**Title**

Single molecule studies of equilibrium and non-equilibrium processes during cotranslational protein folding

**Permalink**

https://escholarship.org/uc/item/3gw733xh

**Author**

Alexander, Lisa M

**Publication Date**

2018

Peer reviewed|Thesis/dissertation

**Single molecule studies of equilibrium and non-equilibrium processes during cotranslational protein folding**

by

Lisa M Alexander


A dissertation submitted in partial satisfaction

of the requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley


Committee in charge:

Professor Carlos Bustamante, Chair
Professor Jamie Doudna Cate
Professor Eva Nogales


Fall 2018

**Single molecule studies of equilibrium and non-equilibrium processes during cotranslational protein folding**

# Abstract

Single molecule studies of equilibrium and non-equilibrium processes during cotranslational protein folding

by

Lisa M Alexander
Doctor of Philosophy in Chemistry
University of California, Berkeley
Professor Carlos Bustamante, Chair

Protein folding is a unique biological problem that single molecule techniques have had success in tackling to gain detailed insight into the folding landscape of purified proteins in solution. The *in vitro* post-translational refolding pathway, however, may not accurately recreate the pathway of folding when a protein is being synthesized by the ribosome. Folding on the surface of the ribosome is difficult to probe since the ribosome itself is composed of both RNA and proteins, so is sensitive to many of the same techniques used to probe the nascent chain. Our lab previously developed a technique to use optical tweezers to watch the nascent chain fold at the single molecule level with minimal perturbation to the ribosome, which showed that the ribosome slows down folding kinetics. Continuing this experimental setup, in this work we explore the folding of a multi-domain calcium-binding protein, calerythrin, on the ribosomal surface both in static stalled complexes and in real-time elongation conditions.

Our results show that off the ribosome, the refolding pathway proceeds via the C-domain, while N-domain folding is the rate-limiting step to the folded state. In contrast, an alternate intermediate that is a misfolded, unproductive state is accessible in a stalled ribosome complex. When we further investigated this misfolded state, we found that it is accessible to partially synthesized, truncated proteins off the ribosome as well. The ribosome decreases the folding rate, as previously reported for native folds, but it also acts to increase the unfolding rate, a combined effect that promotes an escape from misfolding.

Although this effect is interesting, the kinetics of folding, even when delayed, are still quite fast relative to synthesis. These different timescales have been interpreted in the literature as implying that folding during real-time elongation occurs at equilibrium. However, to directly test if this assumption is in fact relevant, we conducted measurements of folding while real-time elongation was ongoing. This gave us the unexpected result that folding on the ribosome during active protein synthesis is not at equilibrium despite the disparate timescales involved. In fact, there is a non-equilibrium relaxation time of 71 seconds prior to the onset of misfolding. In this time, the protein remains unfolded and is incapable of fully exploring its energy landscape. After equilibration occurs, the folding observed matches stalled complexes and is completely reversible and at equilibrium despite further chain elongation or changes to the force. This long delay before folding provides a mechanism for reducing the probability of misfolding significantly by allowing synthesis to terminate (likely for small proteins) or by providing additional time for chaperone recruitment (likely for larger proteins). This study gives us further

insight into the importance of the ribosome not just for synthesis but also for regulating protein function, and opens up interesting questions about further interplay between elongation and protein folding.

We also developed a modified experimental design that provides insight into folding through two parallel channels– optical tweezers and fluorescence. In this assay, fluorophores within the protein can report on local conformational changes independent of the signal from the optical tweezers via FRET, allowing us to correlate global and local changes without using indirect measures of folding intermediates such as mutations. We have conducted proof-of-concept experiments for this design using calmodulin, a eukaryotic signaling protein, that aim to understand the order of events in early folding and to identify any correlation in secondary structure formation between the two domains. Our preliminary results show that the design is feasible and suggest promising avenues for future development to definitively answer the questions posed. In the future, this assay could also be combined with our experiments on ribosome-bound nascent chains to look at dynamics of the unfolded state.

# Table of Contents

# List of Figures

**Chapter 4**

# List of Tables

# Acknowledgements

I would like to thank my family for their support throughout my dissertation. I would also like to thank Carlos for his support. I was very fortunate to have Ignacio Nacho Tinoco as a partial co-advisor until he passed away. I remember his grounded, pragmatic approach to science that was focused and concise. I especially appreciate having both Carlos and Nacho help me develop my work through the early years, since they bring vastly different perspectives and I learned to see questions through both lenses.

I feel particularly fortunate to have worked alongside an inspiring set of scientists who have mentored me and helped me through my journey. I would especially like to thank Daniel Goldman, Liang Meng Wee, Varsha Desai, and Shannon Yan, who were my closest collaborators and colleagues. I appreciate their insight and perspective to all things ribosome-related, as well as their personal friendship. Some of the best times in the laboratory were working through a problem together and brainstorming ideas and questions. They have challenged me to develop my ideas, present my work, and always charge forward.

Other members of the lab I worked with less closely, but were still indispensable. I would like to thank Maya Sen, Pim Dangulwanich, Sam Leachman, Shixin Liu, Ronen Gabizon, Cesar Diaz, and Cristhian Canari.

I have had the opportunity to teach and work with a variety of undergraduates during my PhD as well. I would like to recognize two of my own undergraduates, Apurva Shah and Abigail Keller, in particular. Although I may have taught them in the laboratory, they helped me grow as a scientist and mentor. I also would like to acknowledge Vianca Vianzon and Thy Pham, who I did not directly mentor. I am grateful for their curiosity, perspective, and drive.

Lastly, I would like to thank my peers and friends outside of the laboratory. Graduate school was a long journey, and I would not want to make it without any of you. I am grateful for the opportunity that brought us all together. Thank you to my boyfriend Alec White as well as the rest of my irreplaceable Berkeley family: Noelle Catarineu, Leonel Cosio, Scott Ellis, Samia Hamed, Monatrice Lam, Royce Lam, Michelle Leuenberger, Jacob Olshansky, Walter Ralston, Mark Shapero, Jean Van Buren, and Jon Witte.

# Chapter 1.
## Introduction

### 1.1 Protein synthesis

Proteins are biological polymers that carry out many cellular functions such as signal transduction, catalysis, and regulation. There are twenty amino acids which can be incorporated as monomers into the heteropolymeric polypeptide chain. These amino acids have different side chains that confer different size, charge, hydrophobicity, etc., leading to a wide gamut of chemical and biological properties. A protein's function is innately tied to its sequence of amino acids that yields a specific three-dimensional structure. The process of forming the three-dimensional structure from the polypeptide (folding) will be further discussed in section 1.2.

The sequence of amino acids of a protein is encoded within the cellular DNA. The decoding process of the DNA sequence is the central dogma of biology: DNA is transcribed into a complementary strand of messenger RNA (mRNA), and that mRNA is translated into a protein. The process of protein synthesis (translation) is carried out by the ribosome. The ribosome itself is an enormous molecular machine, 2.3 MDa in size, composed of proteins and ribosomal RNA (rRNA). The ribosome can be further broken into two subunits– the large subunit (50S), which is responsible for catalysis, and the small subunit (30S), which is responsible for mRNA binding and decoding. At the interface between the two subunits are three transfer RNA (tRNA) binding sites: acyl (A), peptidyl (P), and exit (E) (Figure 1.1)
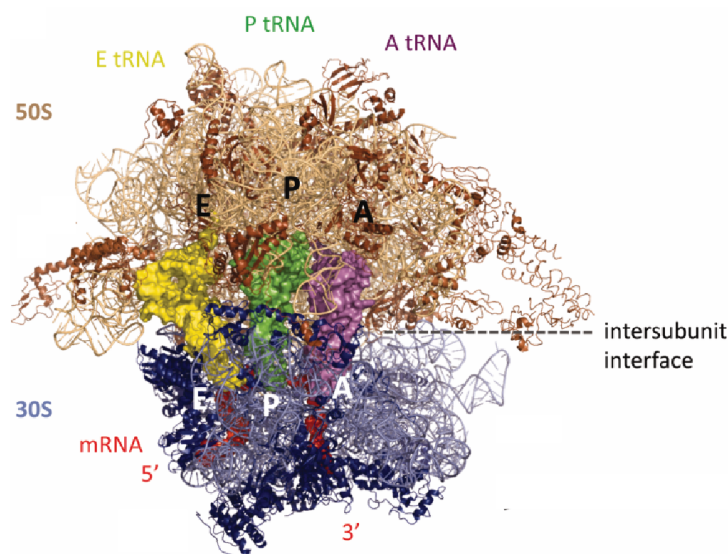
**Figure 1.1: The structure of the bacterial 70S ribosome** From reference (*1*), PDB: 3I8G, 3I8F, 2WRJ. The 50S subunit is brown/tan (protein/rRNA) and the 30S subunit is blue/periwinkle (protein/rRNA). The tRNAs are shown as yellow, green, and purple in the E, P, and A sites, respectively.

A schematic of translation is overviewed in Figure 1.2. There are 3 phases of translation: initiation, elongation, and termination.

During initiation, ribosomes recognize mRNAs through a conserved initiation site of an AUG start codon and a conserved Shine-Dalgarno sequence (AGGAGG) that is 5-7 nucleotides upstream of the start codon in translation initiation. Initiation is dependent on three initiation factors (IF1, IF2, IF3) and a specific initiator tRNA that is charged with formyl-methionine. IF1 and IF3 bind to 30S subunits and ensure subunit dissociation (*2*). This complex then binds with IF2 and fMet-tRNA. After fMet-tRNA binding, the mRNA binds to the 30S subunit. The small subunit of the ribosome recognizes the Shine-Dalgarno sequence of the mRNA through base-pairing interactions with an anti-Shine-Dalgarno sequence in the rRNA, and this helps positions the start codon in the P site. Once the start codon is in the P-site, the fMet-tRNA can form a codon-anti-codon pairing with the mRNA and IF3 dissociates. The 50S subunit is then able to bind and IF1 and IF2 dissociate (*2*). After this process, the ribosome contains a correctly positioned mRNA in reading frame and a P-site initiator tRNA (Figure 1.1A), and the ribosome is ready to enter elongation.

## A: Initiation

E P A

fMet tRNA

Anti-Shine Dalgarno

AUG

## B: Elongation

EF-G•GDP

Ternary complex

Accommodation

EF-Tu•GDP

mRNA movement

Translocation

Peptidyl transfer

EF-G•GTP

## C: Termination

RF1 or RF2, EF-G, RF3

Stop codon (UAA, UAG, UGA)
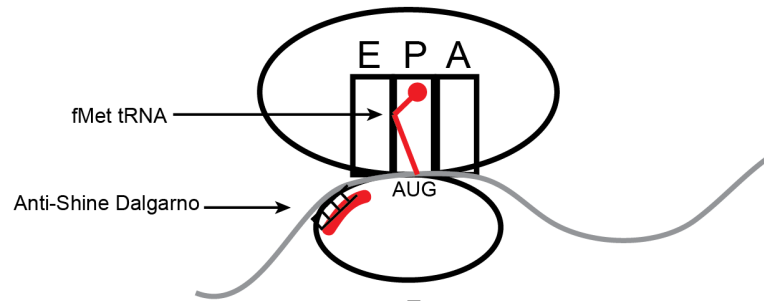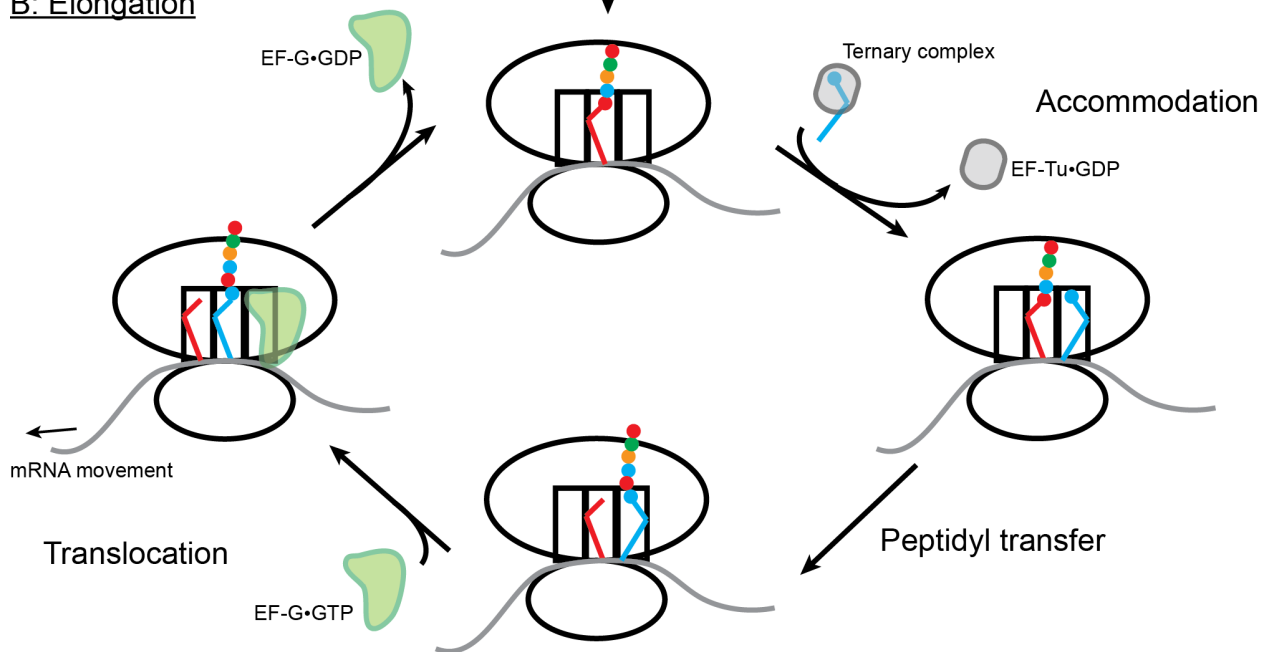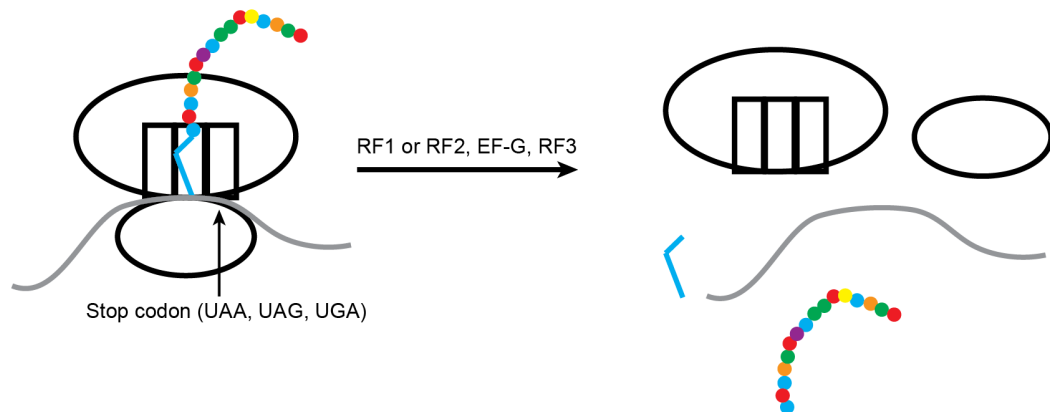
**Figure 1.2: Schematic of the translation cycle.** Each of the main steps of elongation is noted,

although they are each composed of many substeps.

