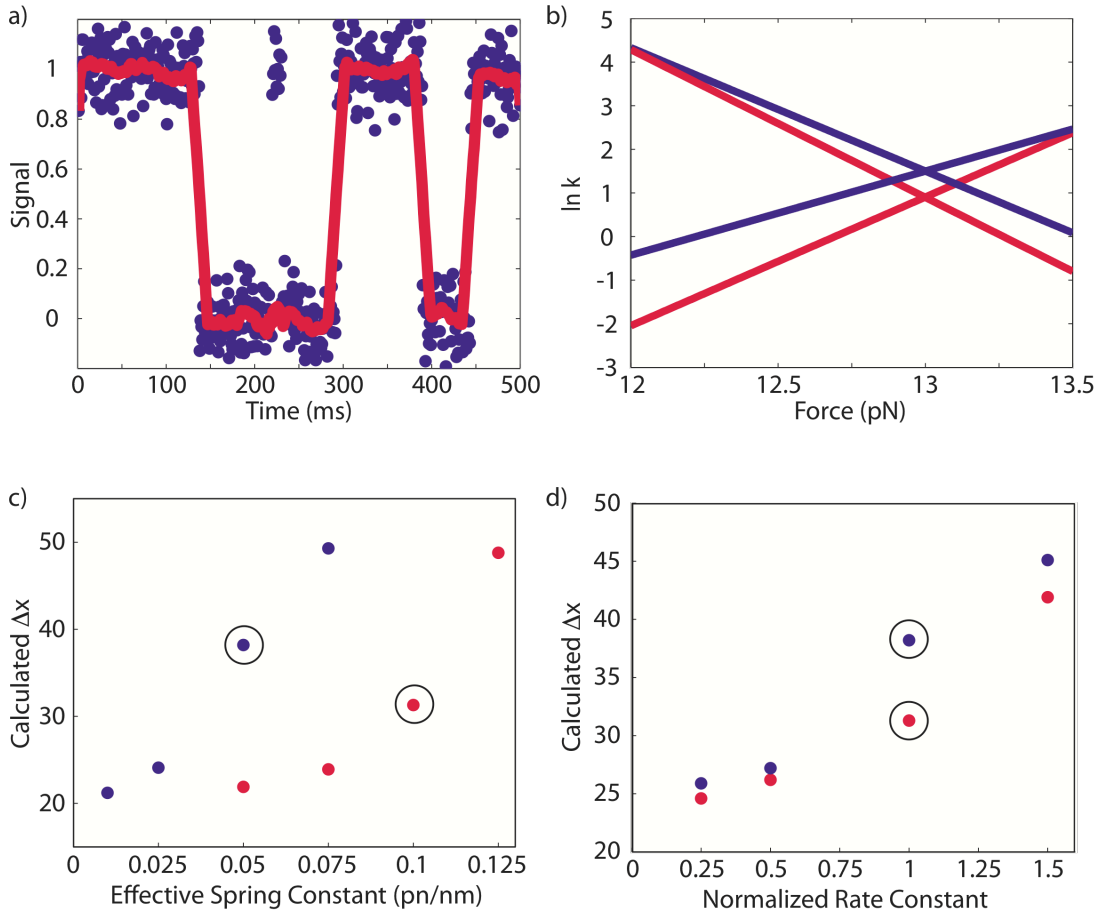

Figure 2.4 Illustration of the missed transition hypothesis and simulation results.

a, Illustration of a missed transition during a constant force-feedback experiment. In blue, the actual signal from the molecule and, in red, the measured behavior of the molecule. **b**, In blue, the ln *k* v. force for the true behavior of the molecule. In red, the ln *k* v force of the measured behavior of the molecule during a constant force experiment. **c** and **d**, Results from the simulation of the nucleic acid hairpin model (in red) and the protein model (in blue) depicting the measured behavior of the molecule as a function of the effective spring constant of the system (**c**) and the normalized rate constants (**d**). The circles indicate the results from the simulation using parameters similar to the constant force experiments.

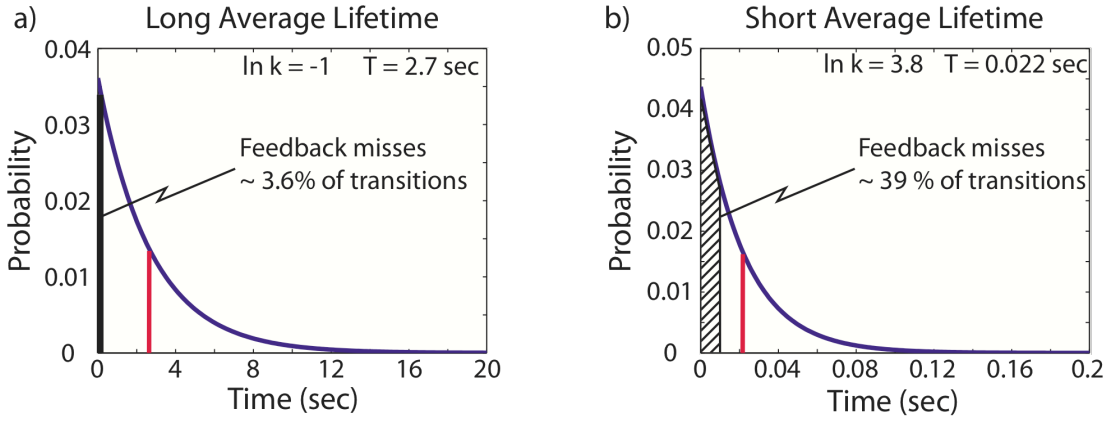

Figure 2.5 Illustration of the number of missed transitions relative to the average lifetime of a state.

The distributions of long (**a**) and short (**b**) lifetimes are shown for two different states, with the average lifetime marked by the red bar. The difference in the average lifetimes represents the change that would occur for a system with a total distance between the two states of 20 nm and an effective spring constant of 0.1 pN/nm, similar to the DNA and RNA hairpin systems. The black (**a**) and striped (**b**) regions indicate the number of missed transitions given a force-feedback cutoff of 10 milliseconds.

force, but also the number of missed transitions. This results in an overestimate of the distance to the transition state (Figure 2.4b).

The trends in our data are consistent with this missed transition hypothesis. For each system, the measured hopping rates are lower in the constant force experiments than in the constant trap position experiments. The distances to the transition state are larger in the constant force experiments than the constant trap experiments and, as previously mentioned, the sums of these distances in the constant force experiment are inconsistent with independent measurements. The missed transition hypothesis also predicts our observation that the greater the coincidental rate constant (measured by the constant trap position experiments), the larger the error in the distances to the transition state as measured by the constant force experiments. A system with a greater coincidental rate constant would, on average, have shorter lifetimes in both states when compared to a slower system. The shorter lifetimes result in more missed transitions and a larger error in the distance to the transition state. Both our DNA and RNA models have similar coincident rate constants and, consequently, have the same magnitude of discrepancy (approximately 34 %) in the total distance change relative to the value obtained by the constant trap position experiment. The protein has the largest coincident rate constant (i.e. the fastest hopping rate) and shows the largest distance error with a difference of 97%.

In order to better explain the discrepancies between the constant force and constant trap position experiments, we carried out a simple simulation. We evaluated two models, one representative of the DNA and RNA systems and the other representative of the protein system using kinetic parameters (the distance to the transition state and the intrinsic rate constants as a function of force) based on those measured by the constant trap position experiments. Constant force experiments were then simulated using the instrumental parameters, such as the frequency and size of the force feedback. These simulated data were then analyzed in the same manner as the experimental data. From these simulations, we modeled the effect of the force-feedback and other experimental parameters, such as the intrinsic rate constants of the molecule (i.e. how fast the molecule hops) and the effective spring constant of the system (i.e. the spring constant as a function of the trap stiffness and the length of the DNA handles), on the measured kinetic parameters. The results of the various simulations are shown for the nucleic acid hairpin model and the protein model in Figure 2.4c and 2.4d and Table 2.2.

The simulation results support the conclusion that the discrepancy between the constant force and constant trap position experiments is primarily a product of missed transitions and the magnitude of the error is a product of the interaction of the feedback with the underlying dynamics of the molecule. The simulations for both models produced similar distances to the transition state from their respective constant force experiments, validating the simulation. Changes in the parameters of the system (i.e. the rate constants or the effective spring constant of the system) changed the calculated kinetic parameters in proportion to the number of missed transitions. For example, decreasing the rate constants results in longer lifetimes and fewer missed transitions and,

Molecule	Effective Spring Constant (pN/nm)	Relative Rate Constant	$\Delta x^{\ddagger}_{Unfolding}$ (nm)[1]	$\Delta x^{\ddagger}_{Folding}$ (nm)[1]	Δx_{Total} (Sum)(nm)[1]	Ratio of calculated to expected Δx_{Total}[2]	$\ln(k_{Coincident})$
Nucleic Acid Hairpin							
Initial Parameters			7.9	11.6	19.5	1	1.5
	0.1	0.25x	12.4 ± 0.7	12.2 ± 0.5	24.6 ± 0.9	1.26	-0.1
	0.1	0.5x	13.9 ± 0.9	12.3 ± 0.6	26.2 ± 1.1	1.34	0.4
	0.1	**1x**	**16.0 ± 1.4**	**15.3 ± 0.8**	**31.3 ± 1.6**	**1.61**	**0.8**
	0.1	2x	27.2 ± 2.8	14.7 ± 1.5	41.9 ± 3.2	2.15	1.0
	0.1	4x	32.4 ± 3.1	24.3 ± 2.1	56.7 ± 3.7	2.91	0.9
	0.05	1x	9.9 ± 0.5	11.9 ± 0.3	21.8 ± 0.6	1.12	1.3
	0.075	1x	11.5 ± 0.7	12.5 ± 0.3	24.0 ± 0.8	1.23	1.2
	0.1	**1x**	**16.0 ± 1.4**	**15.3 ± 0.8**	**31.3 ± 1.6**	**1.61**	**0.8**
	0.125	1x	28.6 ± 3.4	20.3 ± 1.6	48.9 ± 3.8	2.51	0.1
	0.15	1x	45.1 ± 10.3	22.0 ± 2.6	67.1 ± 10.6	3.44	-1.4
Protein							
Initial Parameters			6.1	14.4	20.5	1	3.0
	0.05	0.25x	11.1 ± 1.2	14.8 ± 0.4	25.9 ± 1.3	1.26	1.3
	0.05	0.5x	12.0 ± 1.2	15.2 ± 0.3	27.2 ± 1.2	1.33	1.9
	0.05	**1x**	**22.7 ± 3.3**	**15.5 ± 0.6**	**38.2 ± 3.4**	**1.86**	**2.2**
	0.05	2x	28.1 ± 1.0	16.9 ± 1.6	45.0 ± 1.9	2.20	2.4
	0.05	4x	29.0 ± 3.1	21.2 ± 0.7	50.2 ± 3.2	2.45	2.8
	0.01	1x	7.7 ± 0.5	13.5 ± 0.4	21.2 ± 0.6	1.03	2.8
	0.025	1x	9.9 ± 0.8	14.3 ± 0.3	24.2 ± 0.9	1.18	2.7
	0.05	**1x**	**22.7 ± 3.3**	**15.5 ± 0.6**	**38.2 ± 3.4**	**1.86**	**2.2**
	0.075	1x	32.2 ± 3.6	17.0 ± 1.0	49.2 ± 3.7	2.40	1.7
	0.1	1x	39.9 ± 6.0	22.7 ± 2.2	62.6 ± 6.4	3.05	0.1

Table 2.2 Results from the linear fits of the constant force simulated data.

[1] Average values reported with a 95% confidence interval.
[2] Ratio of the calculated sum of the distances to the transition state to the expected sum of the distances to the transition states set in the initial parameters.

consequently, the coincident rate constants and distances to the transition state were closer to the true values. Increasing the initial rate constants results in more missed transitions and larger deviations in the measured kinetic parameters. A larger effective spring constant results in a larger change in the force and consequently a larger change in the lifetime of the state. This results in more missed transitions, resulting in a larger deviation between the calculated and the true kinetic values.

2.4 Conclusions

In summary, we have identified a previously unreported complication arising from constant-force feedback experiments on systems that hop between two or more states. This study demonstrates that the force feedback for all the systems studied result in an underestimate of the rate constants and an overestimate of the distances to the transition state. We conclude that a constant trap position experiment with sampling frequency limited only by the response of the bead is the best experiment for obtaining the highest quality data. When designing an experiment, a balance must be struck with the choice of the effective spring constant of the system. While a larger spring constant will increase the signal-to-noise ratio and decrease the rate constants of the system, a larger change in the force will occur between the two states resulting in larger changes in the rate constants. If the effective spring constant is too stiff, the range over which lifetimes can be measured will be limited and, taken to an extreme, the system will not hop at all. Further, even in a constant trap position experiment, if a system has lifetimes that are shorter than the response time of the bead, transitions will be missed possibly resulting in errors in the measured kinetic parameters. This emphasizes the importance of confirming that the total measured distance change equals the sum of the distances to the transition state.

2.5 References

1. Smith, S.B., Y. Cui, and C. Bustamante, *Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules*, in *Science*. 1996. p. 795-9.
2. Liphardt, J., et al., *Reversible unfolding of single RNA molecules by mechanical force*, in *Science*. 2001. p. 733-7.
3. Cecconi, C., et al., *Direct observation of the three-state folding of a single protein molecule*, in *Science*. 2005. p. 2057-60.
4. Gebhardt, J.C.M., T. Bornschlögl, and M. Rief, *Full distance-resolved folding energy landscape of one single protein molecule*, in *Proceedings of the National Academy of Sciences*. 2010. p. 2013-8.
5. Block, S.M., L.S. Goldstein, and B.J. Schnapp, *Bead movement by single kinesin molecules studied with optical tweezers.* Nature, 1990. **348**(6299): p. 348-52.
6. Finer, J.T., R.M. Simmons, and J.A. Spudich, *Single myosin molecule mechanics: piconewton forces and nanometre steps.* Nature, 1994. **368**(6467): p. 113-9.
7. Yin, H., et al., *Transcription against an applied force.* Science, 1995. **270**(5242): p. 1653-7.
8. Moffitt, J.R., et al., *Intersubunit coordination in a homomeric ring ATPase.* Nature, 2009. **457**(7228): p. 446-50.
9. Tinoco, I. and C. Bustamante, *The effect of force on thermodynamics and kinetics of single molecule reactions*, in *Biophys Chem*. 2002. p. 513-33.
10. Watkins, L.P. and H. Yang, *Detection of intensity change points in time-resolved single-molecule measurements.* J Phys Chem B, 2005. **109**(1): p. 617-28.
11. McKinney, S.A., C. Joo, and T. Ha, *Analysis of single-molecule FRET trajectories using hidden Markov modeling*, in *Biophys J*. 2006. p. 1941-51.
12. Chodera, J.D., et al., *Bayesian hidden Markov model analysis of single-molecule biophysical experiments: Characterizing metastable states and transition rates under measurement uncertainty.* In preparation, 2010.
13. Bell, G.I., *Models for the specific adhesion of cells to cells*, in *Science*. 1978. p. 618-27.
14. Greenleaf, W.J., et al., *Passive all-optical force clamp for high-resolution laser trapping*, in *Phys. Rev. Lett.* 2005. p. 208102.
15. Wen, J.-D., et al., *Force unfolding kinetics of RNA using optical tweezers. I. Effects of experimental variables on measured results*, in *Biophysical Journal*. 2007. p. 2996-3009.
16. Manosas, M., et al., *Force unfolding kinetics of RNA using optical tweezers. II. Modeling experiments*, in *Biophysical Journal*. 2007. p. 3010-21.
17. Forns, N., et al., *Improving signal-to-noise resolution in single molecule experiments using molecular constructs with short handles.* In preparation, 2010.

18. Elms, P.J., et al., *Characterization of the equilibrium molten globule state of apomyoglobin reveals a large mechanical compliance.* In preparation, 2010.
19. Bustamante, C. and S.B. Smith. *Light-force sensor and method for measuring axial optical-trap forces from changes in light momentum along an optical axis*. 2006.
20. Smith, S.B., Y. Cui, and C. Bustamante, *Optical-trap force transducer that operates by direct measurement of light momentum*, in *Meth Enzymol*. 2003. p. 134-62.
21. Cecconi, C., et al., *Protein-DNA chimeras for single molecule mechanical folding studies with the optical tweezers*, in *Eur Biophys J*. 2008. p. 729-38.
22. Dudko, O.K., G. Hummer, and A. Szabo, *Intrinsic rates and activation free energies from single-molecule pulling experiments*, in *Phys Rev Lett*. 2006. p. 108101.
23. Bustamante, C., et al., *Entropic elasticity of lambda-phage DNA*, in *Science*. 1994. p. 1599-600.
24. Bouchiat, C., et al., *Estimating the persistence length of a worm-like chain molecule from force-extension measurements*, in *Biophys J*. 1999. p. 409-13.

Chapter 3 Characterization of the equilibrium molten globule state of apomyoglobin reveals a large mechanical compliance

3.1 Introduction

The mechanism of how a protein folds into its functional, native state has been a long-standing fundamental question in biology. One aspect of particular interest is the observation that many proteins populate partially structured intermediate states during the folding process. While the role and structure of these states has been hard to determine, many have been shown to be compact and contain secondary structure, but lack any well-packed tertiary interactions. This state, commonly referred to as a molten globule, is thought to be a model for partially folded intermediates populated during folding of many globular proteins [1, 2].

Past mechanical studies on a variety of proteins have indentified several instances of unfolding and folding intermediates [3-11]. Force spectroscopy on recombinant *E. coli* ribonuclease H* Q4C/V155C (herein referred to as RNase H), identified and characterized an intermediate state in the folding process that was similar to the molten globule-like intermediate observed in ensemble studies of the protein [11]. Whereas most natively folded proteins studied to date have a small distance to the transition state (less than 2 nm) [4, 7, 11-19], this intermediate has the unusual property of a large distance to the transition state (5 ± 1 nm), indicating a large mechanical compliance. This observation raised the question of whether this large compliance was a general property of molten globule-like states or specific to the RNase H folding intermediate.

In order to investigate the generality of this property, we studied the mechanical properties of sperm whale apomyoglobin H36Q, which at pH 5.0, populates its equilibrium molten globule state that has been extensively characterized in previous studies [20-28]. We examined both the protein in both its native state at pH 7 and in the molten globule state at pH 5 along two different pulling axes. While unfolding of the native state appears to be brittle with a small distance to the transition state, the unfolding of the intermediate is compliant with a relatively large distance to the transition state. A large distance to the transition state was observed in both pulling axes, suggesting that the molten globule state is isotropic with a large compliance independent of the pulling axis.

3.2 Methods and materials

3.2.1 Protein construction and purification

The plasmid, pMB413b, containing the myoglobin gene was provided by D. Barrick (Johns Hopkins University). Using a PCR Quickchange protocol (Stratagene), a N-terminal six Histadine tag followed by a TEV protease site and the H36Q mutation were inserted into the gene. Two variants of this gene were produced with cystienes either at the N- and C-terminus (hereafter refereed to as the N/C variant) or at residue 53 (A53C) and the C-terminus (hereafter refereed to as the 53/C variant) (Figure 3.1).

The plasmid was transformed into chemically competent BL21 (DE3) pLysS cells. The cells were grown in Luria Broth (200 ug/ml of ampicillian, 34 ug/ml

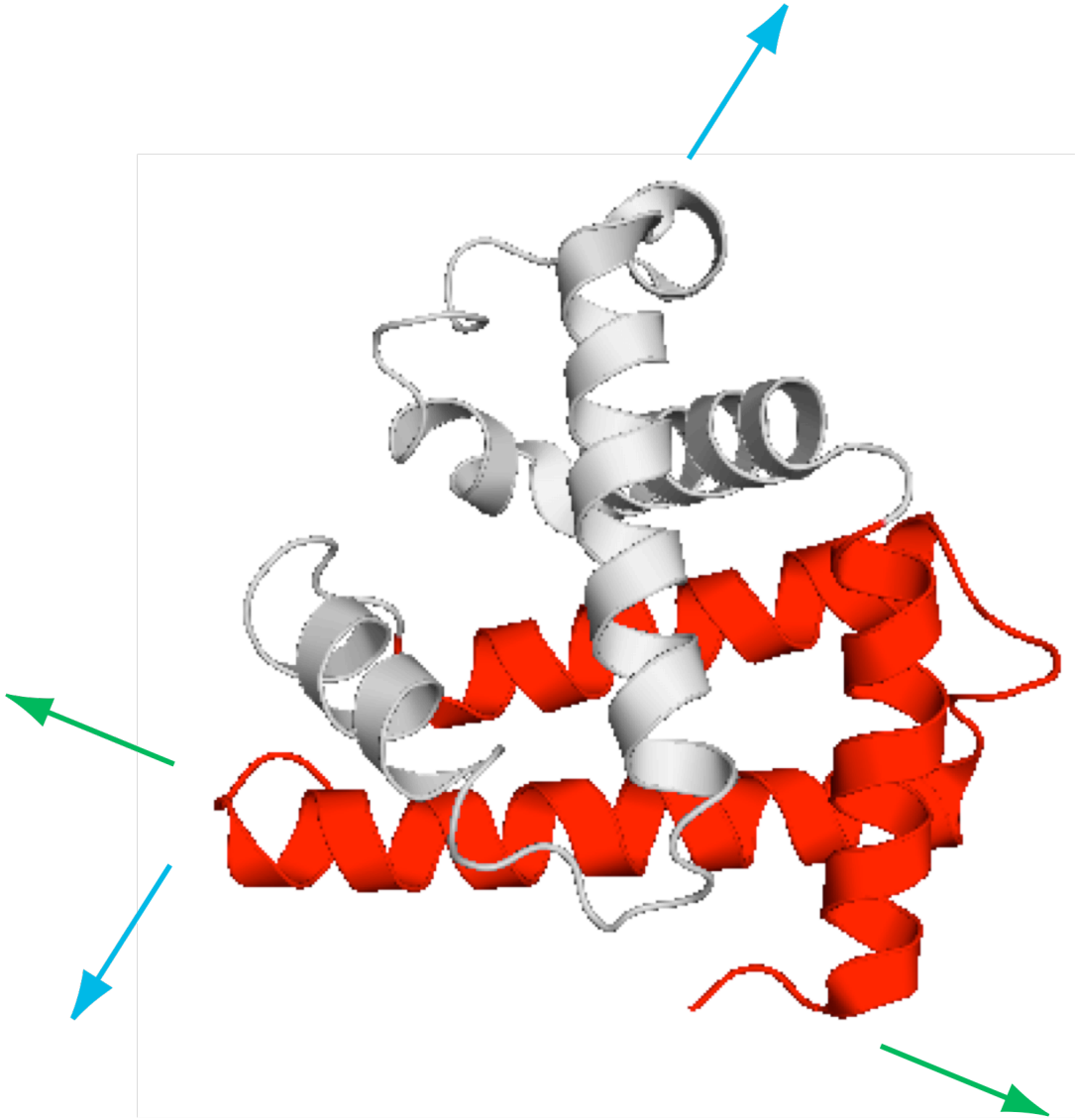

Figure 3.1 Structure of myoglobin.