During elongation, each amino acid is added sequentially via the decoding of 3 nucleotide segments (codons) of the mRNA, where each codon corresponds to a unique amino acid (Figure 1.1B). One full cycle is as follows:

1. tRNA-synthetases recognize a tRNA and its partner amino acid, and acylate the tRNA using energy from the hydrolysis of ATP to AMP, forming an ester-linkage between the tRNA 3' end and the amino acid carboxyl group.

2. Elongation factor Tu (EF-Tu) binds the acylated tRNA and GTP to form the ternary complex.

3. The ternary complex binds to the A site of the ribosome and the mRNA with base pairing between the tRNA decoding loop and the mRNA codon.

4. GTP hydrolysis in the ternary complex allows for accommodation of the correct acyl-tRNA into the ribosome. EF-Tu•GDP dissociates. Incorrect mRNA-tRNA pairings have slower hydrolysis rates and will dissociate prior to accommodation (*3*).

5. Peptidyl transfer occurs whereby the amino group of the A-site tRNA amino acid attacks the ester linkage of the P site tRNA and its amino acid chain. This transfers the entire chain to the A site tRNA and adds one amino acid to the chain. The P site tRNA is deacylated.

6. Translocation occurs via EF-G binding and GTP hydrolysis, whereby the A site tRNA and its chain moves to the P site, the deacylated P site tRNA moves to the E site, and concomitantly the mRNA moves 3 nucleotides (1 codon) down register in the ribosome, placing the next codon in the A site for decoding.

Synthesis ends upon encountering a stop codon (UAA, UAG, or UGA), which will then recruit release factors (RF1 or RF2). The completed protein is released through an EF-G and RF3 dependent mechanism, and the ribosome is recycled (Figure 1.1C). In bacteria, if the mRNA lacks a stop codon, there is a separate rescue system distinct from RF1- and RF2-mediated release. This rescue system uses an ssrA-tag encoded by a transfer-messenger RNA (tmRNA). The tmRNA binds in the empty A-site as both the mRNA and tRNA, and switches the ribosome onto an alternate open reading frame within the tmRNA, the translation of which adds a degradation tag. We take advantage of the independence of stop-codon mediate release and blunt-end mRNA rescue to produce stalled ribosome complexes which are immune to RF1 and RF2 activity (Chapter 3).

The whole process of elongation occurs at rates of 10-20 amino acids/second in bacteria (*4*). Most proteins are at least 100 amino acids, and so synthesis will take a few seconds. This relatively slow timescale is important when considering protein folding in a cellular context. However, more is known about protein folding from in vitro studies of purified proteins so this will be discussed first.

**1.2 Protein folding in solution**

The hierarchy of protein structure has four levels: primary, secondary, tertiary, quaternary. The primary structure is the sequence of amino acids of a given protein. Secondary structures are

local structural formations of alpha helices and beta sheets. The tertiary structure results from the arrangement of secondary structure elements into a stable form. The tertiary structure is the state we are most interested in as it corresponds to fully folded, active protein. Quaternary structure is present for multimeric proteins, and refers to the arrangement of each subunit in the overall structure.

Protein folding is the process by which polypeptides form the specific three-dimensional tertiary structures required for their cellular function. Protein structures can be determined with atom-level precision using X-ray crystallography, cryo-EM, or NMR spectroscopy. These structural studies allowed the classification of protein structures into common motifs and domains, where a domain is a commonly occurring folded structure across many. However, these techniques are probing the static, final folded structure and do not give insight into the process of folding itself nor the intermediate forms the protein may adopt en route to the folded state.

The process of folding can be studied by denaturing the protein through some mechanism (chemical, heat, or mechanical denaturation) and then following refolding as the denaturant is removed. The read-outs are often indirect, for example fluorescence, circular dichroism, heat capacity, or, in optical tweezers, the length of the protein. Since these are indirect, a combination of experimental techniques or mutants needs to be employed to determine key intermediates or regions of interaction. The overall read out of the folding will often give a relatively simple apparent pathway of one or a few transition states. However, the true landscape of the protein is very rugged and multidimensional. The folding may explore many conformations transiently, only a few of which are detectable (Figure 1.3)
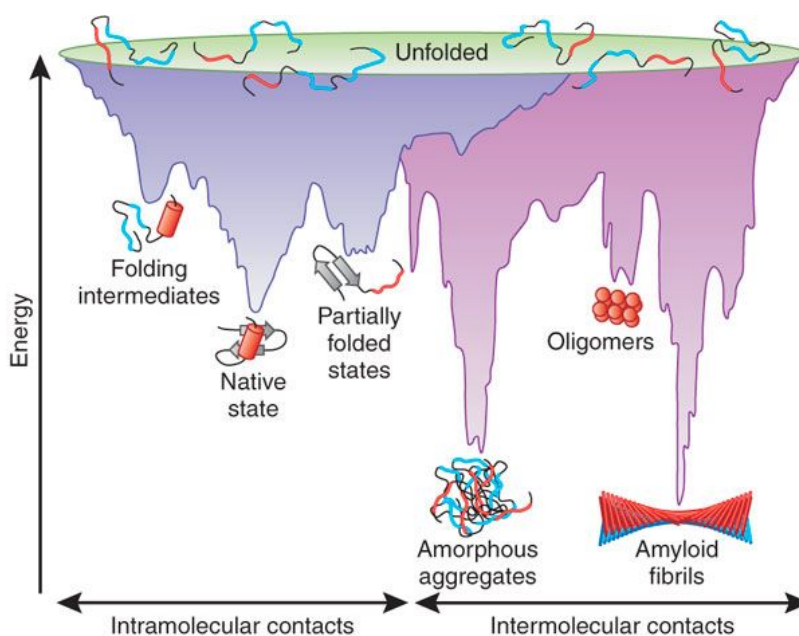


**Figure 1.3: Model of a protein folding landscape** From reference (*5*). In this schematic, there is a broad pool of unfolded conformations that are funneled towards the native state on a rugged landscape. Many shallow minima are less than or similar order as thermal energy. Most techniques would only capture deeper minima like the folding intermediates or partially folded

states. There are competing pathways to productive folding of aggregation or fibril formation. These larger structures may be thermodynamic minima but are often not functionally desired. Chaperones may steer unfolded proteins away from these states.

Many of the model proteins used to study the folding process are single domain proteins with a relatively simple apparent folding pathway. Often, a single domain folds cooperatively in a single step. Depending upon the denaturant used and the readout of the denaturing reaction, the folding, inherently a multidimensional process, is measured along a one-dimensional projection of the true landscape. Thus, folding studies always provide a limited description of all the degrees of freedom of the actual process. This limitation is shared by all techniques; however, the rate-limiting transition state may be the same when employing different techniques (for example, the same transition state was observed using either urea or optical tweezers as denaturants of the SH3 domain, for certain force geometries (6)). Therefore, despite their inherent limitations, it is still possible to gain useful information about the folding rate, protein stability, transition state, potential on-pathway intermediates, or competing off-pathway states from protein folding experiments.

The single domain model proteins are useful for studying hydrophobic collapse and secondary or tertiary structure formation. However, many proteins in the cell are multidomain proteins, which fold reliably *in vivo* but their folding cannot be adequately reproduced *in vitro*. Why are multidomain proteins more difficult to study *in vitro*, and what aspects of cellular folding do our experiments fail to capture? The answers to these questions are inherent in the protein itself: since multidomain proteins contain separately folding "units", when the entire polypeptide is denatured and allowed to refold, the chain can interact with itself in a non-native way, crossing domain boundaries and forming misfolded states (7). Moreover, two or more separate chains can interact, forming aggregated complexes. These complications may be mitigated *in vivo* due to the intersection of the folding process and protein synthesis.


## 1.3 Protein folding on the ribosome

How does the sequential and vectorial process of polypeptide synthesis affect protein folding? This question can be addressed from a kinetic as well as entropic viewpoint. Kinetically, protein synthesis occurs at 12-18 amino acids per second (4). This means that even a small protein of 100 amino acids will require 5-10 seconds for synthesis to be completed. Folding of secondary structures, on the other hand, takes place on a much shorter time scale, from micro- to milliseconds. Thus, initial regions of the polypeptide chain will be in quasi-equilibrium and able to fold before synthesis is complete. However, the ribosome has a narrow tunnel extending from the peptidyl transfer center to the solvent which can accommodate about 33 amino acids (8) (Figure 1.4).
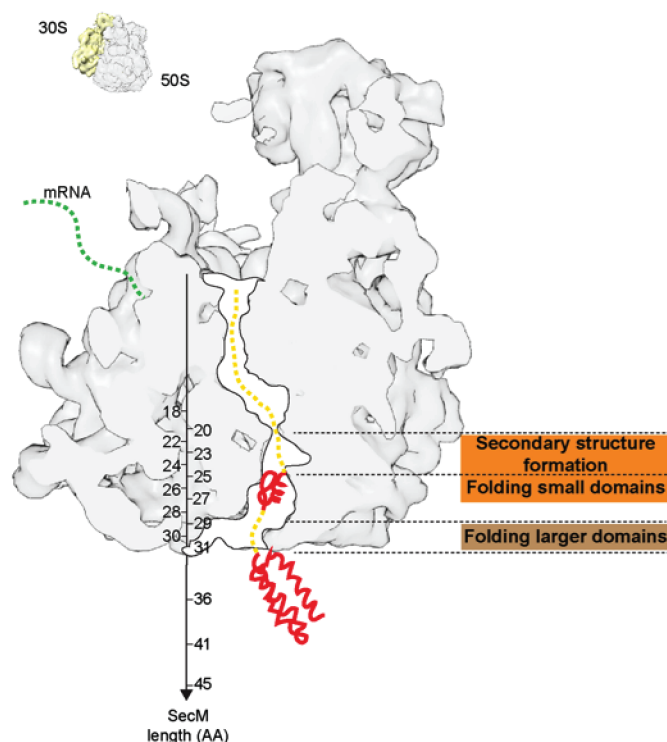
**Figure 1.4: The structure of the ribosome exit tunnel.** A schematic of the ribosome exit tunnel, from reference (*9*). The tunnel extends from the peptidyl transfer center (PTC) to the solvent in 31-33 amino acids. There is a constriction in the tunnel which forces nascent chains to be unfolded (near 20 amino acid marker). However, after this point some amount of folding can occur depending upon the protein structure. After about 26 amino acids, the tunnel widens into the "vestibule" which allows for larger structures to potentially form.

This tunnel can allow small amounts of secondary structure to form, but for the most part forces the nascent chain into an extended conformation. Some studies have shown folding as early as 25 amino acids past a domain boundary, but this is for the case of a very small domain of 29 amino acids (*10*). Other studies show signatures of folding intermediates at 40 amino acids (*11*) or 44 amino acids (*12*). Thus, the polypeptide is sequestered from folding until it is outside of the tunnel. The rate of translation, the sequestering of peptide chain in the tunnel, and its gradual release to the solvent all may affect the folding process as compared to *in vitro* reconstitution post-synthesis, where all of the sequence is available to fold in the solvent. The ribosome itself may affect the chain through electrostatics (*13*, *14*), as it is highly negatively charged (Figure 1.5), or through entropic effect since the chain is essentially tethered to a much larger surface which limits its conformational space (*15*).
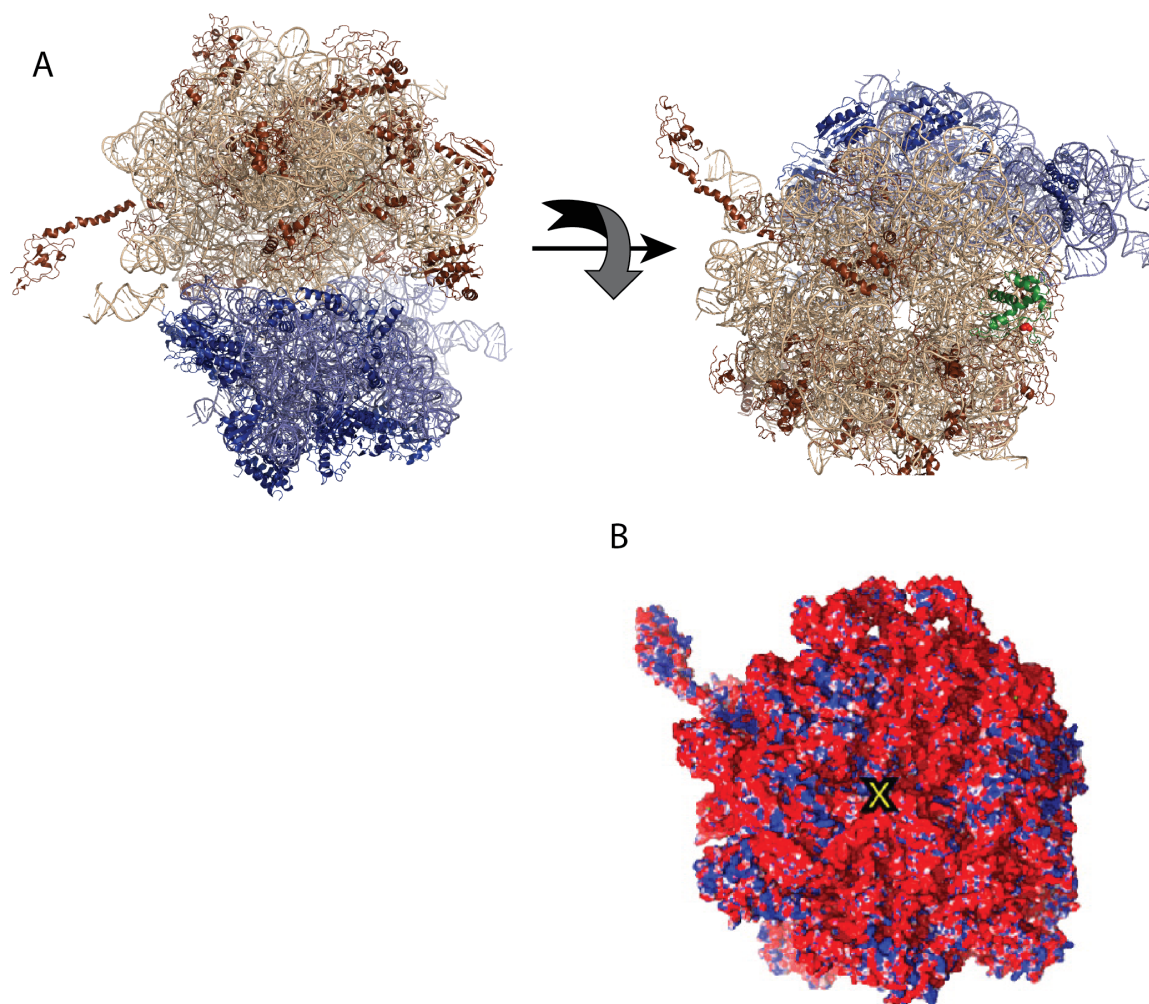
**Figure 1.5: The surface charge of the ribosome.** A) The structure of the 70S ribosome, with same colors as in Figure 1.1 The view is shown from the "front" of the ribosome, where mRNA would bind in between the subunits in a 5'→ 3' direction. The second view is rotated to show the location of the exit tunnel (the hole). The green protein (L17) is the labeled protein discussed in Chapter 3. B) shows a similar angle as in A, but only the 50S and with an electrostatic map (red: -5 kT/e blue +5 kT/e). From reference (*14*). The X marks the location of the exit tunnel.

All of these aspects of protein synthesis could help explain the robust folding of multidomain proteins—chains are prevented from off-pathway self-interactions because later sequence is not yet synthesized or is sequestered in the tunnel. Chains are also less prone to aggregating since they are attached to the ribosome itself. The structures formed on the ribosome may only reflect a subset of accessible structures or different intermediates than observed in solution (*16*). These effects all help promote native, productive folding; schematically, the ribosome can bias the nascent protein into a deeper energy well or reduce the complexity the landscape (Figure 1.6)
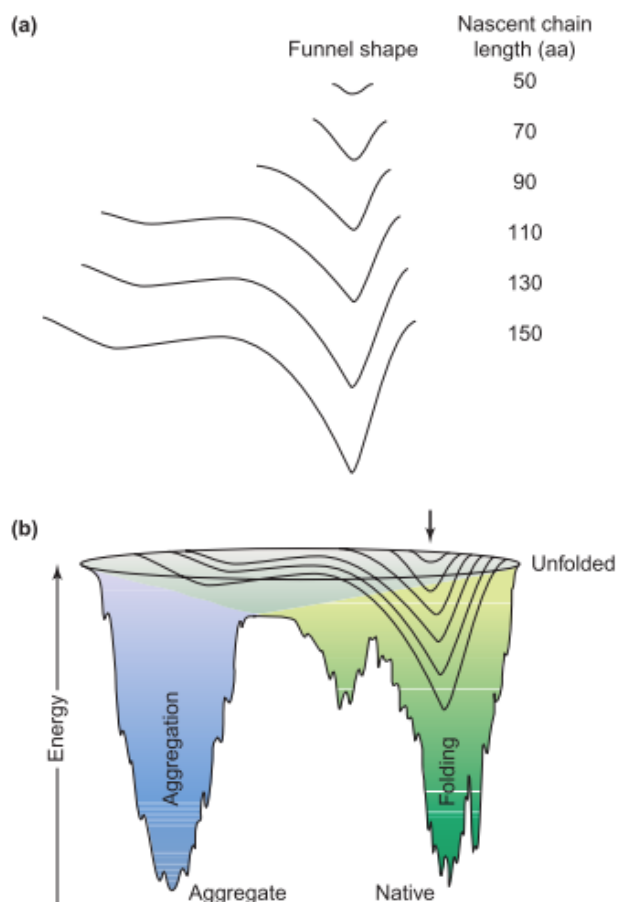
**Figure 1.6: Energy landscapes of folding on the ribosome** From reference (*17*). As the nascent chain lengthens, the landscape changes. With more sequence available, there are more conformations available as well, although not as broad as the protein in solution. In this schematic, the ribosome directs the protein away from a possible local minimum that is not the native state, as well as reducing the minimum of the aggregation funnel.

However, these effects are putative and direct studies of multidomain protein folding on the ribosome or especially during synthesis are limited. There are many questions which are unanswered or only partially answered. How does folding change on the ribosome compared to the full-length, post-translational protein, in either kinetics or the folding trajectory itself? Does the ribosome prevent off-pathway states, as postulated? We seek to answer these questions by studying the folding pathway of a multidomain protein on the ribosome stalled during synthesis (an equilibrium population) and also during synthesis itself (which, as we will show, is a non-equilibrium process, as the amount of sequence available, and thus the folding landscape, changes in time).

Since the ribosome is so large relative to the nascent chain (2.3 MDa versus ~100 kDa) and is itself made up of more then 50 proteins, cotranslational folding is difficult to investigate with many techniques. Most spectroscopic techniques will be dominated by the signal from the ribosome. Furthermore, most denaturants that will affect the nascent chain will also destabilize

the ribosome and thus perturb one of the factors of interest. A few techniques have been developed which allow for probing a ribosome nascent chain complex (RNC). These include:
- fluorophores within the nascent chain (*18*)
- NMR with isotope-enriched nascent chains (*12*)
- Optical tweezers force measurements (*13*)

The benefits of using optical tweezers to investigate cotranslational folding are primarily twofold: the denaturant (force) can be easily manipulated to varying strengths without affecting the ribosome globally, as the force and will only act between the tether points and not on the entire complex (an anisotropic denaturant), and the read-out (molecular length) is observed with high spatial and temporal resolution, allowing direct characterization of the unfolding or refolding kinetics and the identification of structural intermediates. Furthermore, as a single molecule technique we can observe transient intermediates, subpopulations, and detailed statistical kinetics that are all but averaged out when using bulk methods. In particular, this technique is well suited to investigate the co-translational folding of multidomain proteins.

## 1.4 Optical tweezers and instrumentation

Optical tweezers are sensitive to piconewton (pN) and nanometer changes, which is the optimal dynamic range for protein measurements. By attaching the molecule of interest to two polystyrene beads, the molecule can be manipulated and measured by trapping the polystyrene beads in two optical traps to suspend the molecule in the middle. The optical trap can apply force due to the difference in index of refraction between the polystyrene bead and the surrounding medium. When the light from the trap refracts due to the bead, there is a change in momentum of the light, which prompts a change in momentum on the bead as well. This acts as a restoring force on the bead to the center of the laser beam (Figure 1.7).
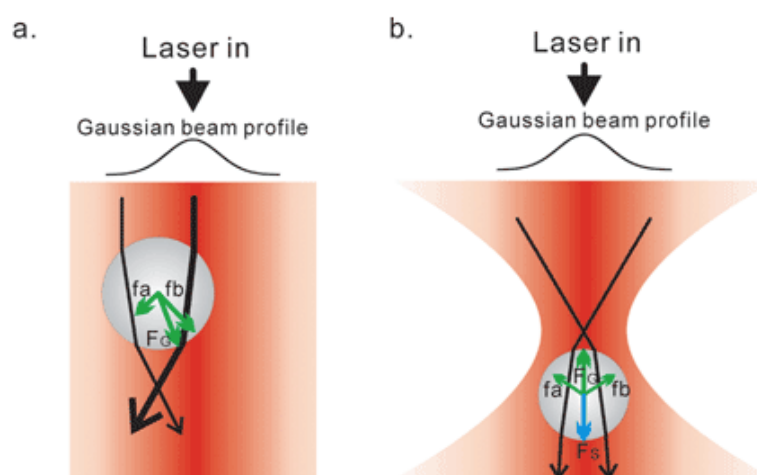


**Figure 1.7: Principles of optical trapping.** From reference (*19*). A) If the bead is displaced from the center of the beam, it experiences uneven forces which act to restore it to the center. Black arrows represent the incoming photons and the outgoing refracted beam, which has a change in momentum. The green arrows are the resulting gradient forces on the bead due to this change in momentum of the light. B) The stable equilibrium (zero net-force) position is slightly behind the beam focal point, where scattering and gradient forces are balanced.

This force is well described as a Hookean spring for small displacements of the bead. If the effective spring constant κ of the trap is known, the force can be measured by measuring the displacement of the bead from the center of the trap.

The spring constant κ is dependent on the laser power and radius of the bead. Although there are multiple ways to calibrate the stiffness, one common method is through measuring the power spectrum (noise vs. frequency) of a bead held in an optical trap (Figure 1.8).
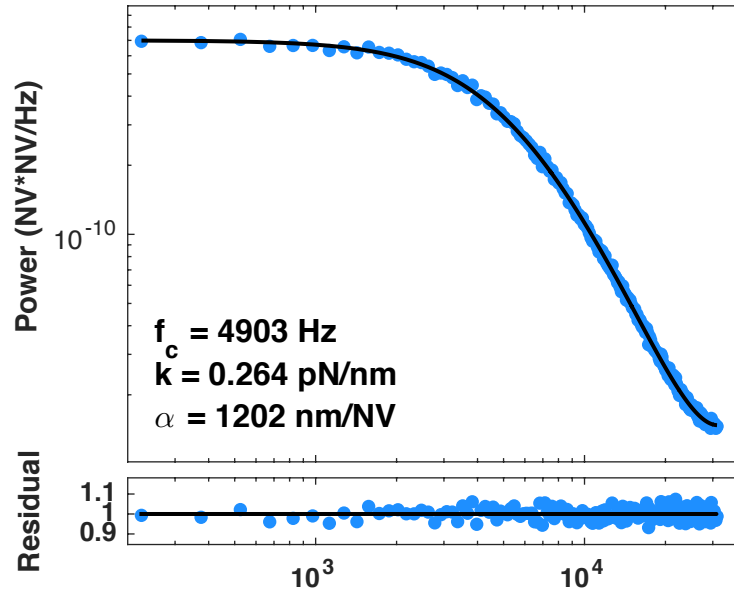


**Figure 1.8: Power spectrum of a trapped bead**. The random motion of a trapped particle fluctuates within the trapping potential. This spectrum allows for the determination of the corner frequency ($f_0$, in the figure as $f_c$) which limits temporal resolution, and κ, which allows for conversion of bead displacement to forces.

The bead, without a tether or external force, will fluctuate according to random thermal (Brownian) motion. The magnitude of these possible fluctuations is constrained by the trapping potential. Thus, by recording the fluctuations in the trap, the strength of the trapping potential can be measured. The power spectrum is described by a Lorenzian function (*20*):

$$\sigma^2 = \frac{k_B T}{\alpha^2 \pi^2 (\gamma)(f_0^2 + f^2)}$$

In this equation, $\sigma^2$ is the power, $f_0$ is the corner frequency of the bead, $T$ is temperature, and γ is the drag coefficient of the bead. This drag coefficient, $\gamma = 6\pi\eta r$, depends on the bead size (*r*) and also the viscosity η. This also fits a parameter α depending on the conversion factor of the detectors (from volts to nanometers). For quick visual comparison, the plateau region of the power spectrum occurs at a value of $\frac{k_B T}{\alpha^2 \pi^2 \gamma (f_0^2)}$, and $f_0$ is near the rollover region. This corner frequency can be further used to obtain κ since $f_0 = \kappa/(2\pi\gamma)$ for random bead motions (*20*). The corner frequency is related to the autocorrelation time of the bead in the trap and thus will determine the upper-limit of temporal resolution. Data collection at frequencies higher than the

corner frequency will give correlated measurements of bead position that do not report on the dynamics of the biological system.

Previous experiments on RNCs with optical tweezers (*13*) employed a dual-propagating trap and a pipette ("mini-tweezers" (*21*)). This system has two lasers that enter the fluidic chamber from objectives on both sides but are focused to the same point. Thus, two lasers are used to generate one dual propagating optical trap. The second "trapped" position is generated via a pipette that uses suction to hold one polystyrene bead stationary. The dual-propagating trap acts to manipulate and measure the molecule relative to the stationary pipette. The benefit of the dual-propagating trap is that much higher forces can be reached with much lower operating power, since the traps are working in tandem. However, these types of set-ups are much more prone to drift, both due to drift of the pipette relative to the trap as well as the traps relative to each other. Therefore, they are not good for measurements over long time scales (tens of seconds).

This type of set-up can overcome some of its inherent drift instability by employing an algorithmic electronic feedback on laser position. This feedback acts to restore the force to a constant value. However, for the feedback to operate in a useful manner and not overcompensate to noise, the signal must be integrated and averaged, and thus feedback must occur at significantly less than the corner frequency of the system. This feedback limits the temporal resolution of the system. In practice, the time resolution of the mini-tweezers is < 1 kHz in passive collection (without feedback) and < 300 Hz with feedback. The mini-tweezers is best suited, therefore, to systems which are considerably slower than 300Hz. For some proteins which have a large hysteresis, this is fine. However, if we are interested in cotranslational folding, we are most interested in events which are considerably faster than translation and thus approach this range. This will be further addressed when looking at the data collected for our model system, calerythrin. It will be shown, however, that we need higher time resolution, lower noise, and great drift stability and thus require a single beam, dual-trap set-up.

The dual trap set-up uses one laser, split via a beam splitter or an acousto-optic modulator (AOM) into two beams. These two beams then generate the two traps, with trap movement controlled either by the frequency of the AOM or by a piezo-mirror which can move one trap. This set-up has higher time resolution due to the increased laser power and lower drift due to isolation of the system from the microfluidic chamber as the beads are not correlated to larger chamber motion as previously (through the pipette). Furthermore, drift of the traps can be limited by minimizing the differential path of all optical components, so most drift of the optics is present in both traps and thus does not cause a net change in extension.