Structure of holomyoglobin (PDB ID: 1BZ6). The regions highlighted in red indicates the regions of the protein that are structured in the molten globule state as determined by hydrogen exchange. The pulling axis for the N/C variant (in green) and the 53/C variant (in blue) are indicated with the arrows.

chloramphenicaol) and the protein was constitutively expressed for 24 hours at 37 degrees Celsius, allowing for the incorporation of the heme. Cells were harvested by centrifugation and re-suspended in 20 mM sodium phosphate, pH 8, 300 mM sodium chloride, and 0.5 mM TCEP. The cells were lysed by sonication and the soluble fraction was isolated by centrifugation followed by filtration through a 0.2 um filter. This lysate was purified over a nickel sepharose column followed by a step elution with 20 mM sodium phosphate, pH 8, 300 mM sodium chloride, 0.5 mM TCEP, and 250 mM imidazole. The elution was dialyzed overnight at 4 degrees Celsius against 20 mM sodium phosphate, pH 8, 300 mM sodium chloride, and 0.5 mM TCEP in the presence of TEV protease (2 mg/L) to remove the N-terminal 6x Histadine tag. The dialysate was then run over a second nickel sepharose column and the flow-through was collected. Analysis by SDS-PAGE demonstrated that the protein was greater than 95% pure. The heme was removed by a Vydac C-18 reverse phase column using 0.1 % trifluoroacetic acid (TFA) and eluting the protein with a linear gradient of 0.1%TFA and acetonitrile on a Shimadzu HPLC system. DNA handles were attached to both holo- and apo-myoglobin using previously reported methods [14, 29].

3.2.2 Optical tweezer experiments

The instrument used in this experiment was a dual beam counter-propagating optical trap [30]. The spring constant of the trap was set to 0.05 pN/nm. A piezo actuator controlled the position of the trap and allowed position resolution to within 0.5 nm [30]. An average force could be maintained to within 0.1 pN over the course of a typical one-minute constant trap position measurement. Because of the limited force precision between fibers (+/- 0.5 pN), each fiber was analyzed separately.

The protein was tethered between two polystyrene beads through functionalized dsDNA attached to the molecule at the sites of cysteine modification, thereby determining the axis along which the force was applied (Figure 3.2). Further, these dsDNA "handles" provided space between both the bead surfaces and the molecule preventing any non-specific interactions with or between the beads from influencing the behavior of the molecule. In this experimental setup, one bead (2.1 um diameter) is held in place on a pipette tip by suction and a dual beam anti-propagating optical trap manipulates the other bead (3.2 um bead diameter). By monitoring the bead in the trap, the force on the tether and its relative extension of the tether were measured.

The molecule was characterized by three different types of experiments: force ramp, force jump, and constant trap position experiments. The unfolding and refolding behavior was first analyzed using force-ramp studies which involved moving the trap position in a cycle between 3 pN and 20 pN at a constant velocity of 100 nm/sec. These force data were smoothed with a sliding window of 10 ms and the difference at intervals of 10 ms was determined. The standard deviation of the difference was then calculated and a threshold for detection of an event was set based off the variance in this signal. A 99.9 % confidence threshold (3.3 σ) was set for the high signal-to-noise high-force events (greater than 7 pN) and a 90% confidence threshold (1.6 σ) was used for the noisier low-force events (lower than 7 pN). The position with the highest difference signal was

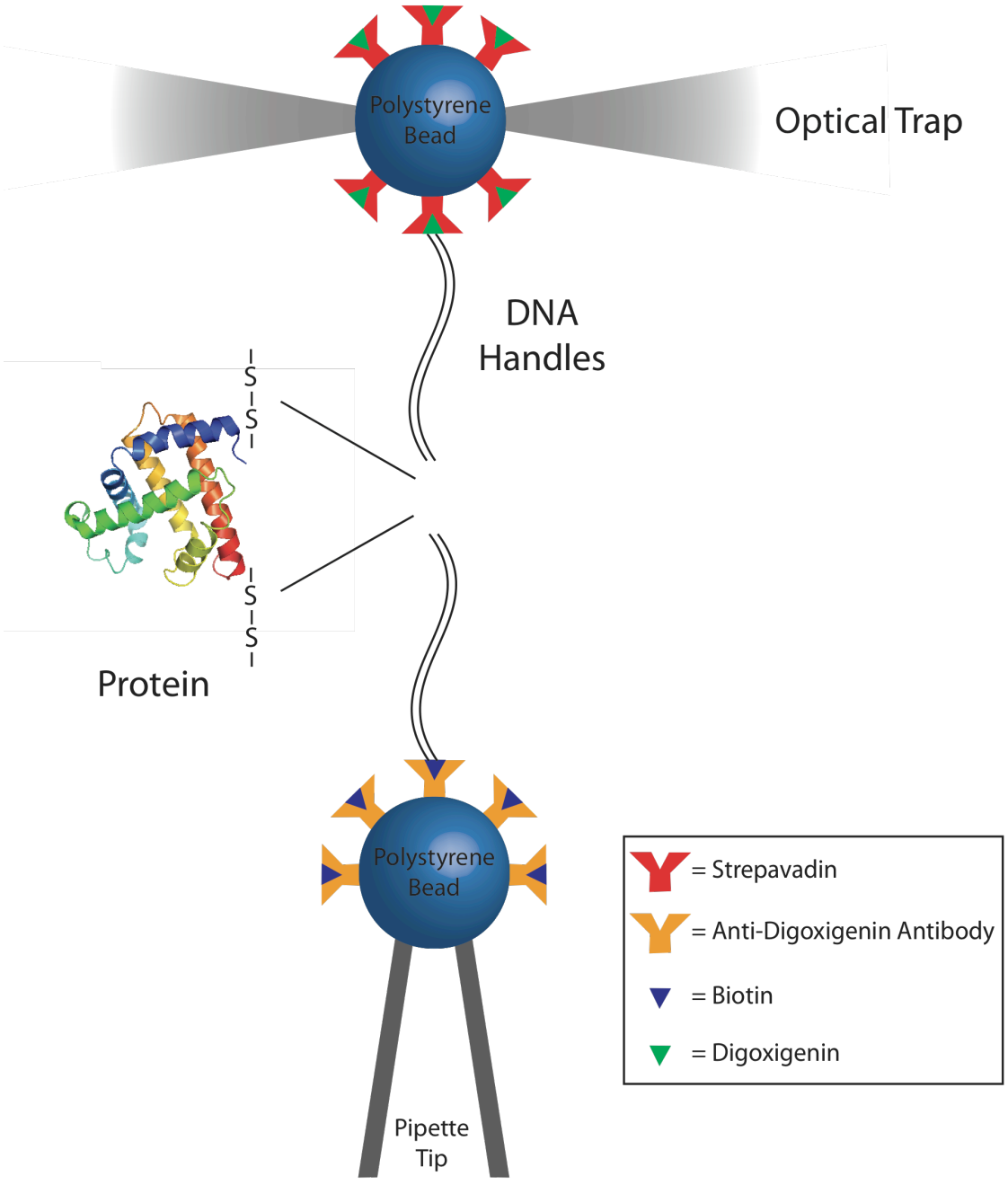

Figure 3.2 Experimental setup in the optical tweezer.

identified as the folding or unfolding event. From this experiment, the unfolding or folding force distribution as a function of the loading rate and the extension change of the molecule as a function of force could be measured.

For force-jump experiments, the molecule was held at a low force of 3 pN for several seconds to ensure that the molecule was in the folded state. The force was then quickly increased to the specified force and held constant by feedback. By adjusting the trap position, the feedback maintained an average force with a standard error of 0.01 pN of the set value. Using the position of the trap, unfolding events were detected using a t-test similar to previous methods [31] and used to determine the lifetime of the folded state of the molecule at the average force.

During constant trap position experiments, the molecule folded and unfolded, hopping between two states, with many transitions observed in a single tether. The force signal was used to identify the state (the higher force indicated the more compact, folded state) and determine its lifetime. The trap was held at each constant trap position for one minute with at least ten different trap positions collected per tether.

The holo- and apo- protein molecules were characterized by force ramp and force jump experiments in 10 mM Tris, pH 7, 250 mM sodium chloride, and 1 mM EDTA. Each tether was then held in the trap and the buffer conditions were changed to 10 mM sodium acetate, pH 5.0, 250 mM sodium chloride, and 1 mM EDTA. The unfolding and folding behavior were then characterized by force ramp and constant trap position experiments at the new pH.

Force ramp experiments at pH 7 revealed no difference in the holo- and apo-proteins behavior excluding the first unfolding event (data not shown) indicating that the non-covalently bound heme disassociates from the holo protein upon initial mechanical unfolding. As only one unfolding event was observed per holo, it was impossible to obtain quantitative data for the unfolding behavior, although qualitatively it appeared to unfold at a higher force than the native apomyoglobin under these conditions.

3.2.3 Constant trap position analysis with a Bayesian hidden Markov model

The Bayesian hidden Markov model (BHMM) approach [32] sampled models over the force measurements, producing estimates of average forces and lifetimes characterizing each state, as well as confidence intervals that characterize the uncertainty in these values due to finite-sample statistics. After sub-sampling the force data to produce Markovian statistics (verified by examination of force-autocorrelation functions; data not shown), the method first fits a maximum-likelihood HMM using standard procedures [33] and then samples models consistent with the data using a Gibbs sampling strategy that assumes the force measurements of each state (including measurement error) are normally distributed about the average force for that state [32]. Here, the number of states was fixed to two after verifying the two-state nature of the data by inspection of the force traces. The first 50 HMM samples after starting from the maximum likelihood estimate were discarded to 'burn-in', and 1000 samples were generated to collect statistics on average forces and lifetimes, as well as generate the 95% confidence intervals reported here.

3.2.4 Determination of the distance to the transition state and the coincidental rate constants using a modified Bell's model

For a given state, a liner fit of the natural log of the rate constants at each average force determined the distance to the transition state using a modified Bell's model [34, 35],

$$k(F) = k_m k_0 \exp\left(\frac{Fx^{\ddagger} + \frac{1}{2}\kappa x^{\ddagger 2}}{k_B T}\right) \tag{3.1}$$

where k_m represents the contribution of experimental parameters such as the bead size, trap stiffness, and handle length to the observed rate, k_0 is the intrinsic rate constant of the molecule in the absence of force, F is the force, $x^{\ddagger}$ is the distance to the transition state, κ is the effective spring constant of the system, k_B is the Boltzmann constant, and T is the temperature in Kelvin. For the constant trap position experiments, the crossing point between the two fits determined the coincident rate constant and force. For the N/C variant, all reported fits had R^2 values greater than 0.9. Because of a lower signal-to-noise ratio for the 53/C variant, all fits had R^2 values greater than 0.7. The reported values were the average of at least five different fibers each analyzed separately.

3.2.5 Equilibrium denaturation by pH monitored by circular dichroism

The circular dichroism signal was measured as a function of pH for the N/C variant apomyoglobin construct. Samples were equilibrated overnight at 25° C in 5 mM citrate, 250 mM sodium chloride, 0.5 mM TCEP, and 0.05 mg/ml of apomyoglobin at a variety of pHs ranging from 4 to 6.5. The signal at 222 nm was averaged over 60 seconds at 1 Hz with a 1 nm bandwidth at 25° C on an Aviv model 410 spectrometer.

3.3 Results:

3.3.1 Equilibrium denaturation by pH monitored by circular dichroism

In order to assure that the cysteine variants did not alter the pH-dependent transition to the molten globule state, pH denaturation studies were carried out by monitoring the circular dichroism signal at 222 nm (Figure 3.3). The acid unfolding to the molten globule state for the N/C variant was similar to that for the parent protein (H36Q apomyoglobin) [36] confirming that at pH 5.0 the protein populates the molten globule.

3.3.2 Unfolding and refolding of apomyoglobin at pH 7 under force

When pulled from the ends (the N/C variant), the protein unfolds and refolds with a notable hysteresis during a force ramp experiment at pH 7 (Figure 3.4 a). A histogram of the unfolding forces at a pulling speed of 100nm/sec is bimodal with peaks centered on 12.5 pN and 6.1 pN (Figure 3.5 a), indicating that the protein was unfolding from two different states. The refolding forces were distributed in a single peak around 4.5 pN. As

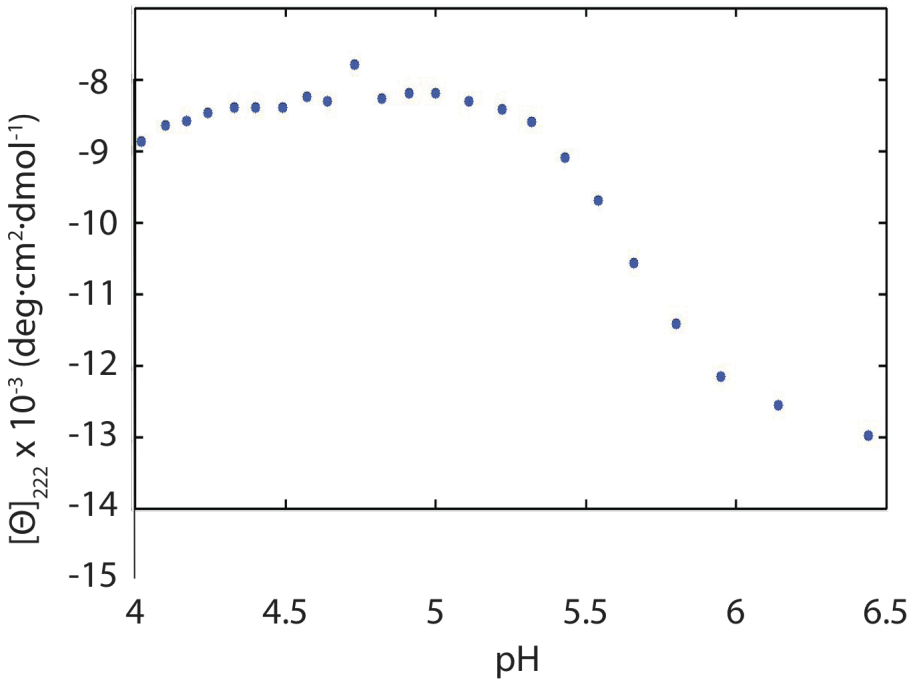

Figure 3.3 Denaturation of the N/C variant of apomyoglobin by pH followed by CD at 222 nm.

The native state is unfolded by pH, populating the molten globule state below pH 5.0 as shown by the mean residue ellipticity at 222 nm as a function of pH.

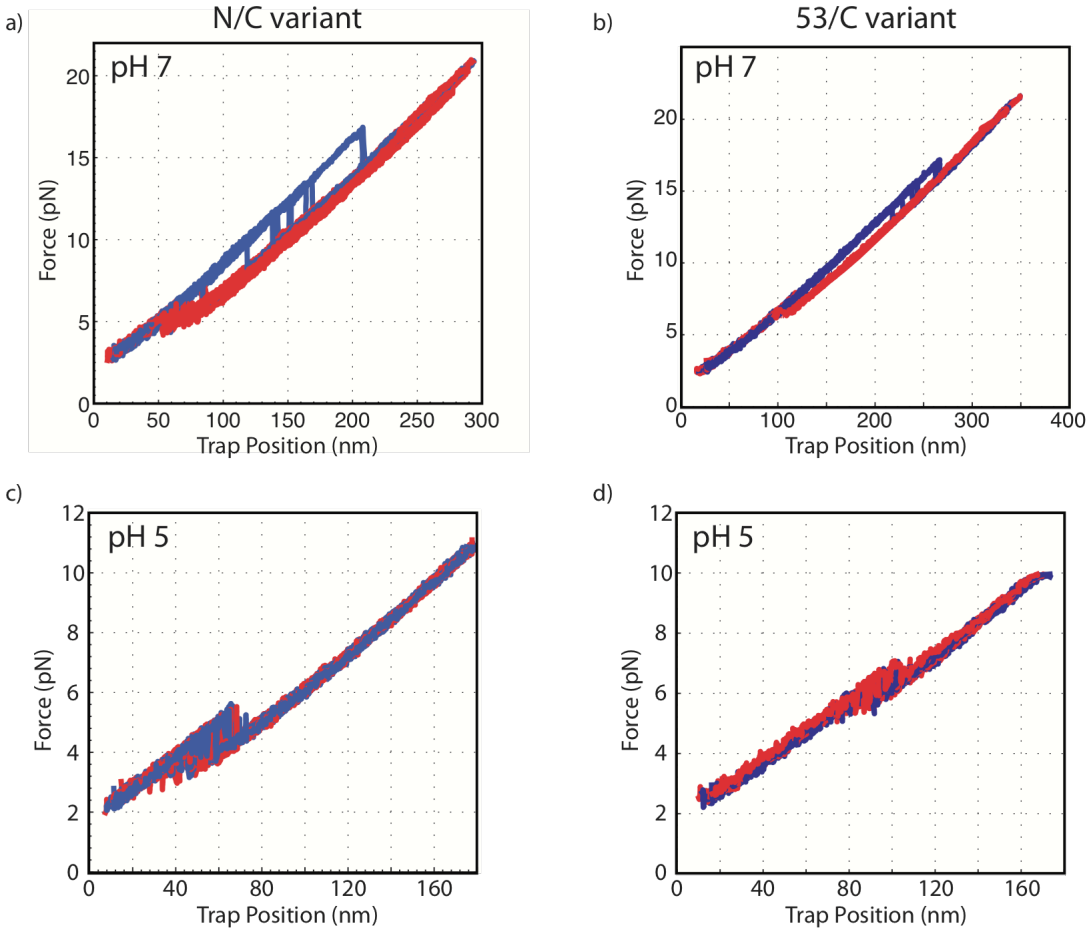

Figure 3.4 Force ramp traces of the N/C and 53/C variant at pH 7 and pH 5.

Force-ramp experiments are depicted showing the force as a function of the trap position with the pulling traces shown in blue and the relaxation traces shown in red. Traces from the N/C variant are shown in **a** and **c** at pH 7 and pH 5, respectively. Traces from the 53/C variant are in **b** and **d** at pH 7 and pH 5, respectively.

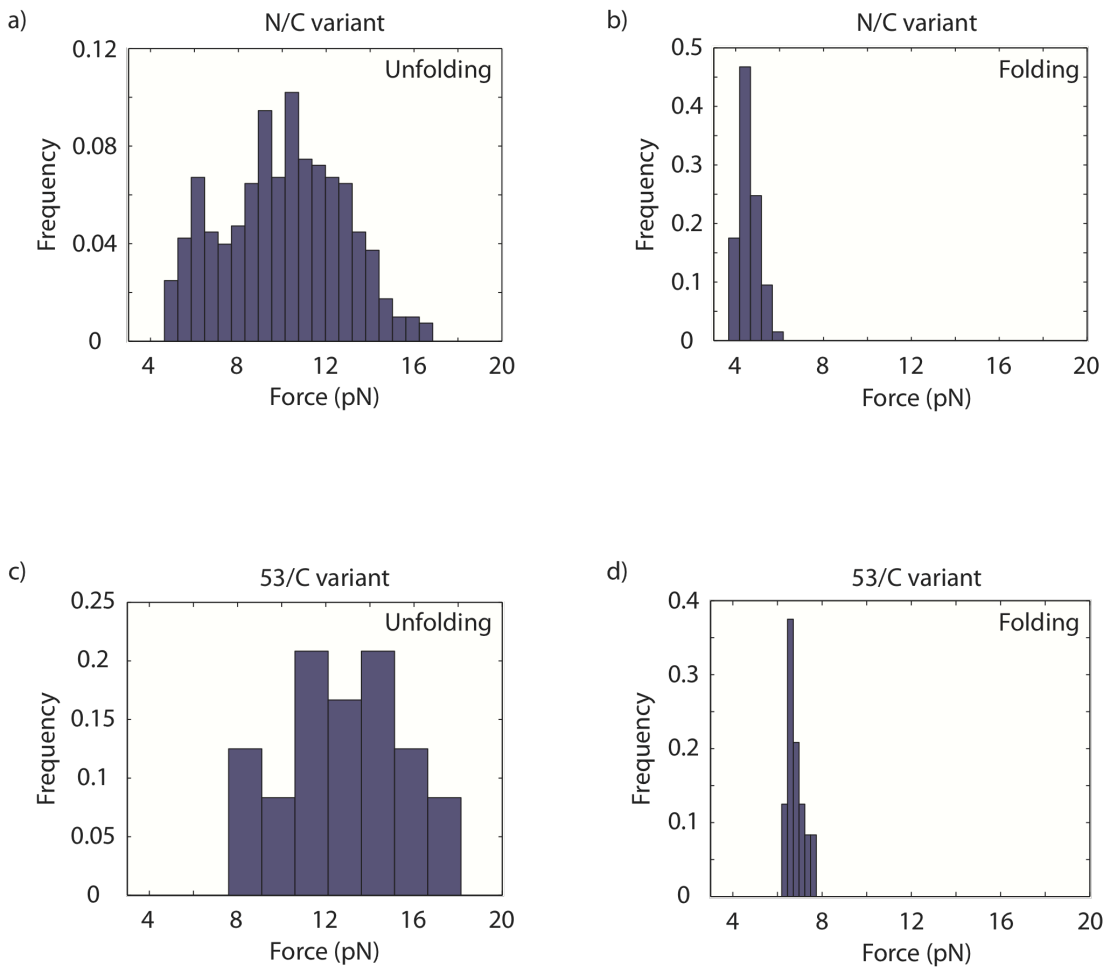

Figure 3.5 Histograms of the unfolding and refolding force distribution for the N/C and 53/C variants.