## 1.5 Extracting protein folding information from single molecule data
i. Overview of experiments

There are primarily two types of information we are concerned about extracting in this work–the size of the folding transition in amino acids and the rates of folding/unfolding. Although other parameters can be extracted (distance to the transition state, transition path time (*22*), landscape reconstruction (*23*), non-equilibrium work (*24*), etc.) we will not discuss them here.

The size of the transition allows us to infer which regions of the protein are folding, in conjunction with structural information or mutants. The data itself is in piconewtons across the tether or the extension of the tether in nanometers. The relationship between the force and the extension will depend upon the elastic properties of the polymer (in our case, a composite of polypeptide, the molecule of interest, and DNA, which is required to move the sample away from the bead surface). As such, the conversion requires knowledge of these polymers' behavior as a function of force.

Both DNA and polypeptide can be modeled as worm-like chain (WLC) polymers, where the polymer is modeled as a continuously flexible rod. The flexibility of the polymer is dictated by its persistence length, which is the correlation length of the direction vector. Broadly, this can be pictured as the average length before the polymer starts to bend. The worm-like chain relationship between force and extension can be approximated as (*25*):

$$\frac{FP}{kT} = \frac{1}{4}(1 - \frac{x}{L})^{-2} - \frac{1}{4} + \frac{x}{L}$$

Where $P$ is the persistence length, $L$ is the total length of the molecule along its backbone (contour length) and $x$ is the distance between the two ends (extension). This model assumes the molecule is behaving as an entropic spring, without stretching or deformation (enthalpic effects). For polypeptide, this is sufficient, but for DNA there is an enthalpic component which can be added by considering the stretch modulus of the polymer K:

$$\frac{FP}{kT} = \frac{1}{4}(1 - \frac{x}{L} + \frac{F}{K})^{-2} - \frac{1}{4} + \frac{x}{L} - \frac{F}{K}$$

This is the extensible WLC (*21, 26*).

DNA is much stiffer than polypeptide, with a persistence length of around 50 nm, compared to the persistence length of polypeptide of 0.6 nm. The relative extension (a percent extension, $x/L$) is shown plotted below for DNA and for polypeptides (Figure 1.9).

As you can see from the plot, polypeptide has a much lower fractional extension in the range of experimental forces. This means that the observed change in the OT will be relatively small and the noise will also be increased relative to a DNA-only tether (noise is inversely proportional to the slope of the curve, *dF/dx*). In order to convert our data to the corresponding size change of the protein, we need to use the WLC to convert extension data to contour length, using known values of contour length per bp (0.34 nm/bp) or contour length per amino acid (0.365 nm/aa). If we observed a signal change of 10 nm at 6 pN in a DNA-based system, this signal would correspond to 93% of the total contour length change, i.e. the instrument has captured the majority of the motion. Calculating back to contour length, this is 10.7 nm, or 31.4 bp of DNA. However, if we observe a 10 nm change in a polypeptide tether at 6 pN, this would correspond to only 40.6% of the contour length, meaning we have much lower sensitivity. The same 10 nm would be 24.5 nm in contour length, roughly 67 amino acids.
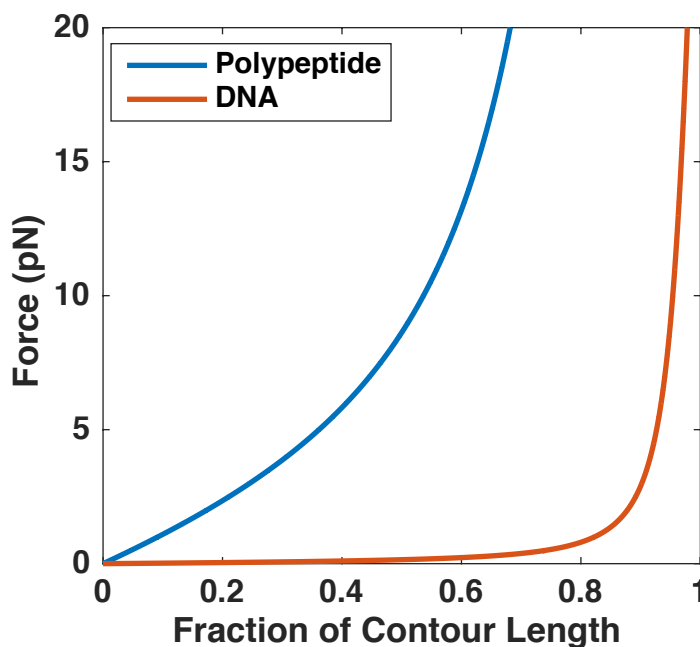
**Figure 1.9: Theoretical force extension behavior of DNA and polypeptide** Calculated WLC values for polypeptide (inextensible) with P = 0.6 nm and DNA (extensible) with P = 35 nm and K = 1200 pN, the values applicable in our experimental conditions.

The above estimates are based on changes assuming constant force (either constant force data or force-extension data can also be analyzed this way by looking at extension of unfolded/folded states at the same force). In passive mode, in order to extract the overall transition size, we need to account for changes in the polypeptide and changes in the handle length. For a folding transition, the protein gets more compact, thus shorter and to higher force. However, the DNA handles will extend further at higher force, canceling out part of the observed signal change. We need to account for both of these factors in our conversion. The inter-trap distance remains constant, however, which allows for calculation of the changes of each part of the system. If we consider the same 67 amino acids that corresponded to a 10nm signal at constant force, in passive a 67 amino acid transition would be a force change of +0.53 pN, or < 5 nm change in extension. Although it would seem that passive has even lower sensitivity, the noise is greatly reduced without feedback, so it is about the same.

In order to extract kinetic parameters, we need to determine the lifetime of a given state as a function of force. Because the initial and final states are directly observable in the OT data, in many cases you can measure the rates associated with all states in the system, without convolution, unlike methods that give sum signals from all states. The rates are described by transition state of crossing a barrier in a potential energy surface from one well to another. At zero force the barrier is very high relative to thermal energy and protein unfolding is rare. Applying force tilts the energy landscape to favor the unfolded state, lowering the barrier (Figure 1.10)
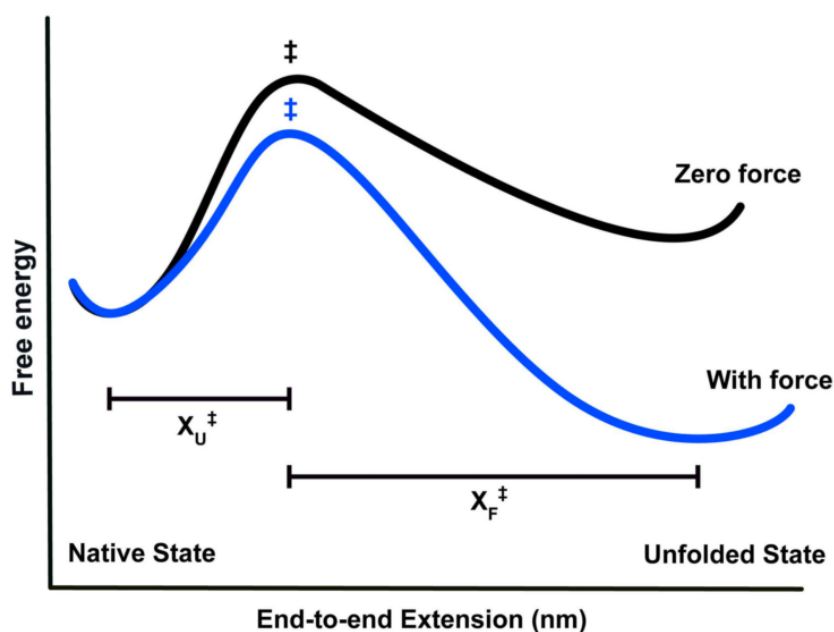
**Figure 1.10: Protein folding energy landscape under force** Figure from reference (*27*). As force is applied, the energy is shifted by a factor of –*Fx*. Thus higher extension states, like the unfolded state, have a larger bias. The unfolded state generally has a broad shallow energy minimum, while the folded state is rather "brittle" (*28*). The distance to the transition state determines how sharply the protein kinetics shift with force.

The simplest way to model this "tilting" is to use Bell's Model, where there is a Boltzmann factor modifying the intrinsic zero-force rate:

$$k(F) = k_m k_0 e^{-\beta F \Delta x^\dagger}$$

Here $k_0$ is the zero force rate and $\Delta x^\dagger$ is the distance to the transition state, while $k_m$ represents contributions from the experimental geometry (beads, handles, etc.). This model predicts a linear dependence on the ln(k) versus force, and can be easily fit to experimental data. We can extrapolate the force-dependent rates to a zero-force rate, however, unless the coefficient $k_m$ can be estimated, the measured rate will likely not match the true rate without handles attached. This simple model is mostly phenomenological, and is limited in its applicability. It assumes the $\Delta x^\dagger$ is constant (the slope of the linear fit). As such, this approximation is most valid for short-range extrapolations.

A more accurate model introduces an additional parameter besides $\Delta x^\dagger$, which is the height of the transition barrier $\Delta G^\dagger$. This additional parameter allows for full characterization of the kinetic barrier without requiring temperature-dependent experiments, as would otherwise be required. This non-linear model is:

$$k(F) = k_m k_0 (1 - \frac{\nu F \Delta x^\dagger}{\Delta G^\dagger})^{\frac{1}{\nu}-1} e^{\Delta G^\dagger[1-(1-\frac{\nu F \Delta x^\dagger}{\Delta G^\dagger})^{\frac{1}{\nu}}]}$$

Here $\nu$ describes the shape of the potential (*29*). For the case of $\nu = 1$, this model reduces to the Bell model described above. For most physical systems, they can be modeled reasonably well by $\nu = 2/3$, which is for a linear-cubic model potential. This equation requires enough spread of the

data in force than the non-linearities are observed. This naturally leads to the question of how data can be collected, and over what range of forces.

There are primarily four types of experiment which can be conducted in optical tweezers to study protein kinetics: force-extension curves, force jumps, constant force, and constant trap position (passive) experiments (Figure 1.11). Constant force and constant position experiments are equilibrium measurements of the folding landscape with a certain bias applied. Force jumps and force-extension are nonequilibrium measurements. I will briefly discuss the advantages and disadvantages of each type of experiment and how the data can be analyzed.

## A: Force-extension curves

Move trap at a constant velocity
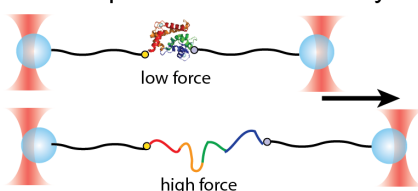
Detect unfolding "rips"
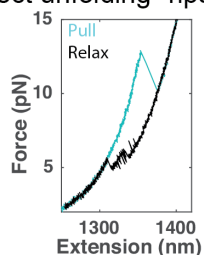
Collect histogram of forces
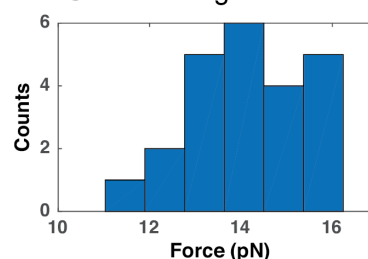
## B: Constant force

Move traps, maintain constant force
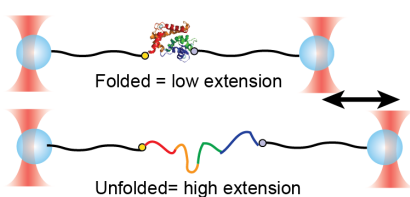
Detect hopping transitions
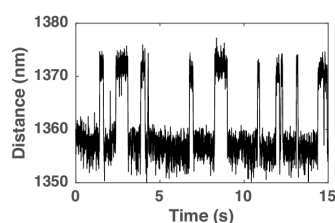
## C: Passive

Stationary traps, fluctuating force
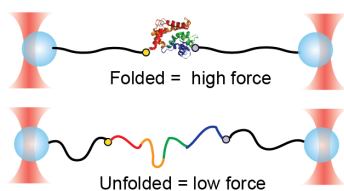
Detect hopping transitions
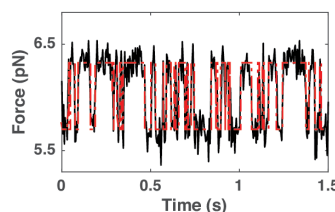
Collect lifetimes $\tau(F)$
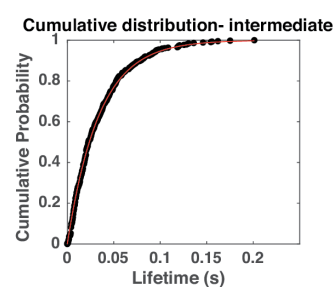
## D: Force jumps

Move trap suddenly
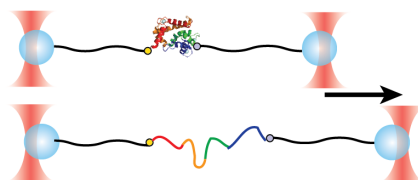
Detect unfolding "rips"
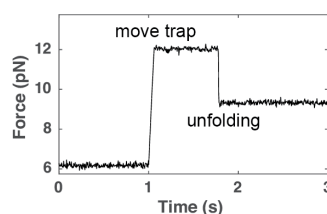
**Figure 1.11 Overview of experimental modes in OT.** A) Force-extension curves move the traps apart at a constant rate (in nm/s), resulting in a gradual force increase and eventual protein unfolding. If the pulling rate is known, the resulting histogram of unfolding force probability can be related to the lifetime of the folded state at a given force. B) Constant force data uses algorithmic feedback to adjust the trap positions, as shown for calmodulin. C) Passive data keeps the traps fixed while force and tether extension fluctuate (shown for calerythrin). D) Force jumps allow for the measurement of kinetics past the equilibrium hopping range by waiting for unfolding(folding) at a high(low) force, then repeating the jumping procedure. All modes B-D will yield distributions of lifetimes directly, unlike A which is indirect.

Force extension experiments (so called FX curves or pulling curves) collect data while applying a constant velocity to one of the traps (Figure 1.9A). The molecule of interest will extend due to the increasing force, which can be modeled with a worm-like chain model. Once the force is high enough to overcome the stabilization of the molecule, it will unfold. The unfolding transition is usually a sharp drop (cooperative) which will have a slope which is equal to the trap stiffness of the beam. To use hairpin unfolding as an example, usually it will unfold as a subset of base pairs all at once rather than 1 base pair at a time. For proteins, a similar cooperativity is often observed, where many single domain proteins unfold in a single transition. Because the molecule is being continuously extended, it does not have time to equilibrate to a given force and samples each force's energy landscape for a very short time. Because of this limited sampling, the force of the transition will vary. The further from equilibrium the experiment is performed (the faster the pulling speed) then the higher the force of the transition will be observed, since the molecule did not have a chance to respond to the force before it changed.

A histogram of rupture forces can be generated from many repeated puling cycles and then fit to determine the zero-force rate if the pulling speed is known and it is a single barrier crossing (*30*). For a normalized probability distribution of unfolding events (PDF), the lifetime can be calculated as:

$$\tau(F) = \int_{F}^{\infty} \frac{PDF(F)df}{v(F)PDF(F)}$$

The pulling speed *v* must be used as a function of force, but is usually constant in terms of nm/s, so the local loading rate must be calculated from the force-dependent response of the molecular construct.

However, if there are multiple transitions, as would be expected in a multidomain protein, only the first transition could be analyzed in this straight forward manner, and later transitions would need a more complicated model which accounts for the undersampling of low forces, since the second transition by definition will only be observed after the first transition (*31*). Force extension measurements are further complicated in the case of highly reversible barriers, where many barriers crossings may occur in one pulling cycle. This makes it harder to identify individual transitions to add to a histogram and also complicates the model presented above.

In cases where the protein does not exhibit large hysteresis, constant force or passive experiments can be conducted. A constant force feedback system computationally corrects the

force in real time to maintain a stable force. When a folding transition occurs, the force begins to increase, and the traps will move together to return the force to the set force, and the folding transition will be monitored as a change in distance (Figure 1.9B). In general, the data is easy to interpret as the size of the transition does not have to be corrected for changes in handle length and thus only reports on the molecule of interest. It also is easy to model, as a constant force corresponds to a linear bias to the free energy landscape where there is an added term of $-Fx$ to the free energy. However, this type of measurement is prone to artifacts because it requires computation feedback and consistently underestimates the true rates (*32*). It also limits the resolution, both spatially and temporally. Temporally the resolution is limited because feedback can only be performed at a frequency which is less than the response time of the bead, otherwise the feedback is acting upon correlated measurements of bead position. Spatially the noise is amplified since feedback is applied to both real dynamics and Brownian noise. As such, passive experiments are preferred due to the more accurate rate determination and resolution.

Passive experiments are performed at a constant trap position. A trap position applies an average force bias to the system, but both the force and extension of the molecule changes in time as it folds and unfolds (Figure 1.9C). For a folding transition, the molecule would compact, shortening the extension and pulling the beads further from the center of the trap (raising the force). However, at higher force the handles also extend more, which would decrease the absolute signal change expected. The folded state also exists at a higher force than the unfolded state did, which promotes subsequent return to the unfolded state. This mode is thus useful in that it promotes frequent, reversible sampling of the energy landscape. Both constant force and passive mode experiments are equilibrium experiments that measure lifetimes with a constant energy landscape. However, since both the force and extension change with each transition in passive, the perturbation to the landscape itself is more difficult to model than the simple linear bias in constant force, and also depends on trap stiffness.

In general, force extension curves and equilibrium measurements may only exhibit transitions across a few piconewtons of forces, requiring a large extrapolation that may lead to greater inaccuracies in the estimated rates. Force jumps experiments (Figure 1.9D) can be used to supplement either type of experiment to sample a much wider range of forces. The unfolded (folded) molecule is jumped to a low (high) force repeatedly, and the time until folding (unfolding) is recorded. This provides a single measurement of the stochastic lifetime of the initial state at that force. Each event is one data point that can be used to construct a probability density function (PDF) or cumulative distribution function (CDF). As such, many repeated force jumps are necessary to accurately measure the kinetics and construct these distributions.

ii. Hidden Markov Models

Hidden Markov models assume there is an underlying Markov chain that accurately describes the transitions in the data. A Markov chain is a system of interconnected states where each state $i$ has a fixed probability of leading to another state $j$ where $0 \leq p_{ij} \leq 1$ (Figure 1.12).

A



B
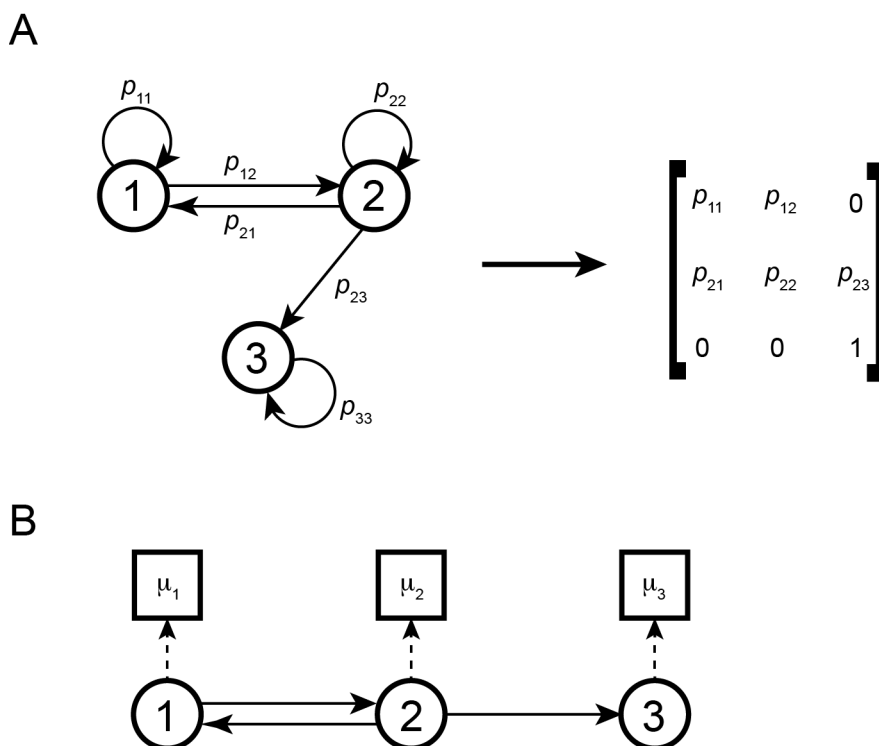


**Figure 1.12: Markov chains and Hidden Markov chains.** A) A simple discrete Markov chain. Each state has a certain transition probability $p_{ij}$, the probability of transiting from state $i$ to $j$ in a discrete time interval $\Delta t$. These probabilities are entries in the transition matrix, and all rows must sum to 1. In this case state 3 is an absorbing state since it cannot exit it once it enters. B) A Hidden Markov chain has an underlying Markov chain (states 1, 2, 3) but the observable values are indirect ($\mu_1$, $\mu_2$, $\mu_3$). These observables may or may not be unique to the true state, rendering the underlying chain "hidden". Arrows for transitions $p_{ii}$ are not shown for simplicity.

Since the data is collected at a certain frequency, we can use a discrete-time Markov chain where the sum of all probabilities out of a given state in a fixed time window is 1 (the state also has a fixed probability to transition to the same state, i.e. no transition occurs within the time window). The time window $\Delta t$ is the sampling frequency of data collection. For a discrete Markov chain, the transition matrix **T** contains the probabilities, which can be related to the rate matrix of rate constants, **K** according to:

$$\mathbf{T} = e^{\Delta t \cdot \mathbf{K}}$$

This can also be approximated by an element-wise computation (*33*), rather than a matrix exponentiation, by the relation:

$$k_{ij} = -\ln(1-p_{ij})/\Delta t$$

If the probability $p_{ii} = 1$ then it is an absorbing state and no further transitions would be observed. A Markov chain is a so-called "memory-less" process because the probability of a given transition only depends on the current state and not on the previous sequence of states (the "state trajectory"). Because events are single steps with a fixed probability, the time until an event is exponentially distributed and the number of events in a certain window is Poisson distributed, meaning the kinetics are first order and have an exponential PDF. Many chemical and biological processes have underlying Markovian statistics. The measurements themselves

may not be Markovian even if the physical phenomenon is– for example in OT, if you sample faster than the corner frequency, the reading contains correlated information about bead position. In this case, each data point has "memory" since it is correlated. The data input needs to be checked that it is uncorrelated and thus can be assumed to represent Markovian transitions prior to analysis with a Markov model.

A "hidden" Markov model (HMM) implies that the process is Markovian, but that the observations are not of the states directly but instead some observation variable. The observation output may obscure identifying the true state. This could be due to, for example, two states with the same observable output (but with different underlying states and thus probabilities) or due to the presence of noise such that one state has a range of output values which could overlap with other states'. Either of these cases could be true for optical tweezers. Because our read out is molecular length, two different folded states could have the same observable but correspond to different parts of the structure. Likewise, the inherent noise of our measurement means that some states cannot be unambiguously assigned. In this case a HMM will find the statistically most likely trajectory that would lead to the observed output. In this process it would also find the transition probabilities.

A HMM with a number of states $n$ has $n^2+n$ degrees of freedom for fitting– $n$ values for the mean observables of each state ($\mu_i$), $n$ values for the widths of each state ($\sigma_i$), and an $n$x$n$ transition matrix **T** with the probabilities into and out each state $p_{ij}$. However, because the sum of each row of the transition matrix must be 1, **T** only contributes $n^2-n$ degrees of freedom. In practical terms, this means that larger numbers of states will require more computation time, non-linearly, to converge to a stable fit. The fitting procedure is iterative, seeking to maximize the log likelihood of an assigned state trajectory. The full algorithmic details behind HMMs are covered very thoroughly elsewhere and are beyond the scope of this work (*34*); we are using a HMM library of code in Python for our analysis.

In the most rigorous sense HMM are a type of machine-learning algorithm and the algorithm should be trained on a subset of data as the learning set, then the algorithm would use this to assign states within a new dataset. For single molecule measurements the training set and the output set are often the same. Because of these limitations in training/optimization, the "optimized" transition matrix is often not truly optimized and a better result can be obtained from ignoring the parameters in **T** and instead analyzing the state trajectory, which has a maximized likelihood. The state trajectory can provide lifetimes and identities of states, which easily can be analyzed to provide kinetic parameters through either a PDF or CDF. The CDF does not require binning and so it the procedure we used here (Figure 1.13).
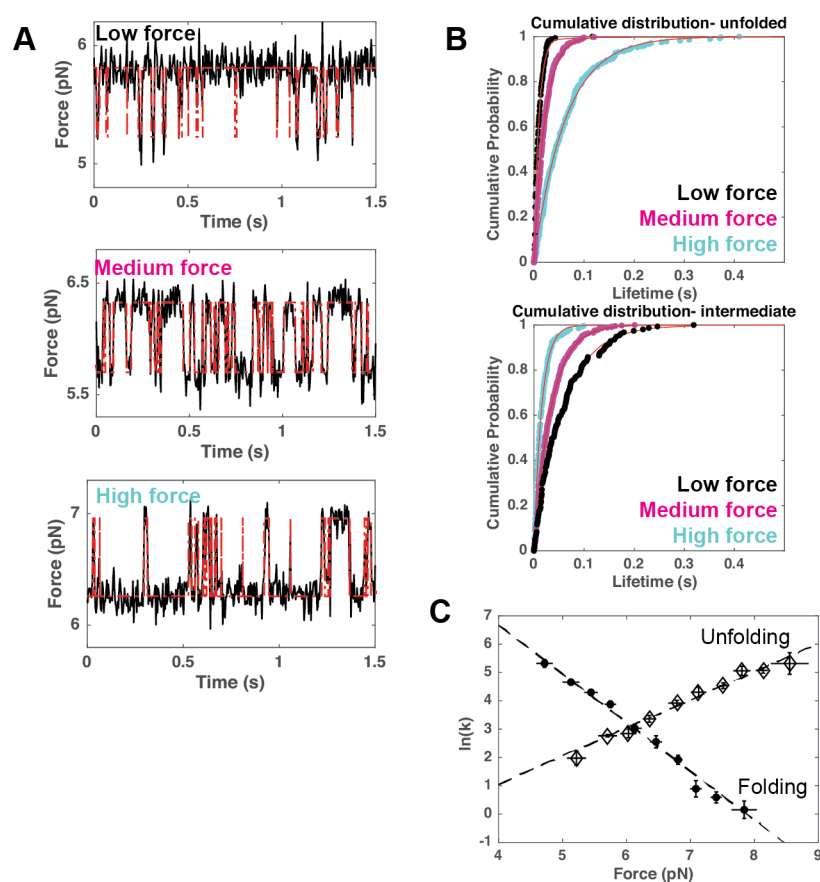
**Figure 1.13: Extracting lifetimes from HMM trajectories.** A) These panels show data collected at 3 different trap positions, which results in a change in the mean force. In all cases, the unfolded state is the lowest force and it has two-state behavior. The HMM trajectory fit is shown in red. B) The lifetimes from the trajectory fit are plotted as an empirical cumulative distribution function. This follows a $1 - e^{-kt}$ relationship. Each fitted value of $k$ is specific to the force at which it is measured. C) The semi-log graph of ln($k$) vs. F is linear, see discussion of Bell's model.

Using the trajectory has the added advantage that it allows verification of the fit– an underfit-HMM will have too few states and not have single exponential lifetimes. Thus, if any of the states are not single exponential either the fit was not converged or there are too few states in the initialization. The actual distribution you observe in the case of an underfit HMM instead depends upon the underlying state network connectivity. In theory it could be calculated from convolution integrals of PDFs but in practice the analytical expressions are not simple except for simple kinetic schemes, such as parallel (unconnected states) which if their distributions were combined would result in a multi-exponential distribution; this case is addressed further when studying the EF123 truncation (Chapter 2.4 and in Chapter 3).

# Chapter 2.

## Characterizing the folding of calerythrin in solution

Portions of this section are reproduced with modifications from co-authored material with permission of the other authors.

### 2.1 Introduction

Calerythrin (CaE) is a two-domain calcium-binding protein from the prokaryote *Saccharopolyspora erythraea* (*35*, *36*), a member of the EF hand family (Figure 2.1).
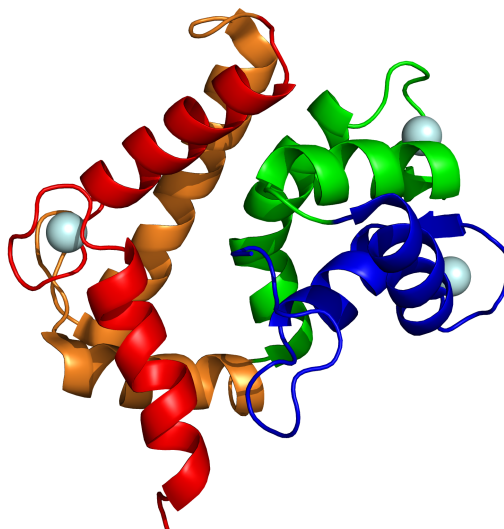


**Figure 2.1: NMR structure of Calerythrin.** EF hands 1, 2, 3, and 4 are red, orange, green, and blue, respectively. Calcium ions are shown as pale blue spheres. EF1 and 2 pair to form the N domain; EF3 and 4 form the C domain. EF2 does not bind calcium due to a non-canonical loop sequence in the EF hand although the α-helices still fold. From PDB 1NYA (*36*).