The different force distributions are depicted for the unfolding (**a**) and refolding (**b**) for the N-terminal construct at pH 7. Similarly, the unfolding (**c**) and refolding (**d**) force distribution are shown for the 53/C variant.

the dwell or delay time between each cycle at 3 pN was increased, the relative population of each peak shifted towards the high force unfolding events. Using long wait times at low force, the properties of the more high-force unfolding transition were investigated with constant force-jump experiments. Using a modified Bell's model, the distance to the transition state from the high unfolding-force state was measured to be 1 ± 1 nm.

The extension change of the molecule was inferred from the difference in the trap positions in the unfolded and folded state at the force at which the event occurred. As the force on the bead is the same in each state, the bead is the same distance from the center of the trap. Therefore, any change in the position of the trap reflects the change in the end-to-end extension of the molecule. The measured extension change is the difference between the unfolded end-to-end extension and the folded end-to-end extension of the molecule.

$$\Delta x(\text{measured}) = x_{\text{Unfolded}}(\text{wlc}) - x_{\text{Folded}} \tag{3.2}$$

Using a worm-like chain model [37] given the persistent length (P=0.65) and temperature, the end-to-end extension of the unfolded state at a given force can be calculated assuming a contour length. Assuming that the less compact conformation is completely unfolded, the contour length was calculated from the number of amino acids between the handle attachment points, in this case 55.4 nm. With this end-to-end extension of the unfolded state, the end-to-end extension of the folded state can be inferred. For both the unfolding and refolding events, the extension changes were consistent with the protein unfolding from a compact state with an end-to-end extension between the cysteines of ~2.5 nm.

Force ramp experiments on the 53/C variant revealed hysteresis between the unfolding and refolding events (Figure 3.4 b) with a single unfolding force distribution centered around 12 pN and a refolding force distribution centered around 6.5 pN (Figure 3.5 c and d). The measured extension changes as a function of force were consistent with complete unfolding of the protein with a contour length change of 36.7 nm. Again, this indicated that the protein was unfolding from a compact state with an end-to-end extension between the cysteines of 3.8 nm. Using a force jump experiment, the distance to the transition state from the folded state was determined to be 1 nm +/- 1 nm, again similar to the other pulling axis.

3.3.3 Unfolding and refolding of apomyoglobin at pH 5 under force

After characterizing the molecule on the optical tweezers at pH 7, the buffer was changed to pH 5, populating the equilibrium molten globule state of the protein. At this pH, a force ramp experiment on the N/C variant showed a single unfolding and refolding force distribution centered on 4.5 pN, with no high force events observed (Figure 3.4 c). As the unfolding and folding transitions were reversible at around 4.5 pN with no hysteresis, a constant trap position experiment was preformed to determine the lifetimes as a function of the average force and the relative distances to the transition state (Figure 3.6 a).

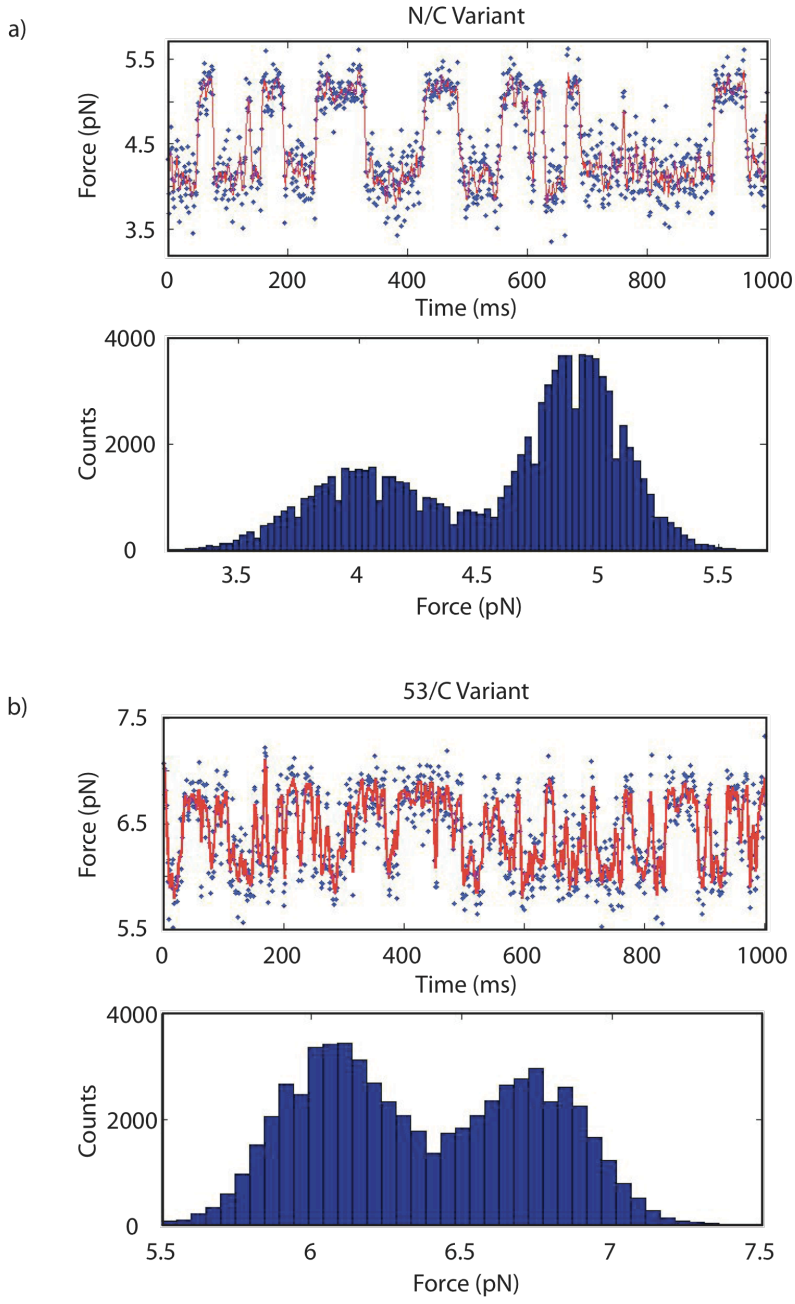

Figure 3.6 Sample traces of a constant trap position experiment for the N/C and 53/C variants.

One-second sample traces of the force averaged down to 1000 Hz (in blue) during a constant trap position experiment for the N/C variant (**a**) and the 53/C variant (**b**). The inferred trajectory of the molecule at 500 Hz is shown in red. The accompanying histogram is of the force measured over one minute depicting the two observed populations.

Individual states were identified and their lifetimes were determined using the Bayesian hidden Markov model (see methods). The natural log of the rate constants as a function of force were fit with a linear Bell's model yielding a distance to the transition state from the folded state of 6.1 ± 0.5 nm and 14.9 ± 1.5 nm for the distance to the transition state from the unfolded state (Figure 3.7 a). The sum of these distances was consistent with the measured extension change between the folded and unfolded states of 19 ± 1 nm under a constant force experiment at 4.5 pN, congruent with a two state model.

Working at pH 5, the 53/C variant showed a similar pattern of behavior compared to the N/C variant. In force ramp experiments, the molecule unfolded and refolded at around 6.5 pN with no hysteresis (Figure 3.4 d). Using a constant trap position experiment (Figure 3.6 b) and Bell's model, the distances to the transition state from the folded and unfolded states were determined be 3.4 +/- 1.2 nm and 7.6 +/- 3.3 nm, respectively (Figure 3.7 b). The sum of theses distances was consistent with the measured extension change of 12 +/- 1 nm measured by a constant-force experiment at 6.5 pN.

3.4 Discussion

The force ramp characterization of the N/C variant at pH 7 revealed a bimodal distribution of cooperative unfolding events, indicating the protein was unfolding from two different states, one more mechanically resistant to force than the other. Apomyoglobin folds through an intermediate in ensemble experiments with folding from the intermediate to the native state occurring on the order of a second. With short times spent at low force, the protein may not have had enough time to completely refold from the intermediate to the native state. Hence the subsequent unfolding event, the protein would unfold from this intermediate state. With longer waiting times at low force, the low force-unfolding event became less prevalent, which is consistent with the protein refolding to the presumably more mechanically resistant native state. This pattern of behavior was similar to the mechanical unfolding of RNAse H, which also showed a bimodal distribution, unfolding from both the native state and an intermediate state [14]. For RNase H, the native state and the intermediate state could be distinguished by their different extension changes. For the N/C variant, however, the changes in the contour length for both the high force and low force transitions were the same. Based on models of the native state and the molten globule state [20], both states are compact with the N- and C- termini helices contacting each other, and therefore similar extension changes would be expected. Competition from the native state at pH 7 precluded a more quantitative characterization of the pH 7 mechanical intermediate.

By waiting at low force for a long period time, the molecule was able to fully refold, allowing the mechanical characterization of the native state using a force jump experiment. From these experiments, the distance to the transition state from the folded state was determined to be small (~ 1 nm), similar to those measured for other natively folded proteins [4, 7, 11-19]. This small distance indicates that the native structure is brittle and that a small deformation in the structure along the defined reaction coordinate, the end-to-end extension of the molecule, results in the unfolding of the protein.

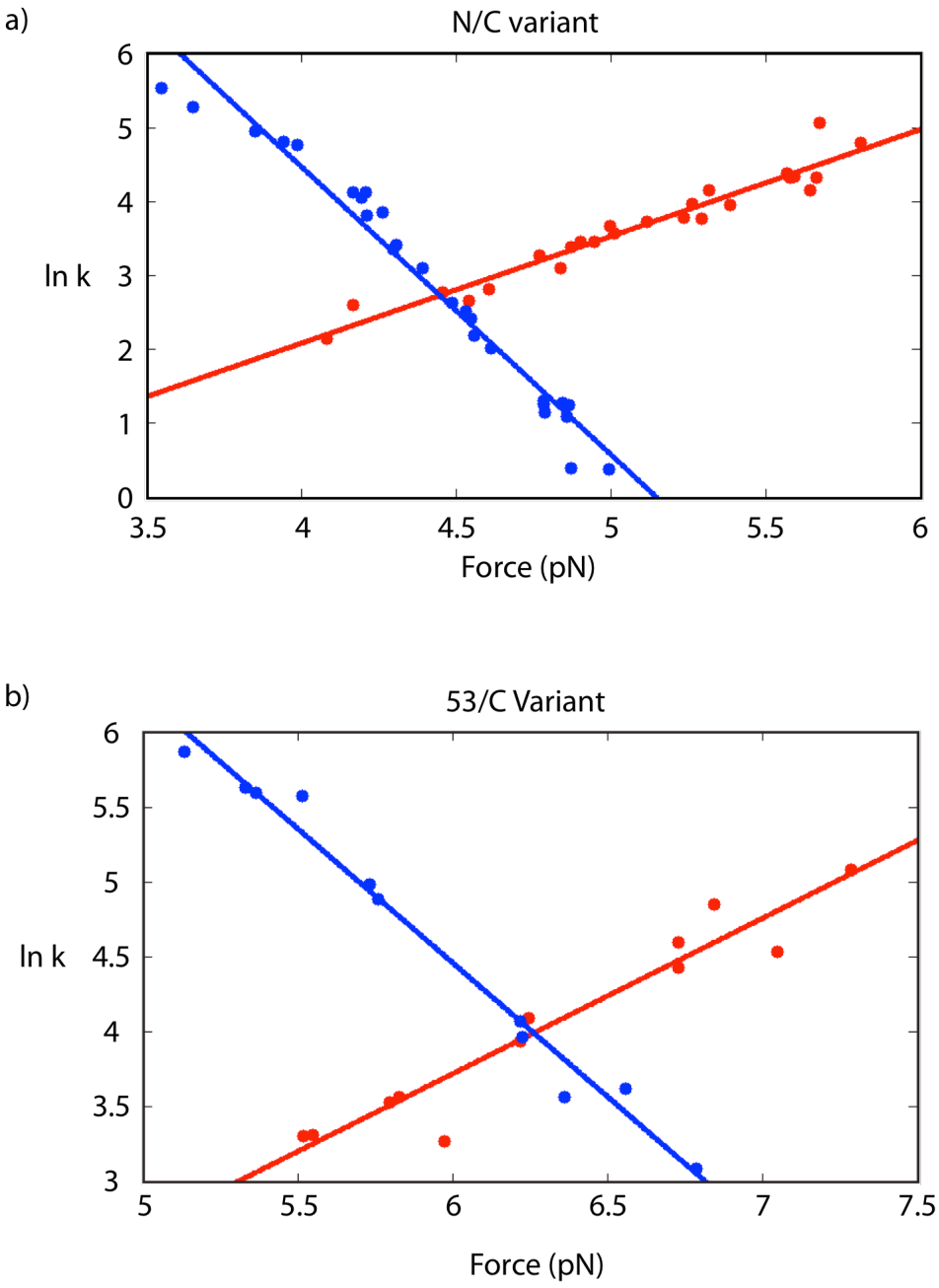

Figure 3.7 Linear fits of the natural log of the rate constants as a function of force.

The fits of the natural log of the rate constants as a function of force are shown for the N/C (**a**) and 53/C (**b**) variants. Using Bell's model, the distance to the transition state is proportional to the slope of the lines.

Characterization of the molten globule state at pH 5 revealed cooperative unfolding and folding at forces similar to the low force events observed at pH 7. The coincidence in behavior under force between the pH 7 intermediate and the molten globule state indicate that the molten globule is a good model for the mechanical intermediate observed at pH 7.

The constant trap experiments on the N/C variant measured a distance of 6.1 ± 0.5 nm to the transition state from the molten globule state, which was significantly larger and in sharp contrast to the native state, which had a distance to the transition state from the folded state of 1 ± 1 nm. These data indicate that the molten globule state is much more compliant and can undergo a larger deformation before reaching the transition state. This property is similar to the RNase H mechanical folding intermediate, which also has a large distance to the transition state (5 ± 1 nm), providing further evidence that the large distance to the transition state may be a general property of a molten globule.

The large distance to the transition state observed for the molten globule state may be specific to the chosen pulling axis. For single domain, natively folded globular proteins, a change in the pulling axis has been seen to result in anisotropic behavior with significant a change in the unfolding behavior [4, 13, 38]. For these proteins, however, there is little change in the distance to the transition state, which remains small, between 0.5 and 2 nm.

To investigate these issues, we pulled along a different axis using the 53/C variant. The molten globule of this variant also showed a large distance to the transition state (3.4 ± 1.2 nm) much larger then the distance to the transition state from the native state (1 ± 1 nm). While this distance was not as large as the distance measured for the N/C variant, the change in the pulling axis changed the reaction coordinate (i.e. the total end-to-end distance change of the molecule) and the 53/C variant end-to-end extension change is smaller (12 nm) than for the N-terminal construct (20 nm).

Comparison of the different pulling axes is difficult because of the different reaction coordinates. A change in the pulling axis may be forcing the protein over very different transition state barriers, exploring different regions of the energy landscape. One approach to comparing different reaction coordinates or analyzing ill-defined reaction coordinates has been to normalize the distance to the transition state by the total distance along the chosen reaction coordinate (Table 3.1). This is analogous to calculating a Tanford β value [39]. Calculating such a normalized distance along the reaction coordinate (i.e. the ratio of the distance to the transition state to the total end-to-end distance change) for natively folded globular proteins typically produces a value between 0.05 and 0.1. For the molten globule state, we determined the relative position of the transition state to be 0.3 for both pulling axes. This indicates that the large distance to the transition state for unfolding of the molten globule state is independent of the force axis.

Protein Pulling Axis	$\Delta x^{\ddagger}_{(F-\ddagger)}$(nm)	$\Delta x^{\ddagger}_{(U-\ddagger)}$ (nm)	$\Delta x_{(F-U)}$ (nm)	$\Delta x_{(F-U)}$(Measured) (nm)	Relative Position of the Transition State
N/C Variant	6.1 +/- 0.5	14.9 +/- 1.5	21.0 +/- 1.6	19 +/- 1	0.29
53/C Variant	3.4 +/- 1.2	7.6 +/- 3.3	11.0 +/- 3.5	12 +/- 1	0.31

Table 3.1 Summary of the distances to the transition state and the normalized position of the transition state.

3.5 Conclusions

This work, combined with the previous work on RNase H, suggests that the large distance to the transition state is a general property of molten globule states. In addition, the choice of pulling axis does not significantly alter this property or the mechanical behavior of the protein. This suggests that the molten globule state is more isotropic than natively folded globular proteins, which while remaining brittle have more significant changes in their behavior with a change in the pulling axis. An important consequence of the large distance to the transition state for the molten globule state is that it is much easier to unfold under force compared to natively folded protein with a small distance to the transition state. One could speculate that this property may play an important role in biology, in particular, for proteins that are required to unfold under force.

3.5 References

1. Kuwajima, K., *The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure*, in *Proteins*. 1989. p. 87-103.
2. Dobson, C.M., *Protein folding. Solid evidence for molten globules*, in *Curr Biol*. 1994. p. 636-40.
3. Fowler, S.B., et al., *Mechanical unfolding of a titin Ig domain: structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering*, in *Journal of Molecular Biology*. 2002. p. 841-9.
4. Dietz, H., et al., *Anisotropic deformation response of single protein molecules*, in *Proc Natl Acad Sci USA*. 2006. p. 12724-8.
5. Dietz, H. and M. Rief, *Exploring the energy landscape of GFP by single-molecule mechanical experiments*, in *Proc Natl Acad Sci USA*. 2004. p. 16192-7.
6. Marszalek, P.E., et al., *Mechanical unfolding intermediates in titin modules*, in *Nature*. 1999. p. 100-3.
7. Williams, P.M., et al., *Hidden complexity in the mechanical properties of titin*, in *Nature*. 2003. p. 446-9.
8. Schwaiger, I., et al., *A mechanical unfolding intermediate in an actin-crosslinking protein*, in *Nat Struct Mol Biol*. 2004. p. 81-5.
9. Li, L., et al., *Mechanical unfolding intermediates observed by single-molecule force spectroscopy in a fibronectin type III module*, in *Journal of Molecular Biology*. 2005. p. 817-26.
10. Li, L., et al., *Stepwise unfolding of ankyrin repeats in a single protein revealed by atomic force microscopy*, in *Biophysical Journal*. 2006. p. L30-2.
11. Shank, E.A., et al., *The folding cooperativity of a protein is controlled by its chain topology*, in *Nature*. 2010. p. 637-40.
12. Brockwell, D.J., et al., *Mechanically unfolding the small, topologically simple protein L*, in *Biophysical Journal*. 2005. p. 506-19.
13. Brockwell, D.J., et al., *Pulling geometry defines the mechanical resistance of a beta-sheet protein*, in *Nat Struct Biol*. 2003. p. 731-7.
14. Cecconi, C., et al., *Direct observation of the three-state folding of a single protein molecule*, in *Science*. 2005. p. 2057-60.
15. Li, H., et al., *Atomic force microscopy reveals the mechanical design of a modular protein*, in *Proc Natl Acad Sci USA*. 2000. p. 6527-31.
16. Liu, R., et al., *Mechanical characterization of protein L in the low-force regime by electromagnetic tweezers/evanescent nanometry*, in *Biophysical Journal*. 2009. p. 3810-21.
17. Oberhauser, A.F., et al., *The molecular elasticity of the extracellular matrix protein tenascin*, in *Nature*. 1998. p. 181-5.
18. Rief, M., et al., *Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles*, in *J Mol Biol*. 1999. p. 553-61.
19. Schlierf, M., H. Li, and J.M. Fernandez, *The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques*, in *Proc Natl Acad Sci USA*. 2004. p. 7299-304.

20. Hughson, F.M., P.E. Wright, and R.L. Baldwin, *Structural characterization of a partly folded apomyoglobin intermediate*, in *Science*. 1990. p. 1544-8.
21. Jennings, P.A. and P.E. Wright, *Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin*, in *Science*. 1993. p. 892-6.
22. Barrick, D. and R.L. Baldwin, *Stein and Moore Award address. The molten globule intermediate of apomyoglobin and the process of protein folding*, in *Protein Sci*. 1993. p. 869-76.
23. Eliezer, D., et al., *The radius of gyration of an apomyoglobin folding intermediate*, in *Science*. 1995. p. 487-8.
24. Eliezer, D., et al., *Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding*, in *Nat Struct Biol*. 1998. p. 148-55.
25. Hughson, F.M., D. Barrick, and R.L. Baldwin, *Probing the stability of a partly folded apomyoglobin intermediate by site-directed mutagenesis*, in *Biochemistry*. 1991. p. 4113-8.
26. Jamin, M. and R.L. Baldwin, *Refolding and unfolding kinetics of the equilibrium folding intermediate of apomyoglobin*, in *Nat Struct Biol*. 1996. p. 613-8.
27. Nishii, I., M. Kataoka, and Y. Goto, *Thermodynamic stability of the molten globule states of apomyoglobin*, in *Journal of Molecular Biology*. 1995. p. 223-38.
28. Nishimura, C., H.J. Dyson, and P.E. Wright, *The apomyoglobin folding pathway revisited: structural heterogeneity in the kinetic burst phase intermediate*, in *Journal of Molecular Biology*. 2002. p. 483-9.
29. Cecconi, C., et al., *Protein-DNA chimeras for single molecule mechanical folding studies with the optical tweezers*, in *Eur Biophys J*. 2008. p. 729-38.
30. Bustamante, C.J. and S.B. Smith, *Light-force sensor and method for measring axial optical-trap forces from changes in light momentum along an optical axis*. 2006: USA. p. 1-20.
31. Carter, B.L., *Antihypertensive drug interactions*, in *Timely Top Med Cardiovasc Dis*. 2005. p. E2.
32. Chodera, J.D., et al., *Bayesian hidden Markov model analysis of single-molecule biophysical experiments: Characterizing metastable states and transition rates under measurement uncertainty.* In preparation, 2010.
33. Baum, L.E., *A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains.* Ann. Math. Statist., 1970. **41**: p. 164-171.
34. Bell, G.I., *Models for the specific adhesion of cells to cells*, in *Science*. 1978. p. 618-27.
35. Greenleaf, W.J., et al., *Passive all-optical force clamp for high-resolution laser trapping*, in *Phys. Rev. Lett.* 2005. p. 208102.
36. Barrick, D., F.M. Hughson, and R.L. Baldwin, *Molecular mechanisms of acid denaturation. The role of histidine residues in the partial unfolding of apomyoglobin*, in *Journal of Molecular Biology*. 1994. p. 588-601.
37. Stigter, D. and C. Bustamante, *Theory for the hydrodynamic and electrophoretic stretch of tethered B-DNA*, in *Biophysical Journal*. 1998. p. 1197-210.
38. Carrion-Vazquez, M., et al., *The mechanical stability of ubiquitin is linkage dependent*, in *Nat Struct Biol*. 2003. p. 738-43.