EF hands are a canonical folding motif with a helix-loop-helix structure. The helices form ~100˚ angles to each other to correctly position the conserved calcium binding site in the loop. The loop has 6 conserved residues coordinating the calcium in a pentagonal bipyramidal geometry, with the 12$^{th}$ position of the loop a conserved glutamate or aspartate that acts as a bidentate ligand and is strictly required for binding. Two EF hands stack and pair to form a stable domain. EF hands are not stable without another binding partner, and proteins with an odd number of EF motifs are observed to dimerize (*37*). The paired EF hands often exhibit calcium-binding cooperativity. In CaE, the C-domain binding sites have cooperative binding effects and measured affinities of $1.2x10^{-8}$ M and $1.3x10^{-9}$ M, while EF1 has a lower affinity of $3.7x10^{-7}$ M (*38*).

EF hand proteins are found in bacteria through humans. The most widely studied member of this family is calmodulin, an archetypal signaling protein that regulates neuronal response (*39*). Aside from signaling, there is another class of EF hand proteins, sarcoplasmic binding proteins (SBPs), which act as calcium buffers and regulate the resting calcium concentration. The function of calerythrin is not fully known, although it is likely that calerythrin is acting as a calcium buffering protein rather than a signaling protein due to its greater similarity to SBPs (*40*). There is not a consensus on the role of calcium in bacteria, and the calcium pathway has not been fully documented, although it must be regulated since resting levels are kept low (nM-level) (*41*).

Members of the EF hand family of proteins has been studied in optical tweezers previously, including calmodulin (*42*) and NCS-1 (*43*). These proteins have been observed to fold at high forces (near 10 pN) and with fast equilibrium kinetics. These attributes made calerythrin an attractive model system for cotranslational folding– the fast kinetics of formation mean that it is likely to fold cotranslationally, or at least that its solution characteristics are competitive with the rate of translation. Furthermore, we can collect data in passive mode quickly, and even if the ribosome suppresses folding to lower forces or slower rates it would still likely be in a resolvable regime (>2 pN). Both NCS-1 and calmodulin showed misfolded states in the optical tweezers studies, suggesting EF hand proteins may be prone to forming non-native interdomain contacts (*42, 44*). This highlights our interest in studying the folding of this family cotranslationally, to see if and how these states may be avoided.

NMR and CD studies with apo-calerythrin suggest it is a molten globule without calcium and upon calcium titration, the C domain folds first in a cooperative manner (*45*). This stimulates further questions, since in the context of synthesis, the C domain is not available for folding until after the protein is released from the ribosome. Therefore, we proceeded with CaE as our model system, first measuring if the folding of the released protein in OT is similar to the folding observed upon calcium titration.

## 2.2 Instrument limitations for fast kinetic systems

In order to study calerythrin in OT, we inserted an N-terminal Avi-tag and a C-terminal ybbR tag. These tags allow us to covalently attach modifications to specific side chains: a biotin to a

lysine residue within the Avi-tag (*46*) and a coenzyme-A moiety to a serine within the ybbR tag (*47*). This geometry allows the application of force from the two ends of the protein: the N terminus by binding the biotin with a streptavidin "bridge" to biotinylated DNA and the C terminus by chemically crosslinking the CoA to a small oligonucleotide (oligo) prior to tagging the protein. This oligo allows for subsequent ligation to a DNA handle or oligo-coated bead.

Our initial experiments were designed with the same geometry as previously used to study T4 lysozyme, where the protein is ligated directly to an oligo-derivatized bead with only a 50 bp spacer, and held on a pipette in a single trap instrument (Figure 2.2) (*13*).
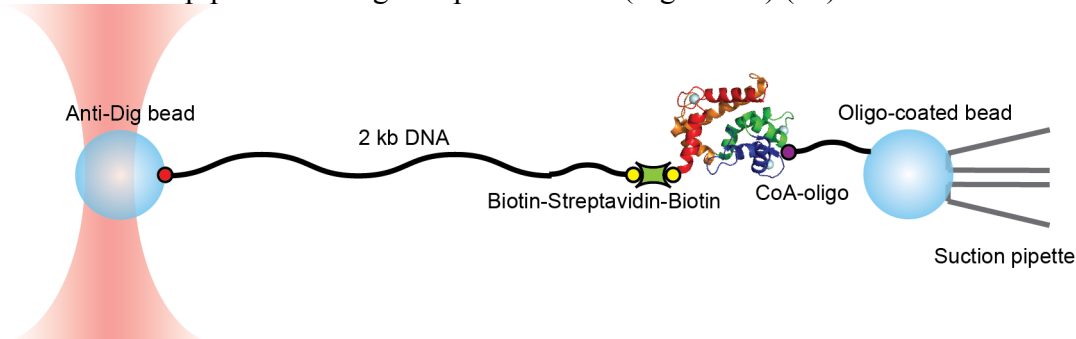


**Figure 2.2: Geometry of single molecule tether in mini-tweezers instrument.**

Data acquired in this setup shows that calerythrin unfolds cooperatively in a single transition at high force and refolds in two steps at lower forces (Figure 2.3). This hysteresis between full folding and refolding suggests the two domains have some critical interactions, likely at the interdomain interface, which provide additional stabilization to the protein. This is especially likely in the context of other EF hand proteins, which often show independent, reversible kinetics of each domain (for example, calmodulin in Figure 2.3A)
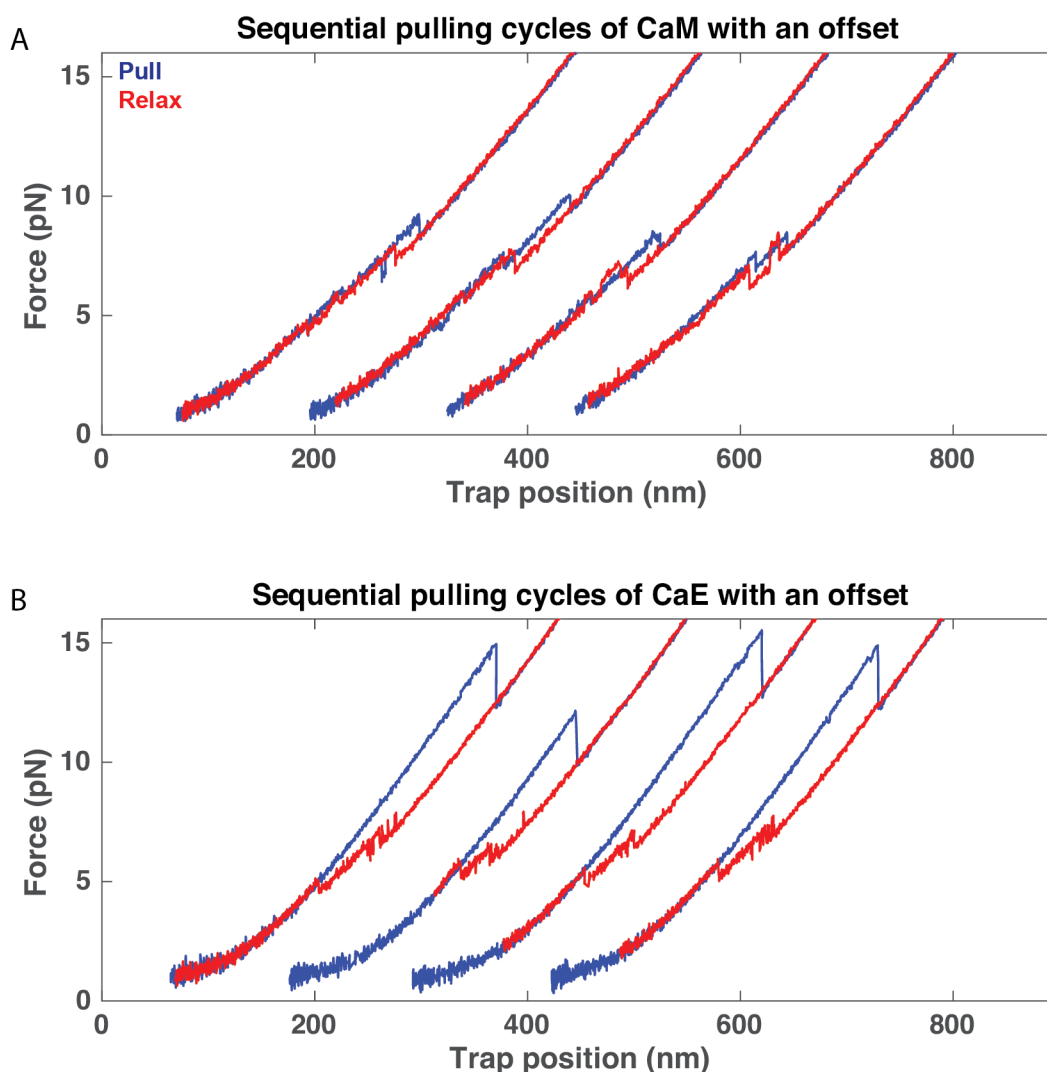
**Figure 2.3: Force-extension curves of Calerythrin vs. Calmodulin.** Cycles of pulling (blue) and relaxing (red) a single molecule. A) Calmodulin can fold and unfold in two equal size steps, for example, in second and fourth pulls, or a larger transition plus a smaller transition if it misfolds (first and third pulls). This matches reported data but at lower forces, likely due to lower $Ca^{2+}$ concentrations (*48*). B) Calerythrin unfolds at high forces in one rip and refolds in two steps.

Because the refolding of CaE is very reversible and occurs in two steps, it is difficult to measure the kinetics from these force-extension curves both from a practical standpoint of scoring transitions in the data as well as a theoretical viewpoint where the rates are convoluted due to sequential barrier crossing. Practically, if a transition crossing occurs many times within one pulling cycle, at what force does one score the transition? The conversion of unfolding forces to force-dependent lifetimes discussed in Chapter 1 requires each unfolding event to be scored, but if unfolding and refolding are happening while the traps are moving, this barrier will not be sampled in an unbiased way even for a single transition, but is further convoluted for 2 transitions. Therefore we proceeded with constant force data (constant force data due to the drift

of the single-trap instrument). However, as shown in Figure 2.4, the force feedback was too slow to clearly resolve transitions.
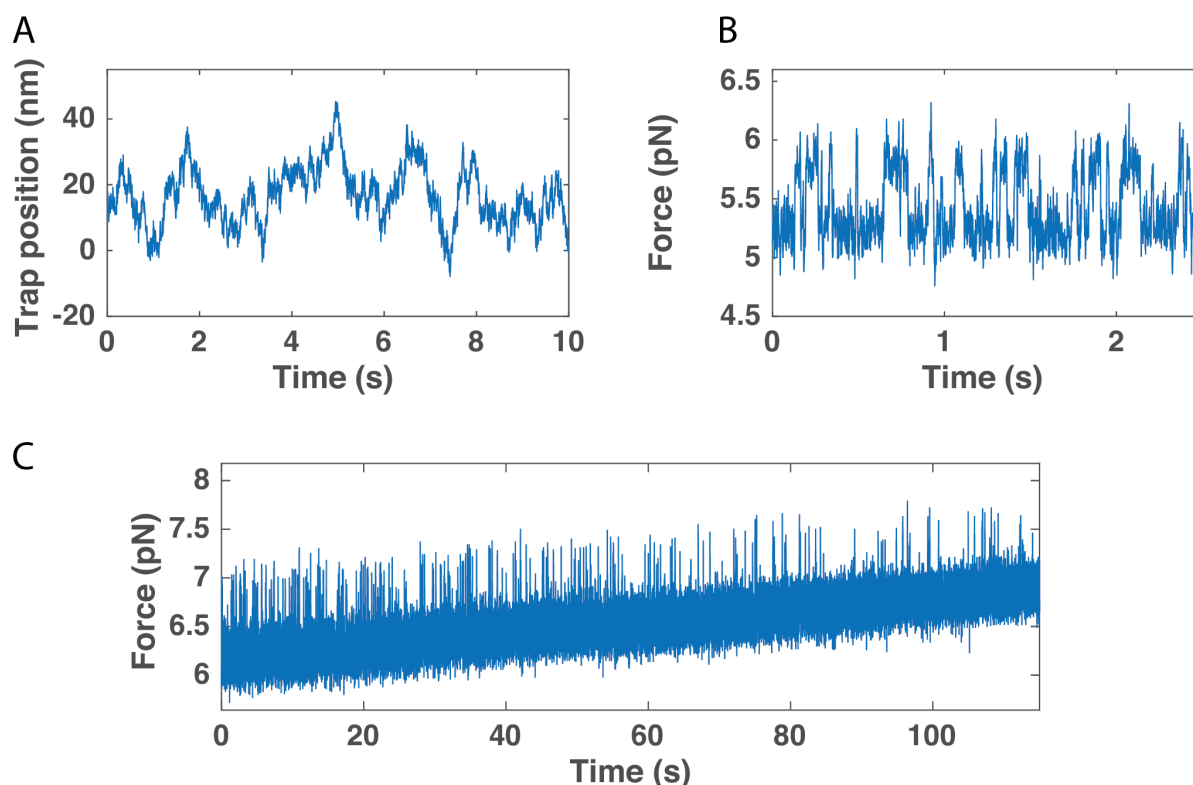


**Figure 2.4: Constant force versus passive data collection of CaE** A) Constant force feedback at 300 Hz, set to 5.2 pN, shows obvious transitions but very irregular features. B) Passive transitions are uniform and show two-state hopping. C) Unfortunately, without feedback like in (A) we are unable to treat the drift of the system. The drift corresponds to physical changes in trap position relative to the pipette, thus leading to a change in force over time which also changes the kinetics over time. Notice the hopping is less frequent at higher forces. This is expected but makes measuring exact kinetics difficult.

The data in Figure 2.4A clearly shows dynamics which are not part of the baseline, but state assignment is impossible. Without any feedback, as in Figure 2.4B, the folding pathway is distinguishable as well behaved "hopping". As mentioned in Chapter 1, the feedback readjusts the trap position every ~3 ms (300Hz). The kinetics in Figure 2.4B are not at this scale but are not orders of magnitude slower either, which leads to artifacts when the feedback is applied. This is an important example of instrument limitations for folding; many proteins (T4L, apomyoglobin (*28*), NCS-1 (*43*), SH3 (*49*), RNase H (*50*)) have been studied in the mini-tweezers without much issue; however, these systems all fold much more slowly than calerythrin. The refolding is quite fast, much faster than calmodulin.

This rate of folding mandates passive data collection, although such data is already preferred due to fewer missed transitions. However, as shown in Figure 2.4C, without drift correction from feedback, the mini-tweezers is not stable enough over a minutes-length timescale in order to

accurately collect data. Each transitions corresponds to a randomly sampled lifetime of the kinetics at that force, but if the baseline is changing the state will not be in equilibrium and thus we cannot tell if it is equilibrated and is a true sampling of rates at that force. Because our drift is relatively slow compared to folding, it is not unreasonable to make the assumption that the sampling is approximately in equilibrium. However, the fast kinetics pose another limitation for this instrumentation: inadequate temporal resolution and missed events even during passive collection.

The auto-correlation function of the mini-tweezers reveals that at 5 pN, the relaxation time of the tethered bead in the trap is 1.54 ms (Figure 2.5). This means that 1kHz data is correlated between data points and thus non-Markovian. We would need to downsample the data to at least a few times the correlation time to have truly Markovian statistics, limiting our time resolution to perhaps 100 Hz or so. The correlation time in a dual trap instrument at 5 pN, however, is 75 μs, essentially not limited by correlation time.
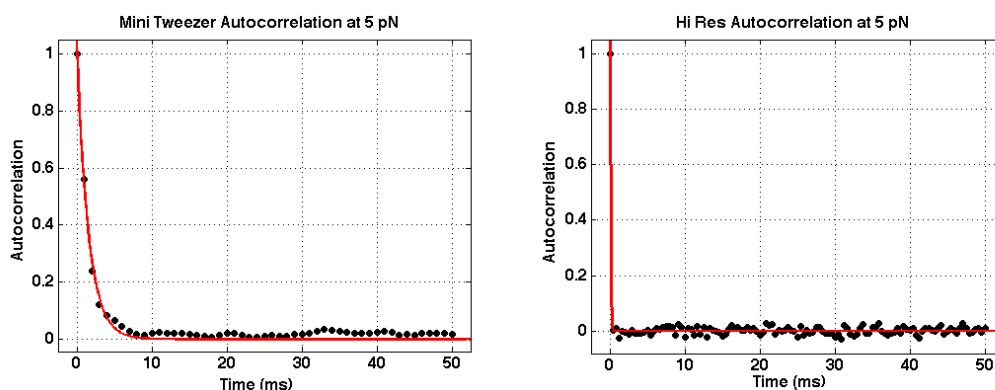


**Figure 2.5: Auto-correlation times at 5 pN for mini-tweezers versus dual-trap instruments**

## 2.3 Wild type protein results

In order to quantitatively measure the kinetics of our fast-folder, we developed a modified geometry which allows for tethering in a dual-trap instrument (Figure 2.6).
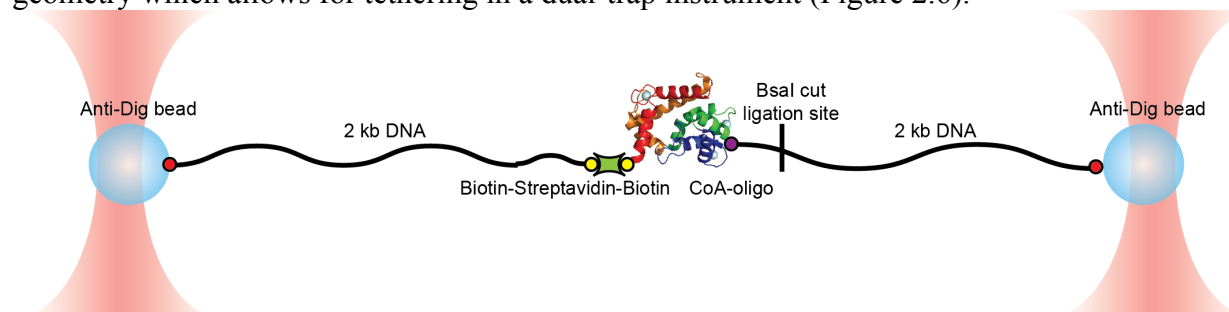


**Figure 2.6: Dual trap optical tweezers geometry.** Each side of the molecule is bound to DNA "handles" which have a digoxigenin-modification to allow binding to antibody-coated beads. The N terminus attaches to a streptavidin protein (SA), and the C terminus ligates to a digested handle.

The dual-trap optical tweezers requires a much higher intensity of light since 1) the beam is split into two traps and 2) there is are not two coincident beams to generate a single trap. Thus, practically, most dual trap instruments use 1064nm light instead of 840nm, due to the stronger lasers available in this range. However, 1064 nm lasers are more prone to heating due to the IR absorption of water, and they also can generate reactive oxygen species from dissolved oxygen, exacerbated by the surface of the polystyrene bead (*51*). We observed photodamage, visible as an irreversible cessation of folding or partial folding, within 1-5 minutes for the measurements performed with direct ligation of the protein to an oligo-coated bead as used in the mini-tweezers. Using the symmetric handle geometry spacing the protein away from the bead surface as shown in Figure 2.6 increases the average lifetime of the protein prior to photodamage to >10 minutes, more than sufficient to collect kinetics across a variety of forces (Figure 2.7).
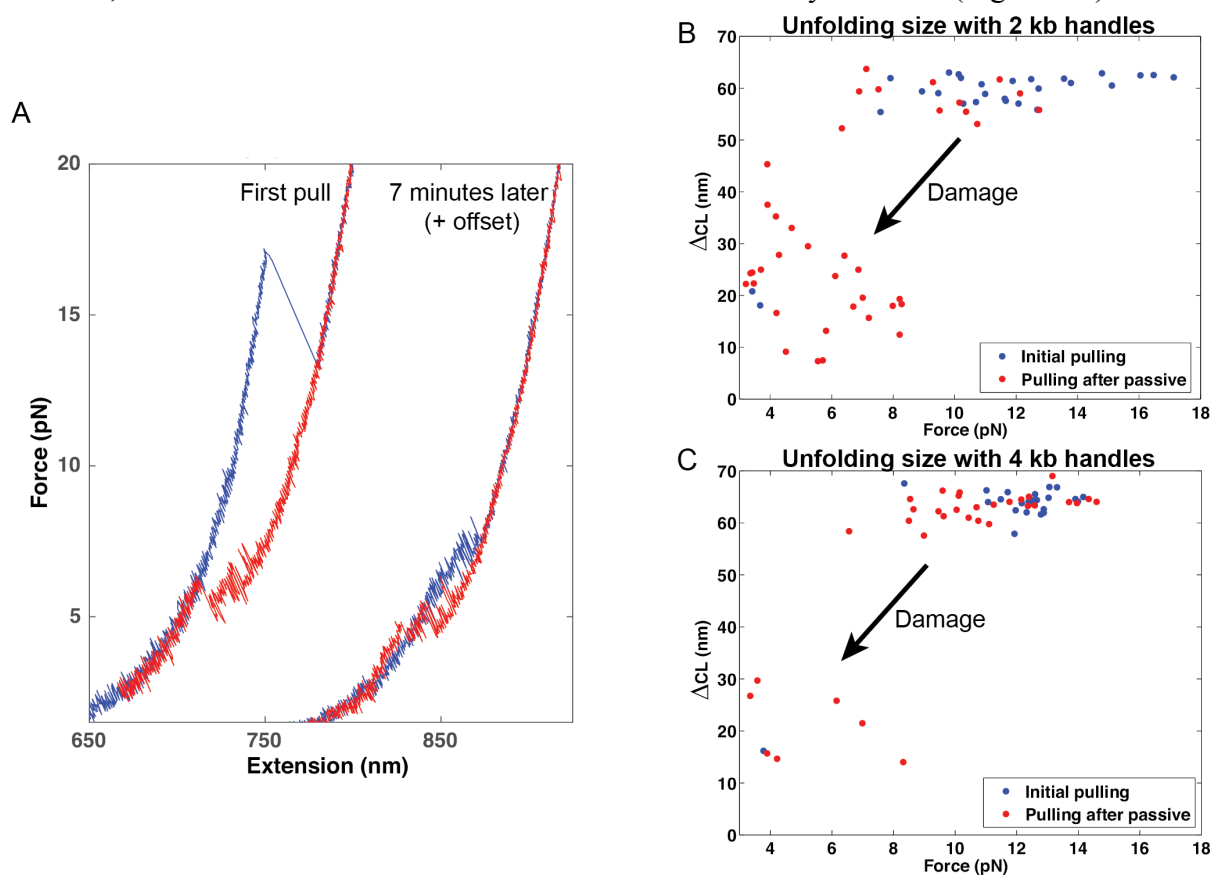


**Figure 2.7: Evidence of photodamage using short, asymmetric handles** A) Force extension curves before collecting passive data and after collecting passive data (approximately 7 minutes later) show a drastic change in the size of unfolding, one hallmark of photodamage. B) A scatter plot of the unfolding size from pulls before and after passive using 2 kb handles. Almost all of the unfolding transitions are centered at 60nm before passive (expected value 63nm), but afterwards they shift to a lower, broad distribution. C) Using 4 kb handles (2kb on each side) we still see some photodamage but the majority of tethers are unchanged after passive. This corresponds to about 1 in 8 tethers damaging within our experimental window (~10 minutes).

The easiest change to document is simply shorter/non-native folding, as shown in Figure 2.7A. However, the effect of photodamage is heterogeneous– it could lead to impaired folding, no folding, or partial folding. This is expected because photodamage is random and corresponds to a discrete physical oxidative change. Tryptophan, histidine, cysteine, tyrosine, and methionine are all documented to be potentially reactive to singlet oxygen species (*52*). Therefore, each time we observe photodamage it is likely to be a unique damaged species. In order to identify and exclude damaged molecules, we measure kinetics at all forces for a given molecule twice– the second measurement should give equivalent kinetic results because the single molecule folding data is ergodic. If, however, the data is not equivalent and if the change is not reversible, we assume it is photodamaged and exclude all data following that point. Often the photodamage can be seen as a discrete event in the data where hopping stops or is altered.

Using our new symmetric tethering geometry, we see qualitatively the same behavior as observed in the mini-tweezers: full-length calerythrin (FL) unfolds cooperatively where both domains unfold in one transition, but the domains refold sequentially. Observing the refolding in passive mode, we see an intermediate state that is highly reversible prior to full protein folding. Once the full protein has folded in passive we do not see subsequent unfolding during most experiments until we increase the force. By characterizing the kinetics of refolding at each force, we determined that the protein refolds through the same obligate intermediate ($FL_{int}$) each time before reaching the fully folded state, meaning the domains do not fold in either order (Figure 2.8).

Because this intermediate has an exponential lifetime, it is a single structural state and not an incorrectly assigned signal of two different states (i.e. observing both the N and C domain and being unable to distinguish them). This demonstrates that two folded states are not co-existing at a given force. The transition is $76 \pm 6$ amino acids; based on sequence assignment to conserved motifs and to known helical regions, the N domain can be defined as amino acids 1-90 and the C domain as amino acids 91-177 (87 residues) (*53*). Neither of these account exactly for the intermediate observed, although the C domain's terminal residues are not resolved well in the NMR structure, suggesting they are flexible and may not contribute to the compact state which is formed in the folding transition. The most likely assignment, based on the size, NMR structure, and CD titrations is that the intermediate is the C domain. However, to confirm this we proceeded to study domain isolates (truncations) and mutants of calerythrin.
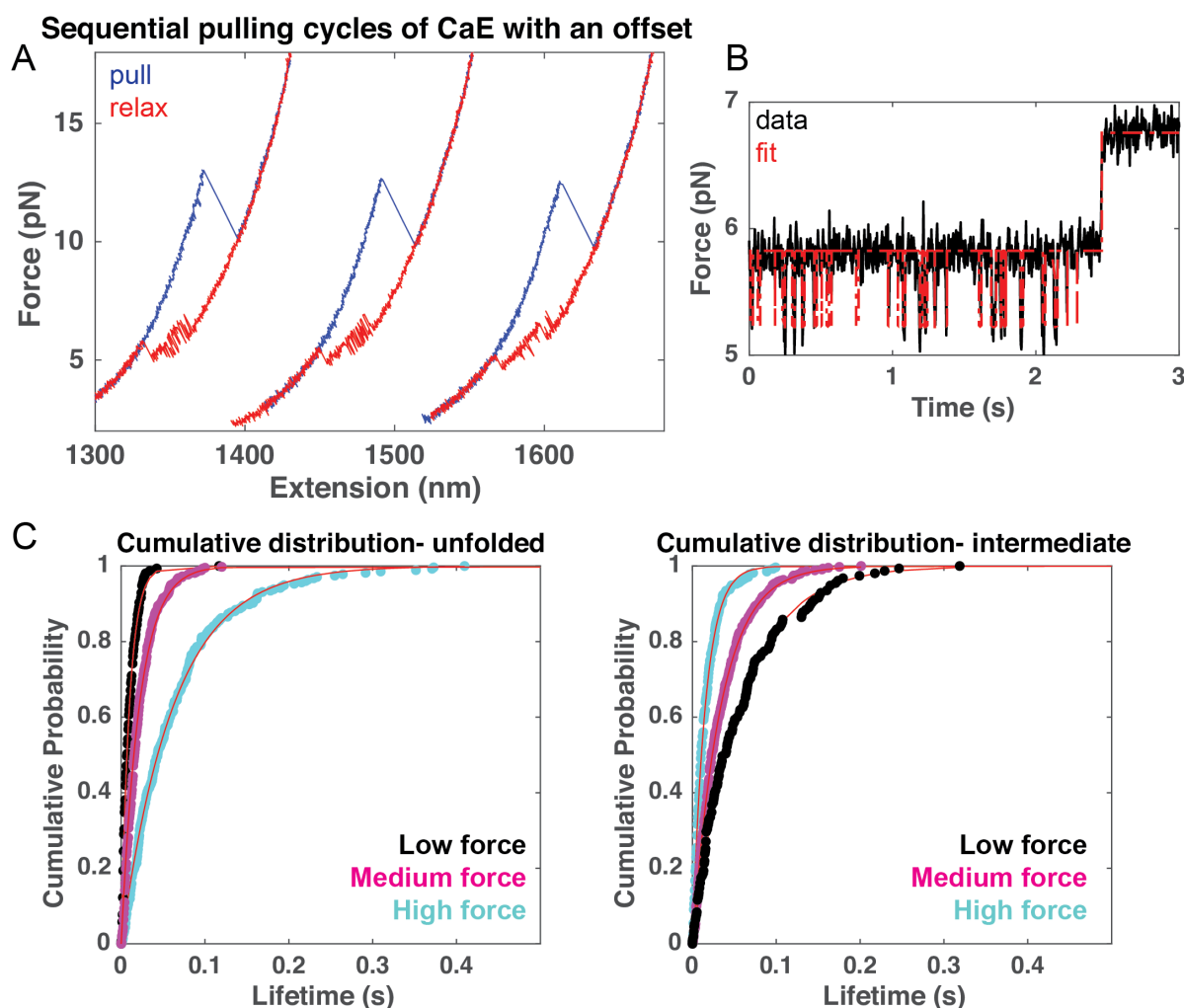
**Figure 2.8: Folding of FL$_{int}$ is one state** A) The pull curves show refolding in two steps. B) Passive data zoomed in to show the intermediate FL$_{int}$ (~5.8 pN) and the fully folded state (~6.8 pN). C) The distributions of lifetimes for the unfolded state and for FL$_{int}$ are well fit by a single exponential (red line) at all tested forces. Furthermore, the kinetics are linear (see Figure 2.9 further demonstrating they are the same state.