39. Fersht, A.R., *Structure and Mechanism in Protein Sci- ence: A Guide to Enzyme Catalysis and Protein folding* 1998: New York: W.H. Freeman and Company.

Chapter 4 Exploring the affect of pulling axis on the transition state of srcSH3

4.1 Introduction

A primary goal of force spectroscopy experiments exploring protein folding and unfolding has been to elucidate the mechanism by which proteins resist force. As stated earlier in this thesis, Bell's model describes the affect of a vectorial force on a transition state given the barrier height and distance to the transition state along a one dimensional potential energy surface. This model however does not provide any mechanistic insight, as the model does not contain any structural information about the rate limiting transition state. Further, there is little information known about the relationship between the pulling axis (i.e. the reaction coordinate) and the structure of the transition state.

Previous work in force spectroscopy exploring the effect of pulling axis has demonstrated that proteins have different mechanical properties (e.g. average unfolding force at a specific loading rate) when the force is applied along different axes. Brockwell et. al. pulled on E2lip3 along two different axes and demonstrated that the two variants unfold at different average forces[1]. Similarly, Carrion-Vazquez et. al pulled on ubiquitin from different attachment points and showed a difference in the unfolding force distribution[2]. More recently, Dietz et. al. unfolded GFP under force exploring the behavior along five different pulling axes[3]. There was significant variation between the average unfolding forces and distances to the transition state between the variants. In addition, one of the variants did not unfold in a single cooperative step but populated an intermediate during the unfolding process, traversing two transition state barriers. This work clearly demonstrated that the application of force along different axes changes the behavior of the protein under force. This work however does not provide any insight into the mechanisms of unfolding or the structure of the transition state.

One potential explanation of how a vectorial force affects the transition state of a protein has been supported by work done on protein translocases, such CLpXP and the mitochondrial import complex[4-6]. In the case of these translocases, in order for a protein to be transported across the narrow pore, it must first be unfolded[6, 7]. The proposed hypothesis suggests that the application of force by these protein complexes to unfold the substrate polarizes the transition state, shifting the transition state to the localized structure in the vicinity of the applied force.

The clearest support for this hypothesis comes from the work on the import mechanism of the mitochondrial import machinery using barnase as a model substrate[5]. The hypothesis was tested by comparing the structures of the transition state for unfolding by chemical denaturation and by force. The structure of the transition state of the ribonuclease barnase in the absence of force had been mapped out previously using a mutational analysis known as phi-value analysis[8]. In this method, single amino acid mutations are made and the affect on the thermodynamic stability and kinetics of folding and unfolding are measured. The phi-value (in the folding direction) is defined as the ratio between the change in the transition state free energy ($\Delta\Delta G_{\ddagger\text{-}U}$) between the mutant and wild type and the change in the free energy ($\Delta\Delta G_{F\text{-}U}$) between the mutant and wild type.

$$\phi = \Delta\Delta G_{\ddagger\text{-}U}/\Delta\Delta G_{F\text{-}U} \quad (4.1)$$

Typically, the values range between zero and one. A phi-value close to zero indicates that the site of the mutated amino acid is not structured in the transition state (shows the same energetic contribution as the unfolded state) where as a phi-value close to one indicates that the site of the mutated amino acid is structured in the transition state (shows the same energetic contribution as the folded or native state).

Using barnase as the substrate for the mitochondrial import machinery, the effect of mutations in barnase on the acceleration of the import rate was determined and used to infer the structure of the transition state. Mutations in the N-terminal had the largest affect on the import process[5]. This was in contrast to the structure of the transition state under chemical denaturation. These results suggested that the transition state shifted toward the N-terminus where the degradation tag was attached and the primary site of interaction with the import machinery. This hypothesis is perhaps not that surprising because to unfold the protein, the unfoldase must first disrupt the local structure close to the site of the applied force in order to unfold the rest of the protein.

In order to better understand the relationship between the affect of an applied force on the transition state of a protein, I developed a protein system to analyze with the optical trap to characterize the mechanical behavior and the transition state under different vectorial forces. An important consideration when comparing work from force spectroscopy and unfoldases is that the geometry and force loading rates may be very different and the affects of the applied force and the mechanisms of unfolding may be very different. This most obvious difference is that for unfolding in an optical trap, the protein is tethered at two points and a tension is applied across the molecule. While in the unfoldases, the force is applied to the protein by pulling the protein from one point of attachment through a small pore. Conclusions from force spectroscopy may be more relevant to proteins that have a force applied across the molecule and resist force in their biological function, such as Ig27 in muscle[9].

The SH3 domain from chicken c-Src domain was chosen as a model system, hereafter referred to as srcSH3 (shown in Figure 4.1). SH3 has several inherent properties that make it a good model system for these studies. First, it has been extensively characterized in ensemble experiments and, at least by optical probes such as CD and fluorescence, it has been shown to fold and unfold in a simple two-state manner. The structure of the transition state under chemical denaturation has been determined using a detailed phi-value analysis [10-13]. The protein is very small (65 amino acids) and is therefore accessible to simulation, which could be used in the future to help verify or interpret any experimental results.

The structure of srcSh3 has some additional features making it a good model for testing the hypothesis that the transition state shift towards a localized force-bearing region. The transition state for srcSH3 in the absence of force is centered on the distal loop and is separated from the N- and C-terminal anti-parallel β-sheet (highlighted in Figure 4.1a). Under the current hypothesis, if force were applied to this β-sheet, the

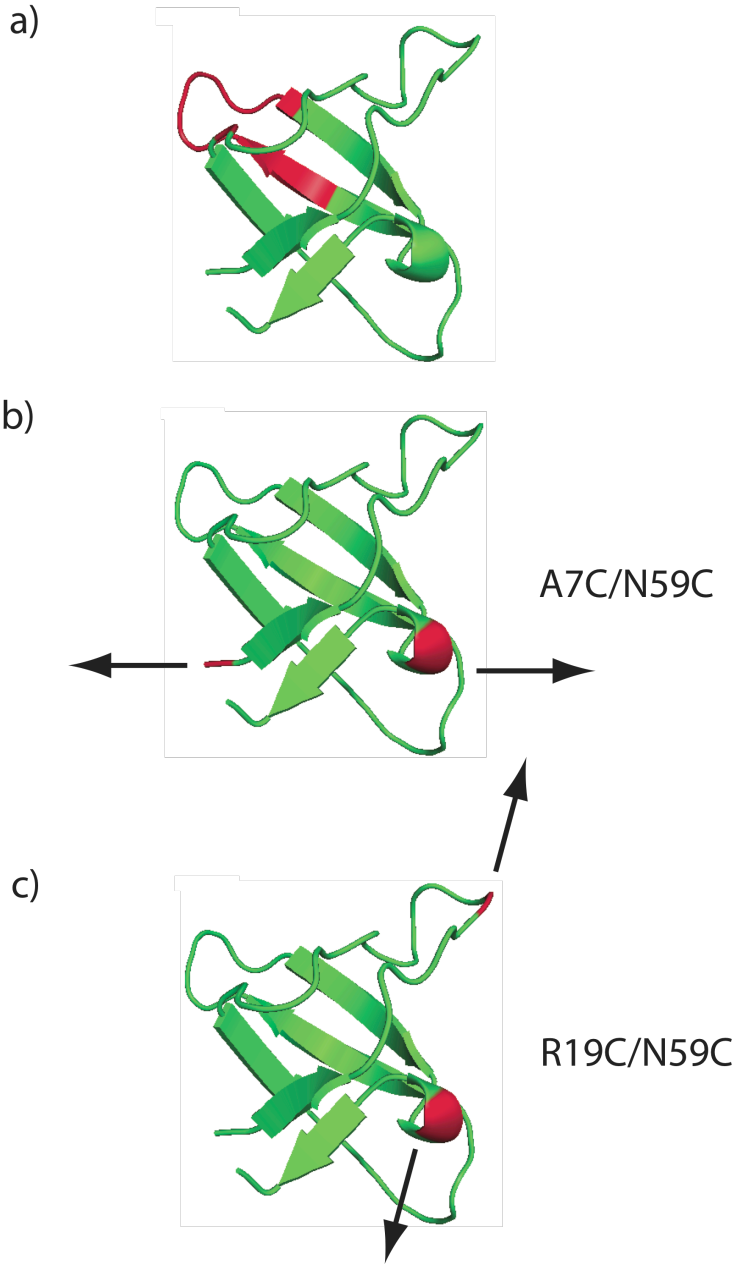

Figure 4.1 – Structure of srcSH3.

a, shows the structure of srcSH3 with the high phi-value (greater than 0.5) residues' backbone highlighted in red. **b** and **c** depict the A7C/N59C srcSH3 variant and the R19C/N59C srcSH3 variant, respectively, with the pulling axis denoted by the black arrows.

transition state would shift to residues in the N- and C-terminal anti-parallel β-sheet. This could potentially allow the differentiation between different transition states in the presence and absence of force if indeed there is a change.

If the transition state does indeed shift towards the N- and C- terminal anti-parallel β-sheet, this system would allow us to further test if and how the axis of the applied force changes the transition state. In this case, a shearing force (a force parallel to the β-sheet) or an unzipping force (a force perpendicular to the β-sheet) would be applied to the system, allowing for the behavior and transition states to be compared and contrasted. The initial hypothesis is that the protein will be more mechanically resistant to the shearing force than the unzipping force.

In this section, I describe the work I have done to develop the model system and report the initial results from force-ramp experiments. As this project is still ongoing, I will also conclude with a brief description of the future work that will be done on the system.

4.2 Methods and materials

4.2.1 Protein purification and handle attachment

In order to test this hypothesis, two srcSH3 variants were engineered with a pair of cysteines engineered at different sites to enable the attachment of the dsDNA handles. Using the Stratagene Quickchange protocol, cysteines were introduced to produce A7C/N59C srcSH3, the shearing force axis, and R19C/N59C srcSH3, the unzipping force axis (Figure 4.1 b and c). The sites of the introduced cysteines were chosen to have a minimal affect on the thermodynamic stability and kinetics of the protein as shown by previous phi-value analysis. The proteins were over-expressed with an N-terminal 6-his tag and TEV protease cleavage site using a T7 promoter system in *Escherichia coli* BL21(DE3) pLysS cells, and purified as previously published except all buffers contained 0.5 mM *tris*(2-carboxyethyl)phosphine (TCEP) to keep the cysteines reduced and prevent any disulfide bond formation[14]. Handles consisting of functionalized 558 bp dsDNA were then produced by PCR and attached to the protein as previously published [15-17].

4.2.2 Thermodynamic stability measurements

To determine if the introduced cyteines affected the thermodynamic stability of the protein, the intrinsic fluorescence of the variants were determined via equilibrium denaturation with guanidinium hydrochloride (GmdCl). The protein (1 μM) was equilibrated at 25 ºC overnight in 10 mM tris(hydroxymethyl)aminomethane (Tris) at pH 7.0, 250 mM sodium chloride, 1 mM ethylenediaminetetraacetic acid (EDTA), and 1mM TCEP with varying concentrations of GmdCl (0 to 5 M). The fluorescence was then measured between 300 and 400 nm with a FluoroMax-3. The center of mass for the fluorescence signal at each denaturant concentration was then calculated by the following equation:

$$\Sigma(\lambda * I)/\Sigma(\lambda) \qquad (4.2)$$

where λ is the wavelength of the light in nm and I is the intensity of the light in arbitrary absorbance units. The center of mass fluorescence as a function of denaturant concentration was fit using a two-state model (Figure 4.2)[18]. Each measurement was performed 4 times. From these data, the average C_m and average m-value were and used to calculate the free energy between the unfolded and folded state.

4.2.3 Force spectroscopy experiments

The srcSH3 tethers were attached to two beads, one held on a pipette tip and the other manipulated in an optical trap as previously described (Figure 1.6). Force ramp experiments were preformed at a constant pulling speed of 100 nm/sec on a dual beam counter-propagating optical trap with a spring constant of 0.1 pN/nm collected at 1000 Hz [19, 20]. The force was cycled between 2 and 40 pN for the A7C/N59C srcSH3 and 2 and 20 pN for the R19C/N59C srcSH3 variant. The force cycle ensured that an unfolding and refolding event were observed during each cycle.

Unfolding and refolding forces were detected by analyzing the smoothed force data and detecting either an increase in the force for the refolding event or a decrease in the force for unfolding event. The data was smoothed with a sliding window of 10 ms and the difference the signal at intervals of 10 ms was determined. The standard deviation was then calculated and a threshold for detection was set based off the variance. A 99.9 % confidence threshold (3.3 σ) was set for the high signal-to-noise unfolding events and a 90% confidence threshold (1.6 σ) was used for the noisier refolding events. In the case of the identification of multiple events above the threshold for the refolding events, the event with the highest signal was selected as the transition as only one refolding occurred in the low force regime. From this experiment, the unfolding or folding force distribution as a function of the loading rate and the extension change of the molecule could be measured.

4.2.4 Equilibrium free energy determination from force ramp experiments

Using Crooks Fluctuation Theorem [21] on the measured unfolding and refolding force distributions, the free energy of the protein could be determined as described in previous publications [17]. Briefly, the area under the force-trap position curves were integrated around the transition with the force bounds determined by the unfolding or refolding force. The work values were then corrected for the energy of the stretching of the protein polymer by integrating a worm-like chain model with parameters defined by the distance between the cysteines and the force of unfolding or refolding. In this case, the contribution of the stretching of the dsDNA handles was ignored because the stretching is an equilibrium transition and the work is measured between equal forces and therefore the dsDNA does not undergo any net conformational change [17].

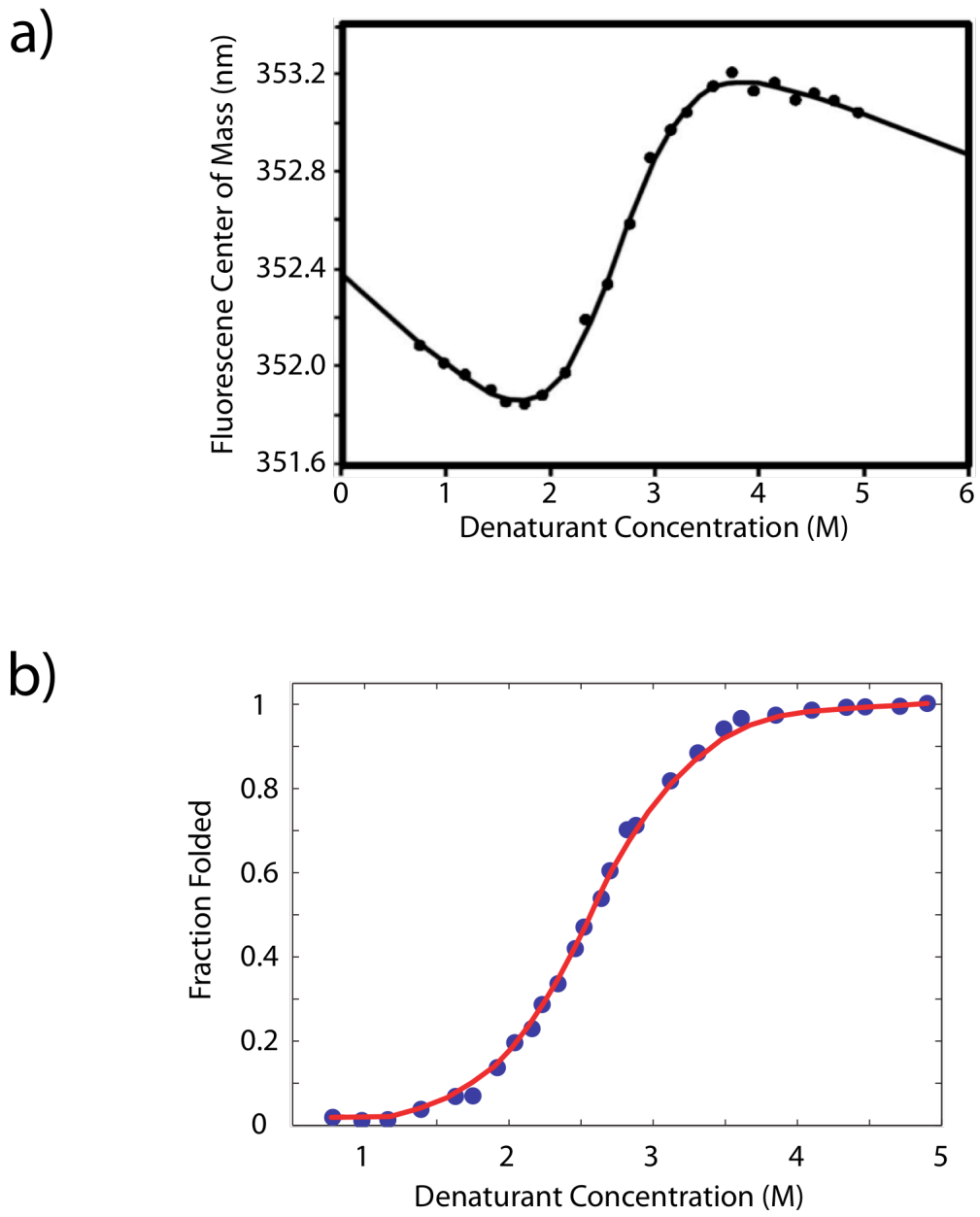

Figure 4.2 Equilibrium denaturation of srcSH3.

a, depicts an equilibrium chemical denaturant melt following the center of mass of the fluorescence as a function of denaturation concentration of the A7C/N59C srcSH3 variant as an example. **b,** shows the fraction folded of the molecule assuming a two state model as a function of the denaturation concentration.

4.3 Results

4.3.1 Mechanical properties of A7C/N59C and R19C/N59C srcSH3

Force ramp experiments were preformed on both variants of srcSH3 in the optical trap at a constant trap speed of 100 nm/sec. Example force-extension traces for each variant are shown in Figure 4.3. In each pulling or relaxation phase, the protein was observed to unfold or refold in a single cooperative step. The extension change for each event was consistent with the expected distance change which was was calculated using the following equation:

$$\Delta x(\text{expected}) = x_{\text{Unfolded}}(\text{wlc}) - x_{\text{Folded}} \tag{4.3}$$

where $x_{\text{Unfolded}}(\text{wlc})$ is the end-to-end extension of the unfolded state as modeled by the worm-like chain model given the force, contour length, persistent length, and temperature [22], x_{Folded} is the end-to-end distance of the folded state as determined by measuring the distance between amino acids' β-carbons in the crystallographic model of the protein (PDB ID: 1SRM). This model assumes that the end-to-end distance of the native state is not deformed under force. Expressed as the contour length, the measured contour length upon unfolding or folding was 18.5 ± 3.1 nm for the A7C/N59C srcSH3 variant and was consistent with the calculated distance change of 18.7 nm. For the R19C/N59C srcSH3 variant, the measured distance change was 16.0 ± 3.7 nm, which was again consistent with calculated distance change of 16.5 nm.

Histograms of the unfolding and refolding force distribution of each variant are shown in Figure 4.4. The average unfolding forces varied greatly between the variants with the shearing construct (A7C/N59C) unfolding at 27.8 ± 6.6 pN and the unzipping construct (R19C/N59C) unfolding at 10.2 ± 3.7 pN. The refolding force distributions did not vary significantly for the shearing and unzipping variants and were 4.1 ± 0.8 pN and 3.9 ± 0.9 pN, respectively.

The free energy of each variant was calculated using the Crooks' Fluctuation Theorem. The free energy for the A7C/N59C variant was 5.0 ± 0.6 kcal/mol and the free energy for the R19C/N59C variant was 3.3 ± 0.4 kcal/mol.

4.3.2 Free energy determination by equilibrium chemical denaturation:

The free energy as determined by 4 independent experiments for the A7C/N59C srcSH3 variant was 3.9 ± 0.5 kcal/mol. The R19C/N59C srcSH3 variant had a stability of 3.8 ± 0.5 kcal/mol. The reported stability of the wild type srcSH3 was 3.7 ± 0.1 kcal/mol [10].

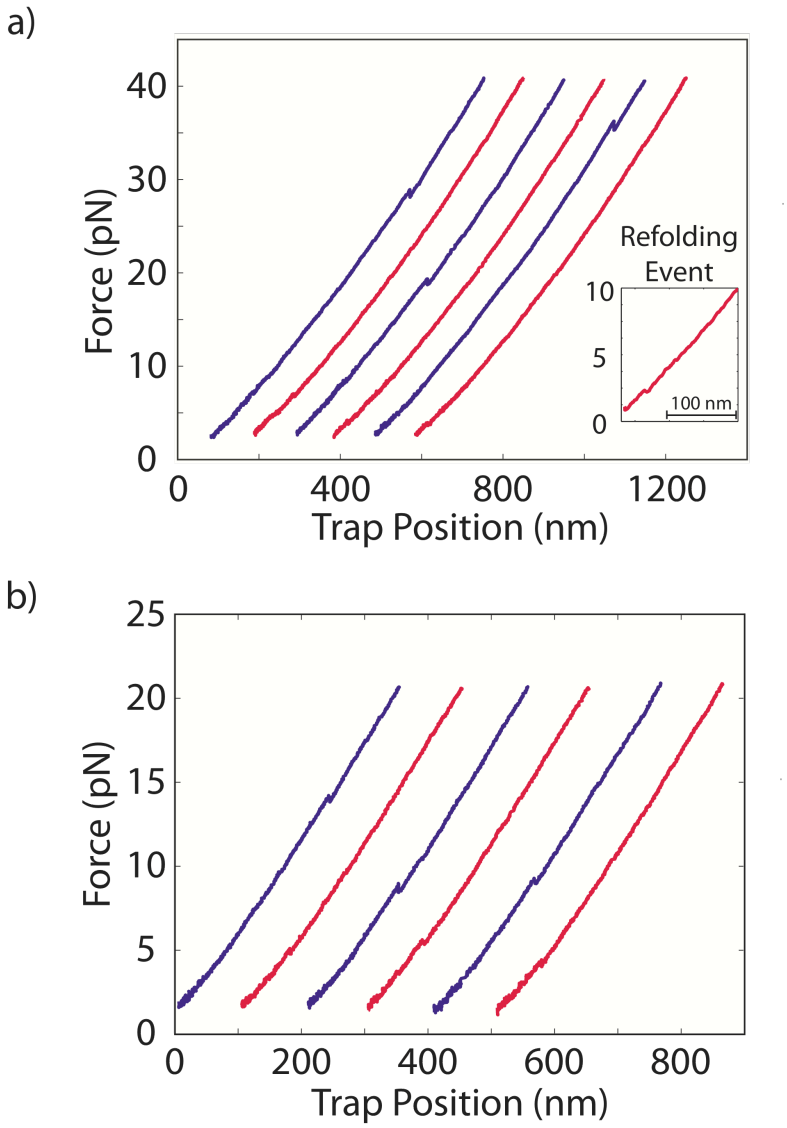

Figure 4.3 Force ramp data on the srcSH3 variants.

a and **b** depict increasing (blue) and decreasing (red) force ramp experiments on the A7C/N59C and the R19C/N59C srcSH3 variants, respectively. The refolding events are hard to detect on this scale because of the low signal-to-noise ratio at low forces. The inset in **a** depicts a close up of the last refolding event.

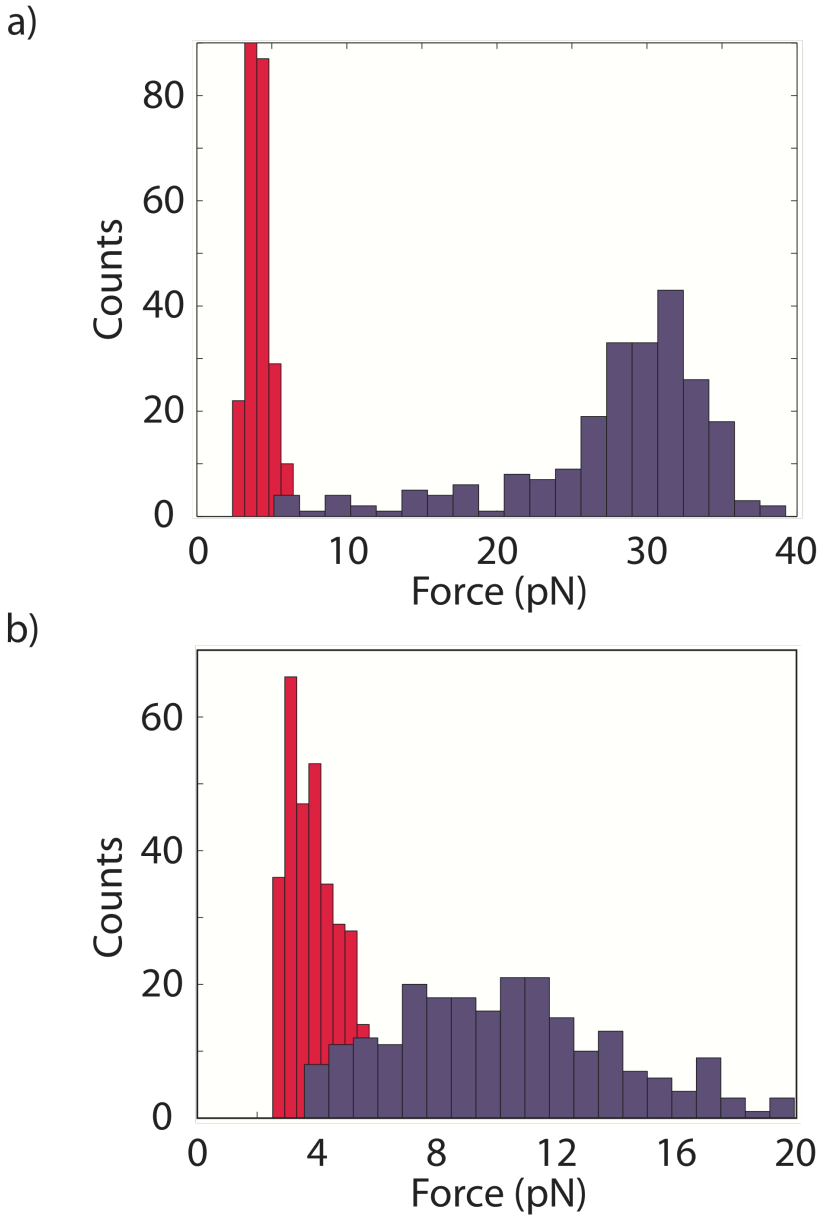

Figure 4.4 Histograms of the refolding and unfolding forces for the srcSH3 variants.

a and **b** present histograms of the unfolding (blue) and refolding (red) forces for the A7C/N59C and the R19C/N59C srcSH3 variants, respectively.

4.4 Discussion:

Clearly, a change in the pulling axis radically changes the mechanical behavior of the protein, consistent with previous studies. There are several potential explanations for the differing mechanical behavior of the srcSH3 variants studied. The first hypothesis is that the cysteine modification and handle attachment altered the thermodynamic stability of the proteins and that this is the root cause of the change in behavior. If the thermodynamic stability of the protein is unaffected, a second potential explanation is that the structure and free energy of the transition state barrier remains unchanged and the difference in behavior arises from the change in reaction coordinate with different distances to the transition state barrier. Another possibility is that the transition states are different between chemical denaturation and force, but the transition state under force for both pulling axes are similar, and the differences in behavior arises from different distances to the transition state for each pulling coordinate. Lastly, all three different methods of denaturation could traverse different regions of the potential energy landscape and proceed through different transition state barriers with different distances to the transition state. The goal of this project is to differentiate between these different hypotheses.

Using the equilibrium chemical denaturation experiements and the Crooks Fluctuation Theorem analysis of the force ramp data, the free energy of the protein variants was determined and used to measure the affect of the mutations and handle attachment on the thermodynamic stability of the protein. The free energies of the A7C/N59C srcSH3 and the R19C/N59C srcSH3 were determined by both methods to have similar stabilities within error of each other (Table 1). The errors in the free energies determined by the Crooks Flucuation Theorem are from the fits and does not account for errors in the measured dissipated work. These results suggest that the significant variation in the mechanical behavior between the variants is not a result of a change in the thermodynamic stability of the proteins because of the introduced mutations or the attachment of the dsDNA handles.

The refolding force distributions were equivalent, occurring around 4 pN for both variants. Refolding force distributions for several different proteins at similar loading rates of 10 pN/sec have been shown to all occur at low forces around 5 pN [15, 17, 23]. This is likely because despite variations in the barrier heights for the different proteins, the large distance to the transitions state from the unfolded state dominates the refolding rate as a function of the force. In other words, given a large distance to the transition state, small variations in the average refolding forces account for any potential variation in the barrier heights.

The mechanical unfolding force distributions for each variant were significantly different with a 17.6 pN difference between the average unfolding forces. This suggests either the protein is traversing different transition state barriers or, on the other extreme, that the difference in behavior can be accounted for solely by different distances to the transition state. If the distances to the transition state from the folded state for both variants are similar to those observed for other natively folded proteins [1-3, 15, 17, 24-32], then they are expected to be small between 0.5 nm and 2 nm. Given the range of

	Equilibrium Denaturation		Crooks Analysis
Protein	Free Energy (kcal/mol)	m-value (kcal/mol·M)	Free Energy (kcal/mol)
srcSH3 wt	3.7 ± 0.1[1]	1.6 ± 0.1[1]	---
A7C/N59C	3.9 ± 0.5	1.5 ± 0.2	5.0 ± 0.6
R19C/N59C	3.8 ± 0.5	1.5 ± 0.2	3.3 ± 0.4

Table 4.1 Summary of free energy determination from equilibrium chemical denaturation and force ramp experiments.

these expected distances, the data strongly suggests a change in the transition state barrier.

The distance to the transition state can be estimated from the force ramp experiments; however, these methods require fitting multiple parameters (k_m, k_0, $\Delta x^{\ddagger}$) to match the unfolding force distribution [33-35]. Therefore, these methods contain significant uncertainty in the reported parameters, specifically, the quantities of interest, the distances to the transition state. An alternative method is to measure the average lifetime as a function of an average force using a constant force jump experiment. This approach requires no assumptions about the pre-exponential or free energy of the transition state.

4.5 Future directions

To resolve the relationship between the transition state and an applied vectorial force for protein folding, phi-value and force jump experiments will be required. This work is currently ongoing and is being continued by Dr. Bharat Jagannathan.

Force jump experiments are currently being performed on the A7C/N59C construct and R19C/N59C construct. Preliminary analysis suggests that the distances to the transition state do not account for the change in the unfolding distributions suggesting a change in the transition state free energy.

Because of the intensive nature of these single molecule experiments, phi-value analysis has been initially targeted to a few residues. These residues will hopefully enable the different hypotheses to be distinguished. If the transition state becomes localized towards the terminal β-sheet, then residues, which have a high phi-value under chemical denaturation, may have a lower phi-value under force denaturation. Potentially, this will demonstrate the change in the transition state structure under force. Targeting residues in the terminal β-sheet, such as F10I and V61A will conceivably allow us to differentiate between the transition state structures of the shearing and unzipping force axes. An important consideration is that the mutational analysis probes side chain interactions and not the hydrogen bonding network in the β-sheet, and so may only be probing one contributing aspect of the transition state. Initial data from the phi-value analysis of S47A indicates a movement of the transition state upon mutation with the distance to the transition state increasing. Another avenue of potential work is simulation work by the Pande group. In combination with the phi-value analysis, the simulations could potentially lead to a clearer explanation of the structure of the transition state under force or suggest future experiments, such as targeting specific sites for phi-value analysis.

4.6 References

1. Brockwell, D.J., et al., *Pulling geometry defines the mechanical resistance of a beta-sheet protein*, in *Nat Struct Biol*. 2003. p. 731-7.
2. Carrion-Vazquez, M., et al., *The mechanical stability of ubiquitin is linkage dependent*, in *Nat Struct Biol*. 2003. p. 738-43.
3. Dietz, H., et al., *Anisotropic deformation response of single protein molecules*, in *Proc Natl Acad Sci USA*. 2006. p. 12724-8.
4. Kenniston, J.A., et al., *Effects of local protein stability and the geometric position of the substrate degradation tag on the efficiency of ClpXP denaturation and degradation*, in *J Struct Biol*. 2004. p. 130-40.
5. Huang, S., et al., *Mitochondria unfold precursor proteins by unraveling them from their N-termini*, in *Nat Struct Biol*. 1999. p. 1132-8.
6. Shin, Y., et al., *Single-molecule denaturation and degradation of proteins by the AAA+ ClpXP protease*, in *Proc Natl Acad Sci USA*. 2009. p. 19340-5.
7. Matouschek, A., et al., *Active unfolding of precursor proteins during mitochondrial protein import*, in *EMBO J*. 1997. p. 6727-36.
8. Matouschek, A., et al., *Mapping the transition state and pathway of protein folding by protein engineering*, in *Nature*. 1989. p. 122-6.
9. Erickson, H.P., *Reversible unfolding of fibronectin type III and immunoglobulin domains provides the structural basis for stretch and elasticity of titin and fibronectin*, in *Proc Natl Acad Sci USA*. 1994. p. 10114-8.
10. Grantcharova, V.P. and D. Baker, *Folding dynamics of the src SH3 domain*, in *Biochemistry*. 1997. p. 15685-92.
11. Riddle, D.S., et al., *Experiment and theory highlight role of native state topology in SH3 folding*, in *Nat Struct Biol*. 1999. p. 1016-24.
12. Grantcharova, V.P., et al., *Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain*, in *Nat Struct Biol*. 1998. p. 714-20.
13. Grantcharova, V.P., D.S. Riddle, and D. Baker, *Long-range order in the src SH3 folding transition state*, in *Proc Natl Acad Sci USA*. 2000. p. 7084-9.
14. Wildes, D. and S. Marqusee, *Hydrogen exchange and ligand binding: ligand-dependent and ligand-independent protection in the Src SH3 domain*, in *Protein Sci*. 2005. p. 81-8.
15. Cecconi, C., et al., *Direct observation of the three-state folding of a single protein molecule*, in *Science*. 2005. p. 2057-60.
16. Cecconi, C., et al., *Protein-DNA chimeras for single molecule mechanical folding studies with the optical tweezers*, in *Eur Biophys J*. 2008. p. 729-38.
17. Shank, E.A., et al., *The folding cooperativity of a protein is controlled by its chain topology*, in *Nature*. 2010. p. 637-40.
18. Pace, C.N. and K.L. Shaw, *Linear extrapolation method of analyzing solvent denaturation curves*, in *Proteins*. 2000. p. 1-7.
19. Smith, S.B., Y. Cui, and C. Bustamante, *Optical-trap force transducer that operates by direct measurement of light momentum*, in *Meth Enzymol*. 2003. p. 134-62.

20. Bustamante, C.J. and S.B. Smith, *Light-force sensor and method for measring axial optical-trap forces from changes in light momentum along an optical axis*. 2006: USA. p. 1-20.
21. Crooks, G.E., *Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences*, in *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*. 1999. p. 2721-6.
22. Bustamante, C., et al., *Entropic elasticity of lambda-phage DNA*, in *Science*. 1994. p. 1599-600.
23. Schlierf, M., F. Berkemeier, and M. Rief, *Direct observation of active protein folding using lock-in force spectroscopy*, in *Biophysical Journal*. 2007. p. 3989-98.
24. Oberhauser, A.F., et al., *The molecular elasticity of the extracellular matrix protein tenascin*, in *Nature*. 1998. p. 181-5.
25. Rief, M., et al., *Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles*, in *J Mol Biol*. 1999. p. 553-61.
26. Li, H., et al., *Atomic force microscopy reveals the mechanical design of a modular protein*, in *Proc Natl Acad Sci USA*. 2000. p. 6527-31.
27. Williams, P.M., et al., *Hidden complexity in the mechanical properties of titin*, in *Nature*. 2003. p. 446-9.
28. Schlierf, M., H. Li, and J.M. Fernandez, *The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques*, in *Proc Natl Acad Sci USA*. 2004. p. 7299-304.
29. Ainavarapu, S.R.K., et al., *Ligand binding modulates the mechanical stability of dihydrofolate reductase*, in *Biophysical Journal*. 2005. p. 3337-44.
30. Brockwell, D.J., et al., *Mechanically unfolding the small, topologically simple protein L*, in *Biophysical Journal*. 2005. p. 506-19.
31. Junker, J.P., F. Ziegler, and M. Rief, *Ligand-dependent equilibrium fluctuations of single calmodulin molecules*, in *Science*. 2009. p. 633-7.
32. Liu, R., et al., *Mechanical characterization of protein L in the low-force regime by electromagnetic tweezers/evanescent nanometry*, in *Biophysical Journal*. 2009. p. 3810-21.
33. Evans, E. and K. Ritchie, *Dynamic strength of molecular adhesion bonds*, in *Biophys J*. 1997. p. 1541-55.
34. Hummer, G. and A. Szabo, *Kinetics from nonequilibrium single-molecule pulling experiments*, in *Biophys J*. 2003. p. 5-15.
35. Dudko, O.K., G. Hummer, and A. Szabo, *Theory, analysis, and interpretation of single-molecule force spectroscopy experiments*, in *Proc Natl Acad Sci USA*. 2008. p. 15755-60.

Chapter 5 A robust approach to estimating rates from time-correlation functions

5.1 Prospectus on Chapters 5 and 6

The next two chapters will focus on two novel approaches for identifying states and their lifetimes from single molecule experiments. The chapter entitled " A robust approach to estimating rates from time-correlation functions" discusses new approach for determining rate constants from single-molecule biophysical experiments. The chapter entitled "Bayesian hidden Markov model analysis of single-molecule biophysical experiments" discusses the method that was used for the analysis of the constant trap position experiments presented in this thesis. Both of these studies were carried out in collaboration with John Chodera. These chapters represent modified versions of manuscripts written mostly by John – my contribution and input was as described below.

Previous methods used for the analysis of single molecule hopping experiments such as the partition method [1-3], require a high signal-to-noise ratio. Using this method for the analysis of constant trap position experiments, with their lower signal-to-noise ratio, would result in an overestimate of the number of transitions and therefore an underestimate of the average lifetimes of a given state. Another way of looking at the problem of requiring a high signal-to-noise ratio for simpler methods to yield good rate estimates is whether the resolved coordinate (here, the bead position in the trap) is a good reaction-coordinate because there may be other relevant degrees of freedom we're not observing. The observed coordinate may still be able to mostly separate the states (and hence be a good order parameter) but may not be an ideal reaction coordinate. Therefore, a better method of analysis was required.

Fortuitously, John Chodera, Frank Noe, Bettina Keller, Aaron Ewall-Wice and Nina Singhal Hinrichs were developing a novel approach for analyzing single molecule data using a Bayesian hidden Markov model. Previous implementations of hidden Markov models rely on maximum-likelihood methods to determine the best parameters that fit the model, such as state identification and transition rates between states [4-6]. However, these methods have a significant shortcoming in that they do not quantitate how well the maximum-likelihood parameters are determined. The method described here uses a Bayesian extension of a hidden Markov model to estimate these uncertainties. In addition, John Chodera, William Swope, Jan-Hendrik Prinz, Frank Noe, and Vijay Pande were developing a second approach using a reactive flux theory-based method to extract rate estimates that are significantly less sensitive to the choice of the partition. Both of these methods are better for analyzing data with a lower signal-to-noise ratio and provide rate estimates despite not having an ideal reaction coordinate. To demonstrate the usefulness of these methods using experimental data, we worked together in collaboration.

In the following sections, I collected the data on the experimental system, the p5ab RNA hairpin. Using these methods requires high fidelity data recorded at a high sampling frequency. At the start of the collaboration, however, the quality and frequency of the data was limited. Due to hardware constraints, the instrument dropped ~40% of the data points at 4 kHz, affecting the value of the data reported and resulting in 2% of

the data being missed at 1 kHz, as previously described. In order to increase both the fidelity and frequency of the data, I bypassed the limiting hardware and collected the data on a second computer that recorded the voltage directly from the position sensitive detectors. This allowed data to be collected up to the response time of the detectors of 100 kHz with no dropped data.

In collaboration with John, I then helped calibrate the voltage-to-force conversion. I determined below what frequency the data was uncorrelated and suitable for analysis by the BHMM method. Then using the analysis software provided by John, I ran the analysis on the experimental analysis data.

5.2 Introduction

The estimation of rate constants from trajectories of microscopic systems is a fundamental problem in classical statistical mechanics. Under conditions where there is a separation of timescales such that a microscopic rate constant exists, the reactive flux theory of Chandler and coworkers demonstrates how the rate constant may be extracted from the plateau region of the reactive flux correlation function [7, 8]. When estimating the rate from trajectories generated in a computed or recorded at finite intervals in a laboratory experiment, the reactive flux correlation function requires an empirical time correlation function to be numerically differentiated to obtain the rate, often introducing an unacceptable amount of noise in the corresponding rate estimate. We present a modified version of reactive flux theory, which does not require numerical derivatives to be computed, allowing rate constants to be estimated directly and robustly from the time-correlation function. We illustrate the approach for a single-molecule force spectroscopy measurement of an RNA hairpin.

5.3 Results and discussion

Suppose we have a population of N non-interacting molecules that can occupy one of two conformational states, denoted A or B with defined associated indicator functions $h_A(x)$ and $h_B(x)$, where x denotes the molecular configurational degrees of freedom, such that $h_A(x)$ assumes the value of unity if the configuration x is in conformation A and zero otherwise; similarly, $h_B(x)$ assumes the value of unity only if x is in conformation B. Together, $h_A(x)$ and $h_B(x)$ form a partition of unity, such that $h_A(x) + h_B(x) = 1$ for all x in the accessible configuration space Γ of the molecule.

If there is a separation of timescales between the short relaxation time within the conformational states and the long time, the system must wait, on average, in one conformational state before undergoing a transition to another state, the asymptotic relaxation behavior of an initial population of $N_A(0)$ molecules in conformation A and $N_B(0)$ molecules in conformation B can be described by a simple linear rate law:

$$\frac{d}{dt}N_A(t) = -k_{A\to B}(t)N_A(t) + k_{B\to A}N_B(t) \tag{5.1}$$

where $k_{A\to B}$ and $k_{B\to A}$ are microscopic rate constants. In terms of non-equilibrium expectations, this is equivalent to:

$$\frac{d}{dt}\langle h_A(t)\rangle_{ne} = -k_{A\to B}\langle h_A(t)\rangle_{ne} + k_{B\to A}\langle h_B(t)\rangle_{ne} \tag{5.2}$$

where $\langle h_A(t)\rangle_{ne}$ denotes the non-equilibrium probability of finding a given molecule in conformation *A* at time *t* given that the fraction of molecules that were initially in conformation *A* was $\langle h_A(0)\rangle_{ne}$. (We hereafter write $h_A(t)$ as shorthand for $h_A(x(t))$.)

Were Eq. 5.2 to govern dynamics at all times, the expected fraction of molecules in conformation *A* as a function of time would be given by an exponential decay function:

$$\langle h_A(t)\rangle_{ne} = \langle h_A\rangle + [\langle h_A(0)\rangle_{ne} - \langle h_A\rangle]e^{-kt} \tag{5.3}$$

where the quantity $k \equiv k_{A\to B} + k_{B\to A}$ is denoted the phenomenological transition rate because it is the effective rate constant that dominates the observed exponential decay governing the asymptotic relaxation behavior. $\langle h_A\rangle$ denotes the standard equilibrium expectation, giving the equilibrium fraction of molecules in conformation *A*. Note that we do not expect Eq. 5.3 to hold for times $t < \tau_{mol}$, where τ_{mol} is the timescale associated with relaxation processes that damp out re-crossings that occur due to imperfect definition of the separatrix between the reactant and product states [9].

Chandler (and subsequent workers) demonstrated how the phenomenological rate could be computed using time-correlation functions by defining the reactive flux correlation function $k_{RF}(t)$ [7, 8, 10, 11]:

$$k_{RF}(t) = -\frac{d}{dt}\frac{\langle \delta h_A(0)\delta h_A(t)\rangle}{\langle \delta h_A^2\rangle} \tag{5.4}$$

where $\delta h_A(t) \equiv h_A(t) - \langle h_A\rangle$ is the instantaneous deviation from the equilibrium population for some trajectory $x(t)$.

$k_{RF}(t)$ measures the flux across the boundary between *A* and *B* that is reactive, in the sense that the system has crossed the dividing surface between *A* and *B* at time zero and is located on the product side of the boundary at time *t*. The reactive flux is bounded from above by the transition state theory estimate k_{TST}, the instantaneous flux across the boundary at time $t = 0$, because re-crossings back to the reactant state will diminish the reactive flux; $k_{RF}(t)$ becomes identical to k_{TST} as $t \to 0^+$. At *t* larger than some τ_{mol}, the timescale of relaxation processes within the conformational states, thermalization processes will cause the molecule to be captured either in its reactant or product states and remain there for a long time. As a result, the asymptotic rate constant (whose existence requires the presupposed separation of timescales) is only obtained at $\tau_{mol} < t \ll \tau_{rxn}$, where $k_{RF}(t)$ reaches a plateau value, decaying to zero at $t \gg \tau_{rxn}$ with a time constant of τ_{rxn} [7, 8]. Subsequent work extends reactive flux theory to the case of multiple conformations [10, 11].

The reactive flux correlation function $k_{RF}(t)$ can, in principle, be used to estimate the phenomenological rate constant *k* and microscopic rate constants $k_{A\to B}$ and $k_{B\to A}$ from one or more observed molecular trajectories collected either in a computer experiment or

a laboratory experiment, but this presents several practical difficulties. For observations recorded at fixed intervals in time, the time derivative of the correlation function (Eq. 5.4) must be estimated by finite-difference methods, but the presence of statistical error in the estimated correlation function often produces unacceptably large noise in the resulting estimate of $k_{RF}(t)$. Alternatively, the correlation function $\langle \delta h_A(0) \delta h_A(t) \rangle$ could be estimated and smoothed by fitting a polynomial to produce a smooth estimate of the derivative, but this introduces a bias due to the functional form of the fit that is difficult to quantify. If the reaction timescales τ_{rxn} is not very long compared to the observation interval, then the plateau region where $k_{RF}(t)$ is identical to the rate may be small and difficult to detect before $k_{RF}(t)$ decays. Lastly, alternative expressions to Eq. 5.4 exist in which the velocity normal to the separatrix at the time of barrier crossing is utilized instead of a time derivative of the empirical correlation function [7, 8], but it is difficult to compute this velocity computationally for complex dividing surfaces and the exact time of barrier crossing are generally not observed when sampling at discrete intervals.

To illustrate several of these pathologies, we computed the reactive flux $k_{RF}(t)$ from a single-molecule force trajectory for the p5ab RNA hairpin in an optical trap. This hairpin has been the subject of previous single-molecule force spectroscopy studies [12-14], and exhibits two-state kinetics as the hairpin folds and unfolds under an external biasing force. The force trace reports on the extension of the polymer within the optical trap; for example, as the hairpin folds, the end-to-end distance contracts, increasing the applied force as the polystyrene bead conjugated to the end of the polymer moves away from the center of the optical trap. At the stationary trap position used for data collection, the hairpin makes many transitions between the two states resolvable from the measured force in the 60-second trajectory, populating each state nearly equally (Figure 5.1a). Data was collected at 50 kHz using a dual-beam counter-propagating optical trap [15, 16], as previously published [14].

While the applied force may not be an ideal reaction coordinate—this is irrelevant for the estimation of the force using the reactive flux formalism. Provided the observed coordinate (here, the observed force) is a suitable order parameter for resolving both states to some degree, the rate estimate will fall to the true rate constant after some initial transient relaxation time.

Figure 5.1b and c shows the reactive flux correlation function $k_{RF}(t)$ (in black) estimated from the observed force trace sub-sampled to 1 kHz (Figure 5.1 b) or at 50 kHz (Figure 5.1 c). The time-derivative in Eq. 5.4 is estimated by one-sided differences. When estimated from 50 kHz data (Figure 5.1 c), the rate smoothly stabilizes after a transient time of $\tau_{mol} \approx 1$ ms, but the numerical derivative introduces a great deal of noise into the estimate (Figure 5.1 c, inset). When estimated from 1 kHz sub-sampled data (Figure 5.1 b), the plateau region is very difficult to detect, and the $k_{RF}(t)$ falls as t reaches times comparable to τ_{rxn}.

An alternative approach allows computation of the phenomenological rate without the need for time derivatives of the correlation function $C_{AA}(t)$. This approach instead estimates the matrix of rate constants implied by the state-to-state transition probabilities for a given lag time, hence these quantities are referred to as the implied rate constants

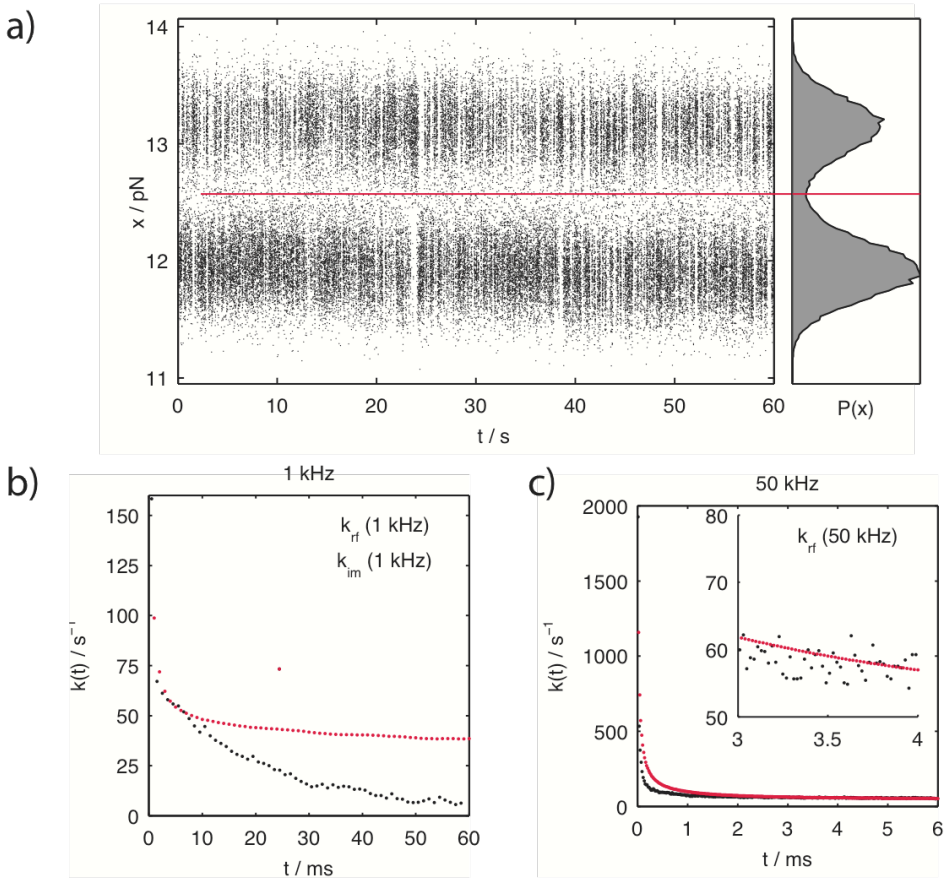

Figure 5.1 Reactive flux correlation function and implied rates from p5ab hairpin single-molecule force trajectory.

a is a 60 s force spectroscopy trace of p5ab RNA hairpin recorded at 50 kHz and sub-sampled to 1 kHz. **b** depicts the implied rate $k_{im}(t)$ (red) and reactive flux rate correlation function $k_{RF}(t)$ (black) from calculated from the 50 kHz force measurement data sub-sampled to 1 kHz. **c** depicts the same as **b**, but with the implied rate $k_{im}(t)$ (red) and reactive flux rate correlation function $k_{RF}(t)$ (black) calculated from the 50 kHz data. Inset: Close view comparing implied rate and reactive flux rate estimates between 3-4 ms for 50 kHz data.

for an observation interval τ. As with the reactive flux correlation function, for times $\tau_{mol} < t \ll \tau_{rxn}$, the phenomenological rate constant (if it exists, by virtue of a separation of timescales) is recovered.

As before, if a separation of timescales exists, relaxation behavior for times $t > \tau_{mol}$ is defined in terms of a rate matrix $\boldsymbol{K}$ by recasting Eq. 5.2 in matrix form:

$$\frac{d}{dt}\mathbf{p}(t) = \mathbf{K}\mathbf{p}(t) \tag{5.5}$$

where $\boldsymbol{p} = [p_A(t)p_B(t)]^T$, with $p_A(t) = \langle h_A(t) \rangle_{ne}$ and $p_B(t) = \langle h_B(t) \rangle_{ne}$ denoting the non-equilibrium occupation probabilities of states A and B at time t, and $\mathbf{K}$ is the matrix of rate constants:

$$\mathbf{K} = \begin{pmatrix} -k_{A\to B} & k_{B\to A} \\ k_{A\to B} & -k_{B\to A} \end{pmatrix} \tag{5.6}$$

The eigenvalues of $\boldsymbol{K}$ are $\lambda_1 = 0$, reflecting conservation of probability mass, and $\lambda_2 = -(k_{A\to B} + k_{B\to A}) = -k$, which governs the recovery toward equilibrium populations at the phenomenological relaxation rate.

The solution to Eq. 5.5 (corresponding to Eq. 5.3) is given by:

$$\mathbf{p}(t) = e^{\mathbf{K}t}\mathbf{p}(0) = \mathbf{T}(t)\mathbf{p}(0) \tag{5.7}$$

where $e^{\mathbf{A}} \equiv \sum_{n=0}^{\infty} \frac{\mathbf{A}^n}{n!}$ is the formal matrix exponential and $\boldsymbol{T(t)}$ can be identified as the column-stochastic transition probability matrix whose elements $T_{ji}(t)$ give the conditional probability of observing the system in conformation j at time t given that it was initially in conformation i at time 0.

The elements of $\boldsymbol{T(t)}$ for a given observation interval t are conveniently given in terms of the correlation function:

$$Tji(t) \equiv \frac{\langle h_i(0)h_j(t)\rangle}{\langle h_i \rangle} \equiv \frac{C_{ij}(t)}{\pi_i} \tag{5.8}$$

For $t > \tau_{mol}$, we have $\boldsymbol{T(t)} \approx e^{\boldsymbol{K}t}$, but we can establish a one-to-one correspondence between the rate matrix $\boldsymbol{K}_{im}(t)$ implied by $\boldsymbol{T(t)}$ for any t:

$$\mathbf{T}(t) = e^{\mathbf{K}_{im}(t)t} \Leftrightarrow \mathbf{K}_{im}(t) = t^{-1}\log\mathbf{T}(t) \tag{5.9}$$

For $t > \tau_{mol}$, all $\boldsymbol{K}_{im}(t) \approx \boldsymbol{K}$, and the rates are identical to those from the reactive flux theory.

Because of their relationship through the exponential (Eq. 5.9), $\boldsymbol{T(t)}$ and $\boldsymbol{K}_{im}(t)$ share the same eigenvectors $\boldsymbol{u}_k$, and their eigenvalues are also simply related [17]:

$$\begin{aligned}\mathbf{T}(t)\mathbf{u}_k &= u_k\mathbf{u}_k \\ \mathbf{K}_{im}(t)\mathbf{u}_k &= \lambda_k\mathbf{u}_k \\ u_k(t) &= e^{\lambda_k(t)t}\end{aligned} \tag{5.10}$$

An estimate of the phenomenological rate constant $k_{im}(t)$ for observation time t can be obtained from the second eigenvalue of $\boldsymbol{K}_{im}(t)$, which for $t > \tau_{mol}$ assumes the value of:

$$\begin{aligned}-(k_{A\to B} + k_{B\to A}) = -k: \\ k_{im}(t) = -\lambda_2(t) = t^{-1}\ln\mu_2(t)\end{aligned} \tag{5.11}$$

We note that $\mu_2(t)$ can be simply expressed:

$$\begin{aligned}\mu_2(t) &= 1-[T_{AB}(t)+T_{BA}(t)] \\ &= 1-\left[\frac{C_{AB}(t)}{\pi_A}+\frac{C_{BA}(t)}{\pi_B}\right] \\ &= 1-C_{AB}(t)\left[\frac{1}{\pi_A}+\frac{1}{\pi_B}\right] \\ &= 1-\frac{C_{AB}(t)}{\pi_A(1-\pi_A)} \\ &= 1-\frac{\pi_A-C_{AA}(t)}{\pi_A(1-\pi_A)} \\ &= \frac{C_{AA}(t)-\pi_A^2}{\pi_A-\pi_A^2} \\ &= \frac{\langle\delta h_A(0)\delta h_A(t)\rangle}{\langle\delta h_A^2\rangle}\end{aligned} \tag{5.12}$$

This is simply the normalized fluctuation autocorrelation function for the indicator function h_A for state A. It assumes the value of unity at $t = 0$ and decays to zero at large t.

In the limit $t \to 0^+$, $k_{im}(t)$ reduces to the transition state theory estimate k_{TST}. To see this, we expand $C_{AA}(t)$ in terms of its behavior near $t = 0$,

$$\begin{aligned}C_{AA}(t) &= C_{AA}(0)+t\dot{C}_{AA}(0)+O(t^2) \\ &= \pi_A+t\dot{C}_{AA}(0)+O(t^2)\end{aligned} \tag{5.13}$$

and so

$$\mu_2(t) = 1+t\frac{\dot{C}_{AA}(0)}{\pi_A-\pi_A^2}+O(t^2) \tag{5.14}$$

Near $t = 0$, $\mu_2(t) \approx 1$, allowing us to expand the argument to the logarithm appearing in $k_{im}(t)$ to first order in t about unity:

$$\begin{aligned}\lim_{t\to 0^+} k_{im}(t) &= \lim_{t\to 0^+} -t^{-1}\ln\mu_2(t) \\ &= \lim_{t\to 0^+} -t^{-1}[\mu_2(t)-1] \\ &= -\frac{\dot{C}_{AA}(0)}{\pi_A(1-\pi_A)}\end{aligned} \tag{5.15}$$

To illustrate the estimation of the rate using the implied timescale $k_{im}(t)$, we again analyze the 60-second force trajectory of the p5ab hairpin considered above. At high sampling rates (Figure 5.1 c), the rate estimates are almost identical to those from $k_{RF}(t)$ for $t > \tau_{mol}$, though there is much less noise than in $k_{RF}(t)$ (inset). At the 1 kHz sampling rate, however, the rate estimate from $k_{im}(t)$ remains stable over several times τ_{rxn}, even though the $k_{RF}(t)$ has already decayed from the plateau region.

To estimate the rates and their uncertainties from a set of equilibrium trajectories X_n, we define two trajectory functionals:

$$\begin{aligned}F[X] &\equiv h_A(x_0) \\ G[X] &\equiv h_A(x_0)h_B(x_\tau)\end{aligned} \tag{5.16}$$

Evaluating these functionals on a dataset of N statistically independent trajectories $X_n(t)$, $n = 1, \ldots, N$, collected from a single temperature β gives us a set of observables:

$$\begin{aligned}F_n &= F[X_n] \\ G_n &= G[X_n]\end{aligned} \tag{5.17}$$

Estimates of their expectations are given by the sample means:

$$\begin{aligned}\pi_A \approx \hat{F} &= \frac{1}{N}\sum_{n=1}^{N} F_n \\ C_{AB}(t) \approx \hat{G} &= \frac{1}{N}\sum_{n=1}^{N} G_n\end{aligned} \tag{5.18}$$

We compute the rate constant for a given observation time t (whose functional dependence we shall suppress) from an estimate of the second eigenvalue $\hat{\mu}_2$:

$$\hat{k}_{im} = -t^{-1}\ln\hat{\mu}^2 \tag{5.19}$$

The second eigenvalue is estimated from Eq. 5.12:

$$\mu_2 = 1 - \frac{C_{AB}(t)}{\pi_A(1-\pi_A)} \Leftrightarrow \hat{\mu}_2 = 1 - \frac{\hat{G}}{\hat{F}(1-\hat{F})} \tag{5.20}$$

The squared uncertainty in $k_{im}(t)$ in terms of the uncertainty in the second

eigenvalue $\mu_2(t)$ can be estimated by simple first-order Taylor series expansion propagation of error:

$$\delta^2 \hat{k}_{im} = \left[\frac{\partial \hat{k}_{im}}{\partial \hat{\mu}_2}\right]^2 \delta^2 \hat{\mu} = \frac{\delta^2 \hat{\mu}_2}{t^2 \hat{\mu}_2^2} \tag{5.21}$$

We apply the first-order Taylor propagation of error to compute the uncertainty in $\delta^2 \hat{\mu}_2$:

$$\delta^2 \hat{\mu}_2 = \left[\frac{\partial \hat{\mu}_2}{\partial \hat{F}}\right]^2 \delta^2 \hat{F} + \left[\frac{\partial \hat{\mu}_2}{\partial \hat{G}}\right]^2 \delta^2 \hat{G} + 2\left[\frac{\partial \hat{\mu}_2}{\partial \hat{F}}\right]\left[\frac{\partial \hat{\mu}_2}{\partial \hat{G}}\right]\delta\hat{F}\delta\hat{G} \tag{5.22}$$

where the derivatives are given by:

$$\frac{\partial \hat{\mu}_2}{\partial \hat{F}} = \frac{\hat{G}(1-2\hat{F})}{\hat{F}^2(1-\hat{F})^2}; \frac{\partial \hat{\mu}_2}{\partial \hat{G}} = +\frac{1}{\hat{F}(1-\hat{F})} \tag{5.23}$$

To estimate $\delta^2 \hat{\mu}_2$, we must first estimate the variance and covariance of the estimators $\hat{F}$ and $\hat{G}$:

$$\delta^2 \hat{A} = \frac{\operatorname{var} A_n}{N}; \delta^2 \hat{B} = \frac{\operatorname{var} B_n}{N}; \delta\hat{A}\delta\hat{B} = \frac{\operatorname{cov}(A_n, B_n)}{N} \tag{5.24}$$

where the sample covariances are used to estimate $varA_n$, $var\ B_n$, and $cov\ (A_n, B_n)$.

Practically, we take advantage of time-reversibility of dynamics, and use a slightly modified set of trajectory functionals that yield the same expectation but average over more snapshots from the trajectory in the case that the trajectory segments are of length $T > \tau$:

$$\begin{aligned} G[X] &= \frac{1}{T-\tau}\sum_{t_0=0}^{T-\tau}\frac{1}{2}\left[h_A(x_0)h_B(x_{t_0+\tau}) + h_B(x_0)h_B(x_{t_0+\tau})\right] \\ F[X] &= \frac{1}{T-\tau}\sum_{t_0=0}^{T-\tau}\frac{1}{2}\left[h_A(x_0) + h_A(x_{t_0+\tau})\right] \end{aligned} \tag{5.25}$$

This appropriately accounts for the fact that all time origins produce equally valid estimates and, for systems with multiple conformational states, ensures satisfaction of detailed balance.

5.4 References

1. Liphardt, J., et al., *Reversible unfolding of single RNA molecules by mechanical force*, in *Science*. 2001. p. 733-7.
2. Woodside, M.T., et al., *Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins*, in *Proc Natl Acad Sci USA*. 2006. p. 6190-5.
3. Wen, J.-D., et al., *Force unfolding kinetics of RNA using optical tweezers. I. Effects of experimental variables on measured results*, in *Biophysical Journal*. 2007. p. 2996-3009.
4. Qin, F., A. Auerbach, and F. Sachs, *A direct optimization approach to hidden Markov modeling for single channel kinetics*, in *Biophysical Journal*. 2000. p. 1915-27.
5. Qin, F., A. Auerbach, and F. Sachs, *Hidden Markov modeling for single channel kinetics with filtering and correlated noise*, in *Biophysical Journal*. 2000. p. 1928-44.
6. Schroder, G. and H. Grubmuller, *Maximum likelihood trajectories from single molecule fluorescence resonance energy transfer experiments.* J Chem Phys, 2003. **119**(18): p. 9920-9924.
7. Chandler, D., *Statistical mechanics of isomerization dynamics in liquids and the transition state approximation.* J. Chem. Phys., 1978. **68**(6): p. 2959-2970.
8. Montgomery, J.A., Jr., D. Chandler, and B.J. Berne, *Trajectory analysis of a kinetic theory for isomerization dynamics in condensed phases.* J. Chem. Phys., 1979. **70**(9): p. 4056-4066.
9. H\"anggi, P., P. Talkner, and M. Borkovec, *Reaction-rate theory: fifty years after Kramers.* Rev. Mod. Phys., 1990. **62**: p. 251-342.
10. Adams, J.E. and J.D. Doll, *Dynamical aspects of precursor state kinetics.* Surface Science, 1981. **111**: p. 492-502.
11. Voter, A.F. and J.D. Doll, *Dynamical corrections to transition state theory for multistate systems: Surface self-diffusion in the rare-event regime.* J. Chem. Phys., 1985. **82**(1): p. 80-92.
12. Liphardt, J., et al., *Reversible unfolding of single RNA molecules by mechanical force.* Science, 2001. **292**: p. 733-737.
13. Wen, J.-D., et al., *Force unfolding kinetics of RNA using optical tweezers. I. Effects of experimental variables on measured results.* Biophys. J., 2007. **92**: p. 2996-3009.
14. Elms, P.J., et al., *Equilibrium force spectroscopy experiments on macromolecules: The problem with force feedback experiments.* In preparation, 2010.
15. Smith, S.B., Y. Cui, and C. Bustamante, *Optical-trap force transducer that operates by direct measurement of light momentum.* Meth. Enzym., 2003. **361**: p. 134-162.

16. Bustamante, C. and S.B. Smith. *Light-force sensor and method for measuring axial optical-trap forces from changes in light momentum along an optical axis*. 2006.
17. Buchete, N.-V. and G. Hummer, *Coarse master equations for peptide folding dynamics.* J. Phys. Chem. B, 2008. **112**: p. 6057-6069.

Chapter 6 Bayesian hidden Markov model analysis of single-molecule biophysical experiments

6.1 Summary

Single-molecule experiments are now able to probe the dynamics of single biological macromolecules or macromolecular assemblies using a variety of techniques, including fluorescence measurements, fluorescent energy transfer (FRET), and optical or atomic force microscopy. While able to report on the behavior of single molecules, the observed signal is usually only an indirect probe of molecular conformation, without the guarantee of a unique correspondence between the observed signal and molecular conformation. This can lead to conformational states with strongly overlapping spectroscopic signatures, making resolution of the instantaneous molecular configuration difficult. Hidden Markov models (HMMs), now a standard approach in machine learning, have been employed to solve this problem by using kinetic information to aid resolution of the observed temporal signal into a sequence of distinct conformational states. These methods suffer from an important drawback: maximum-likelihood fitting procedures do not give a clear picture of how well the model parameters are determined by the data due to instrument noise and finite-sample statistics. Here, we propose a solution to this problem through a simple Bayesian extension of hidden Markov model analysis that allows both the uncertainties in the transition rates and hidden state assignments to be characterized. The method is based on Gibbs sampling, allowing it to be easily extended to other models of observables or to multiple observables by simply "plugging in" new components of the model.

6.2 Introduction

Recent advances in biophysical measurement have led to an unprecedented ability to monitor the dynamics of single biomolecules, such as proteins and nucleic acids [1]. These experiments aim to probe the statistical, heterogeneous dynamics relevant to folding and function. Recent studies have examined the conformational dynamics of large RNA molecules under equilibrium and nonequilibrium conditions by monitoring the energy transfer between two covalently attached fluorophores [2, 3]; the turnover of individual molecules of substrate by enzymes [4]; permeation and gating events of single ion channels [5]; and the fluctuation of nucleic acids under external forces in optical traps [6, 7] or atomic force microscopes [8].

Unlike corresponding ensemble experiments, where spectroscopic observables appear to evolve deterministically after an external perturbation (such as a laser-induced temperature jump) and spectroscopic fluctuations cannot generally be observed at equilibrium, spectroscopic probes of single molecules both in and out of equilibrium exhibit a great deal of stochastic fluctuation. While some of this fluctuation is undoubtedly due to measurement noise, some large component of this fluctuation is due to conformational dynamics of the molecule under study. Often, the dynamics appears to be dominated by stochastic interconversions between two or more strongly metastable states, regions of conformation space in which the system persists for long times before making a transition (often accurately described by first-order kinetics) to another state (a

situation also observed in NMR relaxation-dispersion experiments [9] and molecular dynamics simulations [10]).

While visual inspection of the dynamics may suggest the clear presence of multiple metastable states, characterization of these states is often difficult. First, the spectroscopic observable is unlikely to correspond to a true reaction coordinate easily able to separate all metastable states, and second, measurement noise may further broaden the spectral signatures of individual states. As a result, there is often a large degree of spectral overlap in the signatures of individual states [2, 3]. Attempting to separate these states with simple separation points can often lead to a high degree of state misassignment that corrupts both the distribution of observables and characterization of rates of interconversion between states [11]. Hidden Markov models (HMMs) [12], which use temporal information in addition to the observable to determine which hidden state the system is currently in, have provided an effective solution to this problem. In an HMM, the observed signal is assumed to come from a realization of an underlying Markov chain, where the system makes history-independent transitions among a set of discrete states with probabilities governed by a transition or rate matrix. The experimenter does not know which state the system is in, and can only measure some observable whose value is determined by a probability distribution of observables characterizing each state (which may overlap). Given a set of data, maximum likelihood estimates (MLEs) of the model parameters (transition rates and state observable distributions) and sequence of hidden states corresponding to the observed data can be determined by standard methods [13, 14]. Unfortunately, this approach has a number of serious limitations. Single-molecule experiments often suffer from limited statistics; the events of interest (transitions between states) may occur only a few times during the course of the measurement. As a result, while the MLE may give the most likely set of model parameters, there may be enormous uncertainty in these parameters, and the MLE provides no simple way to characterize them. These uncertainties may also be highly correlated, in that certain combinations of parameters may be well determined in a complex way, despite individual parameters being poorly determined. The high cost (both in terms of instrument and experimenter time) of collecting additional data also means that it is not a simple task to judge how much data need be collected to test a particular hypothesis in a statistically meaningful way.

Here, we present a resolution to this issue in terms of a Bayesian extension of hidden Markov models applicable to single molecule experiments. By sampling over the posterior distribution of model parameters and hidden state assignments instead of simply finding the most likely values, the experimenter is able to accurately characterize the correlated uncertainties in both the model parameters (transition rates and state observable distributions) and hidden state sequences corresponding to observed data. Additionally, prior information (either from independent measurements or physical constraints) can be easily incorporated. The framework we present is based on Gibbs sampling [15, 16], allowing simple swap-in replacement of models for observable distributions, extension to multiple observables, and alternative models for state transitions. Additionally, the Bayesian method provides a straightforward way to model the statistical outcome and assess the utility of additional experiments given some preliminary data, allowing the experimenter a powerful tool for assessing whether the

cost of collecting additional data is outweighed by their benefits.

6.3 Hidden Markov models

We now describe the basic theory behind the maximum likelihood estimate for a hidden Markov model (MLHMM) and corresponding Bayesian extension (BHMM). While any scheme for computing the maximum-likelihood estimator or sampling from the Bayesian posterior can be used to generate these models, the algorithms used in this work are described in detail Section 4.3. Due to the abundance of mathematical notation, we summarize important symbols used throughout in Table I.

O	$o_t^{(n)}$	observed temporal traces
S	$s_t^{(n)}$	hidden sequencs space
T	T_{ij}	transition probability for Δt
E	e_s	state observable distribution parameters
Θ		model parameters $\mathbf{\Theta} \equiv \{\mathbf{T},\mathbf{E}\}$
M	m	number of hidden states
N	n	number of independent observed traces
$L^{(n)}$	t	length of observed trace
ρ	ρ_i	initial state probability distribution
π	π_i	equilibrium state probability
$\varphi(o\|e)$		state observaton probability distribution

Table 1 Summary of important symbols and their elements.

6.3.1 Preliminaries

Suppose we observe N independent temporal traces, where some observable $O(x)$ that is a function of molecular configuration x is observed at temporal intervals Δt. This observable may be, for example, the measured force or extension of a polymer in a force microscopy experiment, an observed FRET efficiency, or an ion current measured by patch-clamp electrophysiology. While we restrict ourselves to consideration of scalar functions $O(x)$, the extension to multidimensional probes (or multiple probes) is straightforward.

Let trace n be denoted by $o_t^{(n)}$, where $t \in \{0,1,2,...,L^{(n)}\}$, collected with uniform sampling interval Δt. We allow the system under observation to either start from equilibrium at the beginning of the observation period (if sufficient time has been allowed for the system to reach equilibrium), or from an out-of-equilibrium initial configuration (such as preparing a protein system by mechanically unfolding it prior to

starting observation).

We presume the system under study has M kinetically metastable states, in the sense that they persist for many observation intervals Δt but may not represent the lowest free energy (most populous) state of the system. (In the language of chemical kinetics, we require that the molecular relaxation time within a state $\tau_{mol} \ll \Delta t$, but the typical reaction time for transitioning between states $\tau_{rxn} \gg \Delta t$ [17-19].) We treat these as the hidden states of the model, because we cannot directly observe the identity of the metastable state in which the system resides.

The hidden Markov model presumes the observed data $O \equiv \{o_t^{(n)}\}$, where $n = 1,...,N$ and $t = 0,...,L^{(n)}$, was generated according to the following model dependent on parameters $\mathbf{\Theta} \equiv \{\mathbf{T}, \mathbf{E}\}$ and prior information about the initial state distribution $\rho(n)$:

$$\begin{aligned} s_0^{(n)} &\sim p_{s_0^{(n)}}^{(n)} \\ s_t^{(n)} \mid s_{t-1}^{(n)}, \mathbf{T} &\sim T_{s_{t-1}^{(n)} s_t^{(n)}}, t \geq 1 \qquad (1) \\ o_t^{(n)} \mid s_t^{(n)}, e_{s_t^{(n)}} &\sim \varphi(o_t^{(n)} \mid e_{s_t^{(n)}}) \end{aligned}$$

In diagrammatic form, the observed state data $o^{(n)}$ and corresponding hidden state history $s^{(n)}$ can be represented

$$\begin{array}{ccccccccc} s^{(n)} \equiv & s_0^{(n)} & \rightarrow & s_1^{(n)} & \rightarrow & s_2^{(n)} & \rightarrow \cdots & s_{L^{(n)}}^{(n)} \\ & \downarrow & & \downarrow & & \downarrow & & \downarrow \qquad (2) \\ o^{(n)} \equiv & o_0^{(n)} & \rightarrow & o_1^{(n)} & \rightarrow & o_2^{(n)} & & o_{L^{(n)}}^{(n)} \end{array}$$

Here, state transitions $(s_{t-1}^{(n)} \rightarrow s_t^{(n)})$ are governed by the discrete transition probability $T_{s_{t-1}^{(n)} s_t^{(n)}}$, while the "emission" of observables from each state $(s_t^{(n)} \rightarrow o_t^{(n)})$ is governed by the continuous emission probability $\varphi(o_t^{(n)} \mid e_{s_t^{(n)}})$. As an alternative to using the transition matrix $\mathbf{T}$ as a model parameter, one could instead use the rate matrix $\mathbf{K}$ related by $\mathbf{T} = e^{\mathbf{K}\tau}$.

The initial state distribution $\rho^{(n)}$ (which may itself be a function of the stationary distribution π of $\mathbf{T}$) simply reflects our knowledge of the initial conditions of the experiment that collected data $o^{(n)}$. In the case that the experiment was prepared in equilibrium, ρ corresponds to the equilibrium distribution π of the model transition matrix $\mathbf{T}$. However, if the experiment was started out of equilibrium, perhaps restricted to some subset of states, then the prior might reflect simple ignorance as to which state the system initially started in by assigning each state in this subset an equal probability. Here, we presume that either ρ is known a priori, or that it is a function of the equilibrium distribution π determined by $\mathbf{T}$.

The Markov property of HMMs prescribes that the probability that a system originally in state i at time t is later found in state j at time $t + 1$ is dependent only on knowledge of the state i, and given by the corresponding matrix element $\mathbf{T}_{ij}$ of the (row-

stochastic) transition matrix **T**.

The probability that a particular value o of the observable is measured is dependent only on the current state *s*, and given by some model of the observable distribution for this state $\varphi(o|e_s)$ parameterized by observable emission parameters *e*.

For example, in the applications to force spectroscopy described in this paper, the observable denotes the measured force exerted on a bead in an optical trap, and the model is taken to be a simple Gaussian distribution parameterized by $e \equiv \{\mu,\sigma^2\}$:

$$\varphi(o \mid e) = \varphi(o \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{(o-\mu)^2}{\sigma^2}\right]. \qquad (3)$$

Given the HMM process specified in Eq. 1, the probability of observing the data *O* given the model parameters **Θ** is therefore

$$P(O \mid \Theta) = \sum_{s} \prod_{n=1}^{N} \rho_{s_0^{(n)}} \varphi(o_0^{(n)} \mid e_{s_0^{(n)}}) \prod_{t=1}^{L^{(n)}} T_{s_{(t-1)}^{(n)} s_t^{(n)}} \varphi(o_0^{(n)} \mid e_{s_t^{(n)}}) \qquad (4)$$

where the sum over hidden state histories **S** is shorthand for

$$\sum_{S} \equiv \sum_{s_0^{(1)}=1}^{M} \cdots \sum_{s_{L^{(1)}}^{(1)}=1}^{M} \cdots \sum_{s_{L^{(N)}}^{(N)}=1}^{M} . \qquad (5)$$

6.3.2 Maximum likelihood hidden Markov model (MLHMM)

The standard approach to construct an HMM from observed data is to compute the maximum likelihood estimator (MLE) for the model parameters **Θ**, which maximize the probability of the observed data *O* given the model,

$$\hat{\Theta} = \arg\max_{\Theta} P(\mathbf{O} \mid \Theta) \qquad (6)$$

Once the MLE parameters **Θ** are determined, the most likely hidden state history that produced the observations *O* can be determined using these parameters:

$$\hat{S} = \arg\max_{S} P(\mathbf{S} \mid \mathbf{O}, \hat{\Theta}) \qquad (7)$$

6.3.3 Bayesian hidden Markov model (BHMM)

Instead of simply determining the model that maximizes the likelihood of observing the data *O* given the model parameters **Θ**, we can make use of Bayes' theorem to compute the posterior distribution of model parameters given the observed data:

$$P(\Theta \mid O) \propto P(O \mid \Theta) P(\Theta) \qquad (8)$$

Here, $P(\mathbf{\Theta})$ denotes a prior distribution that encodes any a priori information we may have about the model parameters **Θ**. This might include both physical constraints

(such as ensuring the transition matrix satisfy detailed balance) and prior rounds of inference from other independent experiments.

Making use of the likelihood (Eq. 4), the model posterior is then given by,

$$P(\Theta \mid O) \propto P(\Theta) \times \sum_{S} \prod_{n=1}^{N} \rho_{s_0^{(n)}} \varphi(o_0^{(n)} \mid e_{s_0^{(n)}}) \prod_{t=1}^{L^{(n)}} T_{s_{t-1}^{(n)} s_t^{(n)}} \varphi(o_t^{(n)} \mid e_{s_t^{(n)}}) \quad (9)$$

Drawing samples of $\boldsymbol{\Theta}$ from this distribution will, in principle, allow the confidence with which individual parameters and combinations thereof are known, given the data (and subject to the validity of the model of Eq. 1 in correctly representing the process by which the observed data is generated). However, due to the sum over all hidden state histories $\mathbf{S}$ appearing in the posterior (Eq. 9), direct sampling of the model parameters $\boldsymbol{\Theta}$ is difficult. Instead, we take the approach of introducing the state histories $\mathbf{S}$ as an auxiliary variable, sampling from the augmented posterior

$$P(\Theta, S \mid O) \propto P(\Theta) \times \prod_{n=1}^{N} \rho_{s_0^{(n)}} \varphi(o_0^{(n)} \mid e_{s_0^{(n)}}) \prod_{t=1}^{L^{(n)}} T_{s_{t-1}^{(n)} s_t^{(n)}} \varphi(o_t^{(n)} \mid e_{s_t^{(n)}}) \quad (10)$$

If we presume the prior is separable

$$P(\Theta) \equiv P(\mathbf{T}) P(\mathbf{E}) \quad (11)$$

we can sample from the augmented posterior (Eq. 10) using the framework of Gibbs sampling [16], in which the augmented model parameters are updated by sampling from the conditional distributions,

$$\begin{aligned} P(\mathbf{S} \mid \mathbf{T}, \mathbf{E}, \mathbf{O}) &\propto \prod_{n=1}^{N} \rho_{s_o^{(n)}} \varphi(o_0^{(n)} \mid e_{s_o^{(n)}}) \prod_{n=1}^{N} T_{s_{t-1}^{(n)} s_t^{(n)}} \varphi(o_t^{(n)} \mid e_{s_t^{(n)}}) \\ P(\mathbf{T} \mid \mathbf{E}, \mathbf{S}, \mathbf{O}) = P(\mathbf{T} \mid \mathbf{S}) &\propto P(\mathbf{T}) \prod_{n=1}^{N} \prod_{t=1}^{L} T_{s_{t-1}^{(n)} s_t^{(n)}} \\ P(\mathbf{E} \mid \mathbf{S}, \mathbf{T}, \mathbf{O}) = P(\mathbf{E} \mid \mathbf{S}, \mathbf{O}) &\propto P(\mathbf{E}) \prod_{n=1}^{N} \prod_{t=0}^{L} \varphi(o_t^{(n)} \mid e_{s_t^{(n)}}) \end{aligned} \quad (12)$$

The equalities on the second and third lines reflect the conditional independence of the hidden Markov model defined by Eq. 1. When only the model parameters $\boldsymbol{\Theta} \equiv \{\mathbf{T}, \mathbf{E}\}$ or the hidden state histories $\mathbf{S}$ are of interest, we can simply marginalize out the uninteresting variables by sampling from the augmented joint posterior for $\{\mathbf{T}, \mathbf{E}, \mathbf{S}\}$ and examining only the variables of interest. In addition, the structure of the Gibbs sampling scheme above allows individual components (such as the observable distribution model $\varphi(o \mid e)$ or transition probability matrix $\mathbf{T}$ to be modified without affecting the structure of the remainder of the calculation.

6.4 Bayesian Experimental Design

The Bayesian treatment equips us with both a model of the parameters given data

(Eq. 9) and a model of the data given the parameters (Eq. 4), allowing an experimenter to use prior knowledge or preliminary experimental data to model the outcome of new experiments, and how the collection of additional experimental data can be expected to reduce model uncertainties. For example, suppose we have conducted an experiment ε_1, which yielded data O_1. Using the information from this experiment, we can model the probability that a yet-to-be- performed experiment ε_2 will yield data O_2,

$$P(O_2 \mid \varepsilon_2, \{O_1, \varepsilon_1\}) = \int d\Theta P(O_2 \mid \varepsilon_2, \Theta) P(\Theta \mid \{\varepsilon_1, O_1\})$$

As a simple illustration of the utility in experimental design, we assume that a prior observation has been made to produce observed dataset O_1, and that the distribution $P_2(O_2|\mathbf{\Theta})$ describes the probability of observing some data O_2 (from a potentially different observable) given the model parameters $\mathbf{\Theta}$. Based on the information gathered from the first observation O_1, the expected information content of the second experiment to collect O_2 can be written as

$$E[I(O_2 \mid O_1)] = H[P_1(\Theta_1 \mid O_1)] - \int dO_2 H[P_2(\Theta_2 \mid O_2, O_1)] \int d\Theta_1 P_2(O_2 \mid \Theta_1) P_1(\Theta_1 \mid O_1) \quad (13)$$

where $H[P(\mathbf{\Theta})] \equiv -\int d\mathbf{\Theta} P(\mathbf{\Theta}) \ln P(\mathbf{\Theta})$ denotes the Shannon entropy or uncertainty of a distribution $P(\mathbf{\Theta})$. While direct computation of Eq. 13 can be challenging, approaches have been developed to compute useful approximations for use in Bayesian experimental design [20].

6.5 Algorithms

Below, we outline the algorithms we use for generating an initial model subject to prior constraints, computing a maximum-likelihood hidden Markov model (MLHMM), and sampling from the Bayesian posterior (BHMM). A Markov model requires that each time point is independent and uncorrelated from the previous time point. Therefore, the data must be sampled at a time interval greater than the relaxation time of the probe, such as the bead in an optical trap. This relaxation time can be determined experimentally by determining the corner frequency of the system from a power spectrum or, alternatively, autocorrelation function of the signal [21].

6.5.1 Generating an initial model

To initialize either computation of the MLHMM or sampling from the posterior for the BHMM, an initial model that respects any constraints imposed in the model prior $P(\mathbf{\Theta})$ must be selected. Here, we employ a Gaussian observable distribution model for $\varphi(o|\mathbf{e})$ (Eq. 3) and enforce that the transition matrix $\mathbf{T}$ satisfy detailed balance. Physical systems that are not driven by an external force or energy reservoir should satisfy detailed balance [22], and its use has been shown to provide a large reduction in transition matrix uncertainty in data-poor conditions [23]. Detailed balance specifies that $\pi_i T_{ij} = \pi_j T_{ji}$ for all i, j, where π is the equilibrium distribution of the row-stochastic transition matrix $\mathbf{T}$.

6.5.2 Observable parameter estimation

We first initialize the observed distributions of each state by fitting a Gaussian

mixture model with M states to the pooled observed data O, ignoring temporal information:

$$P(O \mid \pi, \mathbf{E}) = \prod_{n=1}^{N} \prod_{t=0}^{L^{(n)}} \sum_{m=1}^{M} \pi_m \varphi(o_t^{(n)} \mid \mu_m, \sigma_m^2) \qquad (14)$$

where the state observable emission probability vector $\mathbf{E} \equiv \{e_1,...,e_M\}$ and $e_m \equiv \{\mu_m, {\sigma_m}^2\}$ with μ_m denoting the observable mean and ${\sigma_m}^2$ the variance for state m for the Gaussian observable model. The vector π is composed of equilibrium state populations $\{\pi_1,\ldots,\pi_M\}$ with $\pi_m \geq 0$ and $\sum_{m=1}^{M} \pi_m = 1$.

A first approximation to π and $\mathbf{E}$ is computed by pooling and sorting the observed $o_t^{(n)}$, and defining M indicator functions $h_m(o)$ that separate the data into M contiguous regions of the observed range of o of roughly equal population. Let $N_m \equiv \sum_{n=1}^{N} \sum_{t=1}^{L^{(n)}} h_m(o_t^{(n)})$ denote the total number of observations falling in region m, and $N_{tot} = \sum_{m=1}^{M} N_m$. The initial parameters are then computed as:

$$\begin{aligned} \pi_m &= N_m / N_{tot} \\ \mu_m &= N_m^{-1} \sum_{n=1}^{N} \sum_{t=0}^{L^{(n)}} o_t^{(n)} h_m(o_t^{(n)}) \end{aligned} \qquad (15)$$

$$\sigma_m^2 = N_m^{-1} \sum_{n=1}^{N} \sum_{t=0}^{L^{(n)}} (o_t^{(n)} - \mu_m)^2 h_m(o_t^{(n)}) \qquad (16)$$

This approximation is then improved upon by utilizing the expectation-maximization procedure described by Bilmes [24].

$$\begin{aligned} \pi'_m &= N_{tot}^{-1} \sum_{n=1}^{N} \sum_{t=0}^{L^{(n)}} \chi_m(o_t^{(n)}, \mathbf{E}, \pi) \\ \mu'_m &= (\pi'_m N_{tot})^{-1} \sum_{n=1}^{N} \sum_{t=0}^{L^{(n)}} o_t^{(n)} \chi_m(o_t^{(n)}, \mathbf{E}, \pi) \\ \sigma'^2_m &= (\pi'_m N_{tot})^{-1} \sum_{n=1}^{N} \sum_{t=0}^{L^{(n)}} (o_t^{(n)} - \mu'_m)^2 \chi_m(o_t^{(n)}, \mathbf{E}, \pi) \end{aligned} \qquad (17)$$

where the function $\chi_m(o, \mathbf{E}, \pi)$ is given by the fuzzy membership function:

$$\chi_m(o, \mathbf{E}, \pi) = \frac{\pi_m \varphi(o \mid e_m)}{\sum_{l=1}^{M} \pi_l \varphi(o \mid e_l)} \qquad (18)$$

This iterative procedure is terminated at iteration j when the change in the parameters $\{\pi,\mu,\sigma^2\}$ falls below a certain relative threshold, such as $||\pi^{[j]} - \pi^{[j-1]}||/||\pi^{[j]}|| < 10^{-4}$.

6.5.3 Transition matrix estimation

Once initial state observable emission parameters **E** are determined, an initial transition matrix is estimated using an iterative likelihood maximization approach that enforces detailed balance [25]. First, a matrix of fractional transition counts $\mathbf{C} \equiv (c_{ij})$ is estimated using the membership function:

$$c_{ij} = \sum_{n=1}^{N} \sum_{t=1}^{L^{(n)}} \chi_i(o_{t-1}^{(n)}, \mathbf{E}, \pi) \chi_j(o_t^{(n)}, \mathbf{E}, \pi) \qquad (19)$$

A symmetric $M \times M$ matrix $\mathbf{X} \equiv (x_{ij})$ is initialized by

$$x_{ij} = x_{ji} = c_{ij} + c_{ji} \qquad (20)$$

and a vector of row sums

$$x_{i*} = \sum_{j=1}^{M} x_{ij} . \qquad (21)$$

Then, the iterative procedure described in Algorithm 1 of [25] is applied. For each updated iteration, we first update the diagonal elements of **X**:

$$x'_{ii} = \frac{c_{ii}(x_{i*} - x_{ii})}{c_{i*} - c_{ii}} \qquad (22)$$

where

$$c_{i*} = \sum_{j=1}^{M} c_{ij} \qquad (23)$$

followed by the off-diagonal elements:

$$x'_{ij} = x'_{ji} = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad (24)$$

where the quantities a, b, and c are computed from **X** and **C** as

$$\begin{aligned} a &\equiv c_{i*} - c_{ij} + c_{j*} - c_{ji} \\ b &\equiv c_{i*}(x_{j*} - x_{ji}) + c_{j*}(x_{i*} - x_{ji}) - (c_{ij} + c_{ji})(x_{i*} - x_{ij} + x_{j*} - x_{ji}) \\ c &\equiv -(c_{ij} + c_{ji})(x_{i*} - x_{ij})(x_{j*} - x_{ji}) \end{aligned} \qquad (25)$$

Once a sufficient number of iterations j have been completed to compute a stable estimate of **X** (such as the relative convergence criteria $||\mathbf{X}^{[j]} - \mathbf{X}^{[j-1]}||/||\mathbf{X}^{[j]}|| < 10^{-4}$, the maximum likelihood transition matrix estimate **T** is computed as

$$T_{ij} = \frac{x_{ij}}{x_i} \qquad (26)$$

Note that the equilibrium probability vector π computed during the Gaussian mixture model fitting is not respected during this step.

6.5.4 Fitting a maximum likelihood HMM

The HMM model parameters $\mathbf{\Theta} \equiv \{\mathbf{T},\mathbf{E}\}$ are fit to the observed data O through use of the expectation-maximization (EM) algorithm [26]. This is an iterative procedure, where the model parameters are subsequently refined through successive iterations.

During each iteration, the Baum-Welch algorithm [13] is used to compute, for each trace n, $\Xi^{(n)} \equiv (\xi_{tij}^{(n)})$, which represents the probability that the system transitions from hidden state i at time $t-1$ to hidden state j at time t, and $\gamma_{ti}^{(n)}$, the probability that the system occupied state i at time t. This is accomplished by first executing the forward algorithm, which proceeds (suppressing superscripts) as

$$\alpha_{tj} = \begin{cases} p_j \varphi(o_0 \mid e_j) & t = 0 \\ \varphi(o_t \mid e_j) \sum_{i=1}^{M} \alpha_{(t-1)i} T_{ij} & t = 1,...,L \end{cases} \qquad (27)$$

followed by the backward algorithm,

$$\beta_{ti} = \begin{cases} 1 & t = L \\ \sum_{j=1}^{M} T_{ij} \varphi(o_{t+1} \mid e_j) \beta_{(t+1)j} & t = (L-1),...,0 \end{cases} \qquad (28)$$

The $L \times M \times M$ matrix $\mathbf{\Xi}$ is then computed for $t = 0,...,(L-1)$ as

$$\xi_{tij} = \alpha_{ti} \varphi(o_{t+1} \mid e_i) T_{ij} \beta_{(t+1)j} / \sum_{i=1}^{M} \alpha_{Ti} \qquad (29)$$

$$\gamma_{ti} = \sum_{j=1}^{M} \xi_{tij} \qquad (30)$$

In practice, the logarithms of these quantities are computed instead to avoid numerical underflow.

The aggregate matrix of expected transition counts $\mathbf{C} \equiv (c_{ij})$ is then computed from the $\Xi^{(n)}$ computed for all traces as

$$c_{ij} = \sum_{n=1}^{N} \sum_{t=0}^{L^{(n)}-1} \xi_{tij}^{(n)} \qquad (31)$$

This count matrix is used to update the maximum-likelihood transition matrix **T** using the method of Prinz et al. [25] described in the previous section.

The state observable distribution parameters **E** are then updated from the γ_{ti}, using all trace data, as

$$\begin{aligned}\mu_i' &= \frac{\sum_{n=1}^{N}\sum_{t=0}^{L} o_t^{(n)}\gamma_{ti}^{(n)}}{\sum_{n=1}^{N}\sum_{t=0}^{L^{(n)}}\gamma_{ti}^{(n)}} \\ \sigma_i'^2 &= \frac{\sum_{n=1}^{N}\sum_{t=0}^{L^{(n)}}(o_t^{(n)}-\mu_i')^2\gamma_{ti}^{(n)}}{\sum_{n=1}^{N}\sum_{t=0}^{L^{(n)}}\gamma_{ti}^{(n)}}\end{aligned} \tag{32}$$

Once the model parameters have been fitted by iteration of the above update procedure to convergence (which may only converge to a local maximum of the likelihood), the most likely hidden state sequence can be determined given the observations O and the MLE model $\hat{\Theta}$ using the Viterbi algorithm [14]. Like the forward-backward algorithm employed in the Baum-Welch procedure, the Viterbi algorithm also has a forward recursion component that is applied independently to each trace n (again suppressing superscripts),

$$\begin{aligned}\varepsilon_{jt} &= \begin{cases}\rho_j\varphi(o_t \mid e_j) & t=0 \\ \varphi(o_t \mid e_j)\max_i \varepsilon_{i(t-1)}T_{ij} & t=1,...,L\end{cases} \\ \Phi_{jt} &= \begin{cases}1 & t=0 \\ \arg\max_i \varepsilon_{i(t-1)}T_{ij} & t=1,...,L\end{cases}\end{aligned} \tag{33}$$

as well as a reverse reconstruction component to compute the most likely state sequence $\hat{S}$,

$$\hat{s}_t = \begin{cases}\arg\max_i \varepsilon_{it} & t=L \\ \Phi_{\hat{s}_{t+1}(t+1)} & t=(L-1),...,0\end{cases} \tag{34}$$

6.5.5 Sampling from the posterior of the BHMM

Sampling from the posterior of the BHMM (Eq. 9) proceeds by rounds of Gibbs sampling, where each round consists of an update of the augmented model parameters {**T**,**E**,**S**} by sampling

$$\mathbf{S}' \mid \mathbf{T},\mathbf{E},\mathbf{O} \sim P(\mathbf{S}' \mid \mathbf{T},\mathbf{E},\mathbf{O})$$
$$\mathbf{T}' \mid \mathbf{S}' \sim P(\mathbf{T}' \mid \mathbf{S}')$$
$$\mathbf{E}' \mid \mathbf{S}',\mathbf{O} \sim P(\mathbf{E}' \mid \mathbf{S}',\mathbf{O})$$

where the conditional probabilities are given by Eq. 12.

6.5.6 Updating the hidden state sequences

In the first part of each sampling round, we use a modified form of the Viterbi process to generate an independent sample of the hidden state history S given the transition probabilities **T**, state observable distribution parameters **E**, and observed data O. Like the Viterbi scheme, a forward recursion algorithm (Eq. 33) is applied to each trace $o^{(n)}$ separately, but instead of computing the most likely state history on the reverse pass, a new state history is drawn from the distribution $P(s|o,\mathbf{T},\mathbf{E})$. The forward recursion is identical to the Viterbi case (Eq. 33):

$$\varepsilon jt = \begin{cases} \rho_j \varphi(o_t \mid e_j) & t = 0 \\ \varphi(o_t \mid e_j) \max_i \varepsilon_{i(t-1)} T_{ij} & t = 1,...,L \end{cases} \qquad (35)$$

The hidden state sequence s_t corresponding to observation trace $o^{(n)}$ is then sampled according to $P(s_t|s_t+1, \ldots, s_L)$ in order from $t = L$ down to $t = 0$:

$$P(s_t \mid s_{t+1},...,s_L) \propto \begin{cases} \varepsilon_{s_t t} & t = L \\ \varepsilon_{s_t t} T_{s_t s_{t+1}} & t = (L-1),...,0 \end{cases} \qquad (36)$$

6.5.7 Updating the transition probabilities

If no detailed balance constraint is used and the prior $P(\mathbf{T})$ is Dirichlet in each row of the transition matrix **T**, it is possible to generate an independent sample of the transition matrix from the conditional distribution $P(\mathbf{T}'|\mathbf{S}')$ by sampling each row of the transition matrix from the conjugate Dirichlet posterior using the transition counts from the sampled state sequence $\mathbf{S}'$ [23]. However, because physical systems in the absence of energy input through an external driving force should satisfy detailed balance, we make use of this constraint in updating our transition probabilities, since this has been demonstrated to substantially reduce parameter uncertainty in the data-limited regime [23].

The transition matrix is updated using the reversible transition matrix sampling scheme of Noé [23, 27]. Here, an adjusted count matrix $\mathbf{C} \equiv (c_{ij})$ is computed using the updated hidden state sequence $\mathbf{S}'$:

$$c_{ij} = b_{ij} + \sum_{n=1}^{N} \sum_{t=1}^{L} \delta_{i,s_{n(t-1)}} \delta_{j,s_{nt}} \qquad (37)$$

where the Kronecker $\delta_{i,j} = 1$ if $i = j$ and zero otherwise, and $\mathbf{B} \equiv (b_{ij})$ is a matrix of prior pseudocounts, which we take to be zero following the work of Noé et al. [10]. Using the adjusted count matrix **C**, a Metropolis-Hastings Monte Carlo procedure [28] is used to update the matrix and produce a new sample from $P(\mathbf{T}' \mid \mathbf{S}')$. Two move types are attempted, selected with equal probability, and 1000 moves are attempted to generate a

new sample $\mathbf{T}'$ that is approximately uncorrelated from the previous $\mathbf{T}$. Prior to starting the Monte Carlo procedure, the vector of equilibrium probabilities for all states π is computed according to

$$\mathbf{T}\pi = \pi \qquad (38)$$

The first move type is a reversible element shift. A pair of states (i,j), $i \neq j$, are selected with uniform probability, and a random number Δ is selected uniformly over the interval

$$\Delta \in \left[\max\left(-T_{ii}, -\frac{\pi_j}{\pi_i} T_{jj} \right), T_{ij} \right]$$

The changed elements in the proposed transition matrix $\mathbf{T}'$ are then given by:

$$T'_{ij} = T_{ij} - \Delta \quad ; \quad T'_{ji} = T_{ji} - \frac{\pi_i}{\pi_j}\Delta$$

$$T'_{ii} = T_i + \Delta \quad ; \quad T'_{jj} = T_{jj} + \frac{\pi_i}{\pi_j}\Delta$$

This move is accepted with probability

$$P_{accept}(\mathbf{T} \to \mathbf{T}') = \min\left\{ 1, \sqrt{\frac{(T'_{ij})^2 + (T'_{ji})^2}{(T_{ij})^2 + (T_{ji})^2}} \times \left(\frac{T'_{ii}}{T_{ii}}\right)^{c_{ii}} \left(\frac{T'_{ij}}{T_{ij}}\right)^{c_{ij}} \left(\frac{T'_{jj}}{T_{jj}}\right)^{c_{jj}} \left(\frac{T'_{ji}}{T_{ji}}\right)^{c_{ji}} \right\} \qquad (39)$$

This move will leave the vector of stationary probabilities π unchanged.

The second move type is a row shift. A row i of $\mathbf{T}$ is selected with uniform probability, and a random number α chosen uniformly over the interval

$$\alpha \in \left[0, \frac{1}{1 - T_{ii}} \right]$$

and used to update row i of $\mathbf{T}$ according to

$$T'_{ij} = \begin{cases} \alpha T_{ij} & j = 1,...,M, \quad j \neq i \\ \alpha(T_{ii} - 1) + 1 & j = i \end{cases} \qquad (40)$$

This move is accepted with probability

$$P_{accept}(\mathbf{T} \to \mathbf{T}') = \min\left\{ 1, \alpha^{(M-2)} \alpha^{(c_{i*} - c_{ii})} \left(\frac{1 - \alpha(1 - T_{ii})}{T_{ii}} \right)^{c_{ii}} \right\} \qquad (41)$$

The row shift operation will change the stationary distribution of π', but it may be efficiently updated:

$$\pi' = \frac{\pi_i}{\pi_i + \alpha(1-\pi_i)} \quad ; \quad \pi'_j = \frac{\alpha\pi_j}{\pi_i + \alpha(1-\pi_i)}$$

Since this update scheme is incremental, it will accumulate numerical errors over time that cause the updated π to drift away from the stationary distribution of the current transition matrix. To avoid this, π is recomputed from the current sample of the transition matrix in regular intervals (here, every 100 sampling steps).

6.5.8 Updating the observable distribution parameters

Following the update of the transition matrix **T**, the observable distribution parameters **E** are updated by sampling **E** from the conditional probability $P(\mathbf{E}'|\mathbf{S}',O)$. The conditional probability for the observable distribution parameters for state m, denoted e_m, is given in terms of the output model $\varphi(o|e)$ by Bayes' theorem:

$$P(\mathbf{E} \mid \mathbf{O},\mathbf{S}) = \left[\prod_{n=1}^{N}\prod_{t=0}^{L^{(n)}} \varphi(o_{nt} \mid e_{s_t^{(n)}}\right] P(\mathbf{E}) \qquad (42)$$

An important choice must be made with regards to the prior, $P(\mathbf{E})$. If the prior is chosen to be composed of independent priors for each state, as in

$$P(\mathbf{E}) = \prod_{m=1}^{M} P(e_m) \qquad (43)$$

then the full BHMM posterior (Eq. 9) will be invariant under any permutation of the states. This behavior might be undesirable, as the states may switch labels during the posterior sampling procedure; this will require any analysis of the models sampled from the posterior to account for the possible permutation symmetry in the states. On the other hand, breaking this symmetry (e.g., by enforcing an ordering on the state mean observables) can artificially restrict the confidence intervals of the states, which might additionally complicate data analysis.

Here, we make the choice that the prior be separable (Eq. 43), which has the benefit of allowing the conditional probability for **E** (Eq. 42) to be decomposed into a separate posterior for each state. For each state m, collect all the observations $o_t^{(n)}$ whose updated hidden state labels $s_t^{(n)'} = m$ into a single dataset $o \equiv \{o_n\}_{n=1}^{N_m}$, where N_m is the total number of times state m is visited, for the purposes of this update procedure. Then, the observable parameters **e** for this state are given by

$$P(e \mid o) = P(o \mid e)P(e) = \left[\prod_{n=1}^{N_m} \varphi(o_n \mid e)\right] P(e) \qquad (44)$$

In the application presented here, we use a Gaussian output model (Eq. 3) for the state observable distributions $P(o|\mathbf{e})$, where $\mathrm{e} \equiv \{\mu,\sigma^2\}$, with μ the state mean observable and σ^2 the variance (which will include both the distribution of the observable characterizing the state and any broadening from measurement noise). Other models (including

multidimensional or multimodal observation models) are possible, and require replacing only the observation model $\varphi(o|\mathbf{e})$ and corresponding prior $P(\mathbf{e})$.

We use the (improper) Jeffreys prior [29] which has the information-theoretic interpretation as the prior that maximizes the information content of the data [30], (suppressing the state index subscript m),

$$P(e) \propto \sigma^{-1} \qquad (45)$$

which produces the posterior

$$P(e \mid o) \propto \sigma^{-(N+1)} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(o_n-\mu)^2\right] \qquad (46)$$

The conditional distribution of the mean μ is then given by

$$P(\mu \mid \sigma^2, \mathbf{o}) \equiv \frac{P(\mu,\sigma^2 \mid \mathbf{o})}{\int d\mu P(\mu,\sigma^2 \mid \mathbf{o})} \propto \exp\left[-\frac{1}{2(\sigma^2/N)}(\mu-\hat{\mu})^2\right] \qquad (47)$$

where $\hat{\mu}$ is the sample mean for o, the samples in state m,

$$\hat{\mu} \equiv \frac{1}{N}\sum_{n=1}^{N} o_n \qquad (48)$$

This allows us to update μ according to

$$\mu' \sim \mathcal{N}(\hat{\mu}, \sigma^2/N) \qquad (49)$$

The conditional distribution of the variance σ^2 is given by

$$\begin{aligned} P(\sigma^2 \mid \mu, \mathbf{o}) &= \frac{p(\mu,\sigma^2 \mid \mathbf{o})}{\int d\sigma^2 p(\mu,\sigma^2 \mid \mathbf{o})} \\ &\propto \sigma^{-(N+1)} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(o_n-\mu)^2\right] \qquad (50) \\ &\propto \sigma^{-(N+1)} \exp\left[-\frac{N\hat{\sigma}^2}{2\sigma^2}\right] \end{aligned}$$

where the quantity $\hat{\sigma}^2$, which is not in general identical to the sample variance, is given by

$$\hat{\sigma}^2 \equiv \frac{1}{N}\sum_{n=1}^{N}(o_n-\mu)^2 \qquad (51)$$

A convenient way to update $\sigma^2|\mu,o$ is to sample a random variant y from the chi-square distribution with $N-1$ degrees of freedom,

$$y \sim \chi^2(N-1) \qquad (52)$$

and then update σ^2 as

$$\sigma'^2 = \frac{N\hat{\sigma}^2}{y} \qquad (53)$$

Note that μ and σ^2 can be updated in either order, but the updated values of μ or σ^2 must be used in sampling the not-yet-updated σ^2 or μ, and vice-versa.

6.6 References

1. Ritort, F., *Single-molecule experiments in biological physics: methods and applications.* J. Phys.: Condens. Matter, 2006. **18**: p. R531-R583.
2. Smith, G.J., et al., *A large collapse-state RNA can exhibit simple exponential single molecule dynamics.* J. Mol. Biol., 2008. **378**: p. 941-951.
3. Qu, X., et al., *Single-molecule nonequilibrium periodic Mg$^2+$-concentration jump experiments reveal details of the early folding pathways of a large RNA.* Proc. Natl. Acad. Sci. USA, 2008. **105**: p. 6602-6607.
4. Min, W., et al., *Fluctuating enzymes: Lessons from single-molecule studies.* Acc. Chem. Res., 2005. **38**: p. 923-931.
5. Harms, G., G. Orr, and H.P. Lu, *Probing ion channel conformational dynamics using simultaneous single-molecule ultrafast spectroscopy and patch-clamp electric recording.* Appl. Phys. Lett., 2004. **85**(10): p. 1792-1794.
6. Li, P.T.X., et al., *Probing the mechanical folding kinetics of TAR RNA by hopping, force-jump, and force-ramp methods.* Biophys. J., 2006. **90**: p. 250-260.
7. Woodside, M.T., et al., *Direct measurement of the full, sequence-dependent folding landscape of a nucleic acid.* Science, 2006. **314**: p. 1001-1004.
8. Mickler, M., et al., *Revealing the bifurcation in the unfolding pathways of GFP by using single-molecule experiments and simulations.* Proc. Natl. Acad. Sci. USA, 2007. **104**(51): p. 20268-20273.
9. Baldwin, A.J. and L.E. Kay, *NMR spectroscopy brings invisible protein states into focus.* Nature Chem. Biol., 2009. **5**(11): p. 808-814.
10. No\'e, F., et al., *Constructing the full ensemble of folding pathways from short off-equilibrium simulations.* Proc. Natl. Acad. Sci. USA, 2009. **106**: p. 19011-19016.
11. Chodera, J.D., et al., *A robust approach to estimating rates from time-correlation functions.* in preparation, 2010.
12. Eddy, S.R., *What is a hidden Markov model?*, in *Nat Biotechnol*. 2004. p. 1315-6.
13. Baum, L.E., et al., *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains.* Ann. Math. Statist., 1970. **41**: p. 164-171.
14. Viterbi, A.J., *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.* IEEE Trans. Info. Theory, 1967. **13**: p. 260-269.
15. Geman, S. and D. Geman, *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images.* IEEE Trans. Pattern Anal., 1984. **6**: p. 721-741.
16. Liu, J.S., *Monte Carlo strategies in scientific computing*. 2002, New York: Springer-Verlag.
17. Chandler, D., *Statistical mechanics of isomerization dynamics in liquids and the transition state approximation.* J. Chem. Phys., 1978. **68**(6): p. 2959-2970.

18. Adams, J.E. and J.D. Doll, *Dynamical aspects of precursor state kinetics.* Surface Science, 1981. **111**: p. 492-502.
19. Voter, A.F. and J.D. Doll, *Dynamical corrections to transition state theory for multistate systems: Surface self-diffusion in the rare-event regime.* J. Chem. Phys., 1985. **82**(1): p. 80-92.
20. Ryan, K.J., *Estimating expected information gains for experimental designs with application to the random fatigue-limit model.* Journal of Computational and Graphical Studies, 2003. **12**: p. 585-603.
21. Berg-Sørensen, K. and H. Flyvbjerg, *Power spectrum analysis for optical tweezers*, in *REVIEW OF SCIENTIFIC INSTRUMENTS*. 2004. p. 594-612.
22. van Kampen, N.G., *Stochastic processes in physics and chemistry*. 1997: Elsevier.
23. No\'e, F., *Probability distributions of molecular observables computed from Markov models.* J. Chem. Phys., 2008. **128**: p. 244103.
24. Bilmes, J.A., *A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models*. 1998, University of California, Berkeley.
25. Prinz, J.-H., et al., *Markov models of molecular kinetics: Generation and validation.* J. Chem. Phys., 2010: p. submitted.
26. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum-likelihood from incomplete data via the EM algortihm.* J. Royal Statist. Soc. B, 1977. **39**: p. 1-38.
27. Chodera, J.D. and F. No\'e, *Probability distributions of molecular observables computed from Markov models: II. Uncertainties in observables and their time-evolution.* J. Chem. Phys., 2010. **to appear**.
28. Hastings, W.K., *Monte Carlo sampling methods using Markov chains and their applications.* Biometrika, 1970. **57**(1): p. 97-109.
29. Jeffreys, H., *An invariant form for the prior probability in estimation problems.* Proc. Royal Soc. A, 1946. **186**: p. 453-461.
30. Goyal, P., *Prior probabilities: An information-theoretic approach*, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. 2005, American Institute of Physics. p. 366-373.