## 2.4 Truncations results

If the domains fold independently in the full-length protein, the isolated domains should have similar kinetics when measured without the complementary natural domain. Thus, we designed and cloned molecular constructs of truncated proteins corresponding to the domain isolates. The domain boundaries were assigned according to the helices from the NMR structure (*53*). This criterion establishes the N domain as amino acids 1-90; although there is a short section of amino acids prior to the beginning of the EF3 helix, the C domain construct included all the remaining amino acids 91-177 in order to span the entire protein. The truncations contain the same N and C terminal linkers used for the full-length protein: an N-terminal Avi tag with a 21-amino acid serine-glycine linker and a C-terminal ybbR tag without any additional linker.

Folding of the isolated C domain is indistinguishable from $FL_{int}$ in the folding and unfolding kinetics, (Figure 2.9). In contrast, folding of the isolated N domain is slower and typically occurs at lower forces than $FL_{int}$ or the C domain ($p<0.05$) (Figure 2.9). The data for the fits in Figure 2.9C is listed in Table 2.1 at the end of the chapter.
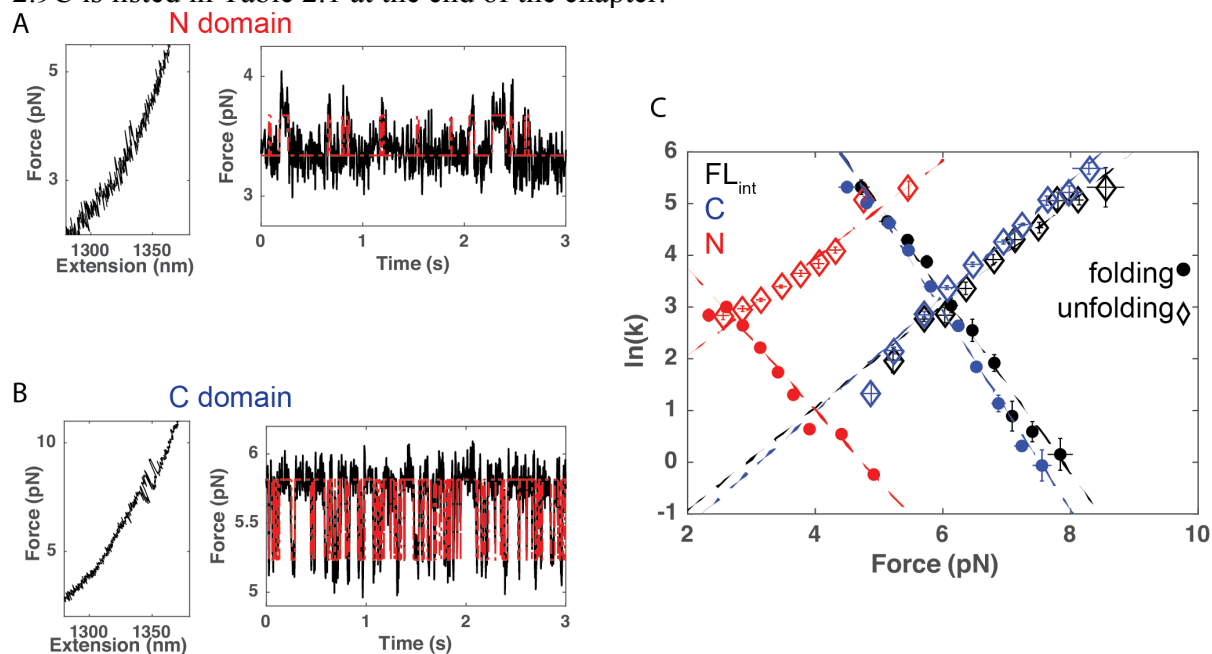


**Figure 2.9: Folding of the N and C domain isolates.** A) The N domain isolate folds around 4 pN, seen in both force-extension and passive. B) The C domain isolate folds around 6.5 pN, similar to $FL_{int}$. C) The force-dependence of the kinetics of folding and unfolding is indistinguishable between the C domain and $FL_{int}$. The N domain folds significantly more slowly (p<0.05).

Additionally, the OT also allows us to measure the size of the transitions, which corroborates the kinetic results as well. From our data, we find that the C domain transition is $72 \pm 8$ amino acids, $FL_{int}$ is $76 \pm 6$ amino acids, and the N domain is $98 \pm 12$. This size if a little larger than expected for the N domain (90 amino acids), but the forces of the N domain are lower and thus may be less accurate for conversion; it is also possible it is not the native structure (less compact) without the C domain present or interacts somewhat with the linkers. Following these kinetic and size measurements, we can establish that the folding pathway of the isolated full-length molecule proceeds from the unfolded state through the C domain to the native structure. This was also confirmed via mutants, described in section 2.5.

The other truncation that we designed was for EF123: a partial C terminal deletion that lacks the fourth EF hand, only consisting of hands 1-3. This truncation would be relevant in the course of translation as a transient intermediate while the ribosome is synthesizing the fourth EF hand. Although of course there are many truncations that would be relevant to translation, this one was chosen as an initial test case since it uses known helical boundaries, rather than a truncation in the middle of a helix. The final amino acid in the second helix of EF3 would be at residue 133;

we used 134 as the cutoff (within an unstructured region between the helices of EF3 and EF4; the first helix of EF4 does not start until amino acid 137).

This truncation does not fold to the N domain state; instead, a misfolded state that corresponds to the folding of the entire available sequence is formed. As shown in Figure 2.10, the pulling curves of EF123 are significantly different than N domain folding.
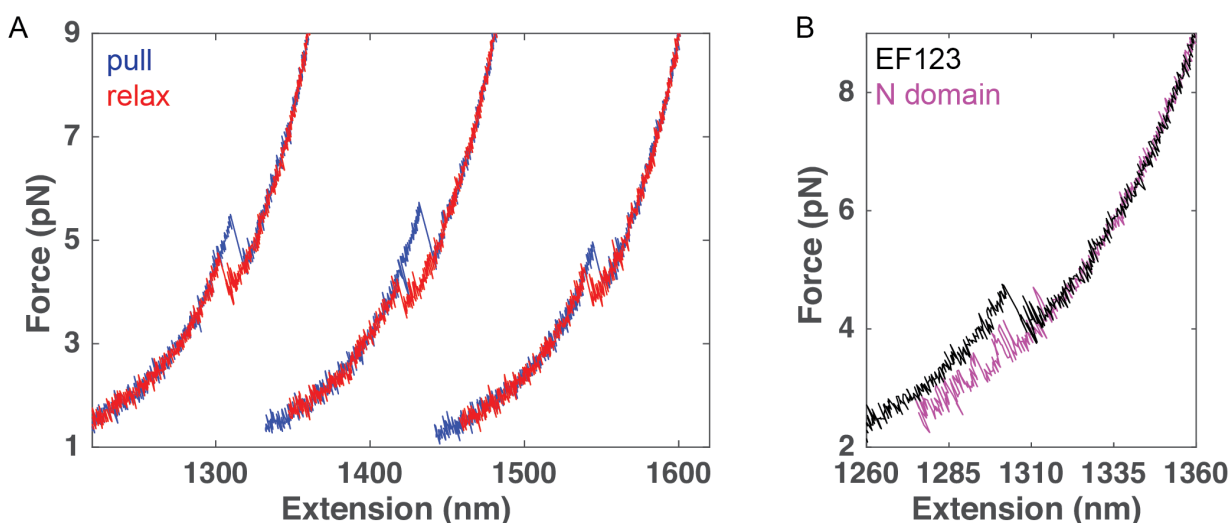


**Figure 2.10: Force-extension curves of EF123** A) Sequential pulls on one molecule of EF123 show a slightly hysteresis between unfolding and refolding. B) The refolding of EF123 (black) is larger than the N domain (magenta), although transient states prior to stable folding are similar to the N domain.

Although the pulling curves always show the same folded transition during unfolding, in passive the data cannot be fit to a two state model of hopping between a folded and unfolded state. Instead, the dominant structure is a misfolded state ($I_1$) of $141 \pm 8$ amino acids and there is also a less stable, transient folded state ($I_2$) of $83 \pm 3$ amino acids. Based on the forces of folding for $I_2$ and the size, it is likely that it corresponds to the N domain state, although we cannot measure kinetics for a wide range of forces since the $I_1$ state will dominate the data at forces below 4 pN, where the N domain folding is best resolved according to our domain isolates.
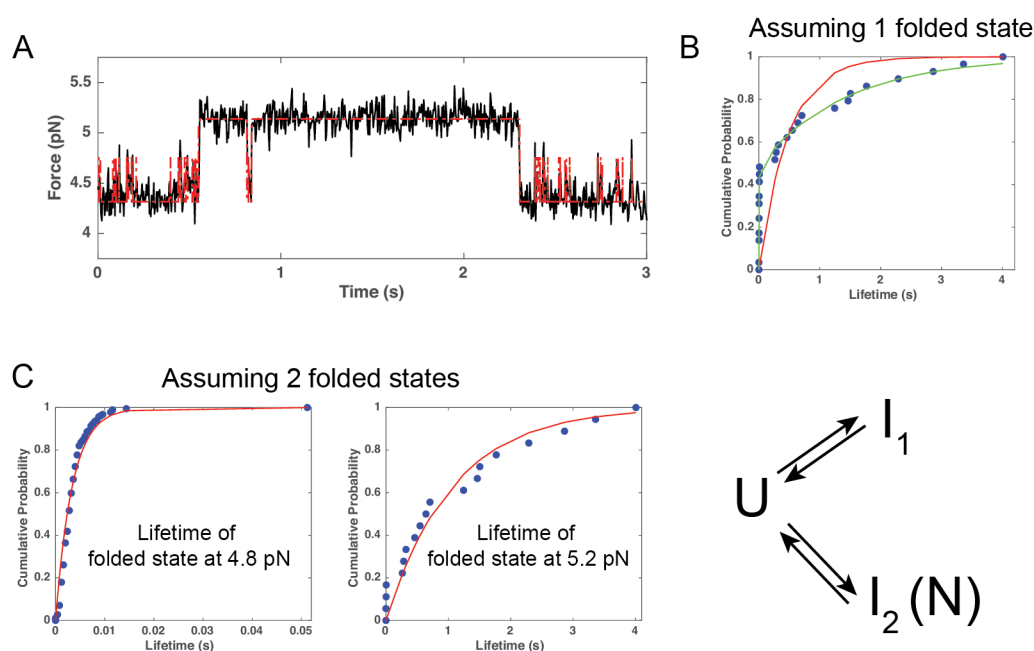
**Figure 2.11: Folding of EF123 is two-state** A) An example of passive data for EF123. A second, transient folded state is observable as short spikes in the data. B) If a HMM restricted to two states is fit to the data, it cannot converge to exponential fits, and the lifetime CDF shows a mixture of two exponentials, which suggests two parallel states are misassigned. The exponential fit is shown in red and a double exponential fit is shown in green. C) A HMM with three states (two folded states) fits the data robustly. The state connectivity of the best-fit HMM is shown, where $I_1$ and $I_2$ (the N domain) do not interchange.

For a simple 3-state system, we can solve or approximate for the expected CDFs of various state connectivity schemes where two states are grouped together into one observed CDF– this is discuss further in the methods (section 2.6, subheading "Expected distributions for under-fit HMMs"). Briefly, our "two-state" HMM applied to EF123 yields a double-exponential distribution, which results from parallel states. In a similar vein, a 3-state fit to our data is best described by a model where the unfolded state can fold to either $I_1$ or N domain, but the two folded states cannot directly transition to each other, i.e. they are independent and mutually exclusive folded states and one is not an intermediate to form the other (Figure 2.11). Because the misfolded and folded N domain states of EF123 do not interchange without first reverting to the fully unfolded state, this suggests the misfolded state could be due to mispaired EF hands, similar to a previously reported misfolded state of calmodulin (*54*). In this case, EF1 and EF3 could pair and form a more stable domain than EF1 and EF2. This is somewhat surprising, but EF2 is non-canonical and so may be out-competed for pairing with EF1 by the EF3 fold if EF3's binding partner EF4 is not available (not present, as in this case, or potentially sequestered in the tunnel in the context of translation). This misfolded state is off-pathway for the nascent chain, requiring non-native interdomain contacts that preclude the correct folding of the individual domains.

## 2.5 Mutants results

Glutamine mutants

As further confirmation that the intermediate is the C domain, we individually mutated the conserved $12^{th}$ position glutamate/aspartate to glutamine in the EF-hand loops of hands 1, 3, or 4, thus impairing calcium binding and destabilizing a single EF-hand fold (*55*) (Figure 2.12). These mutants can be labeled as QEF1, QEF3, and QEF4, respectively.
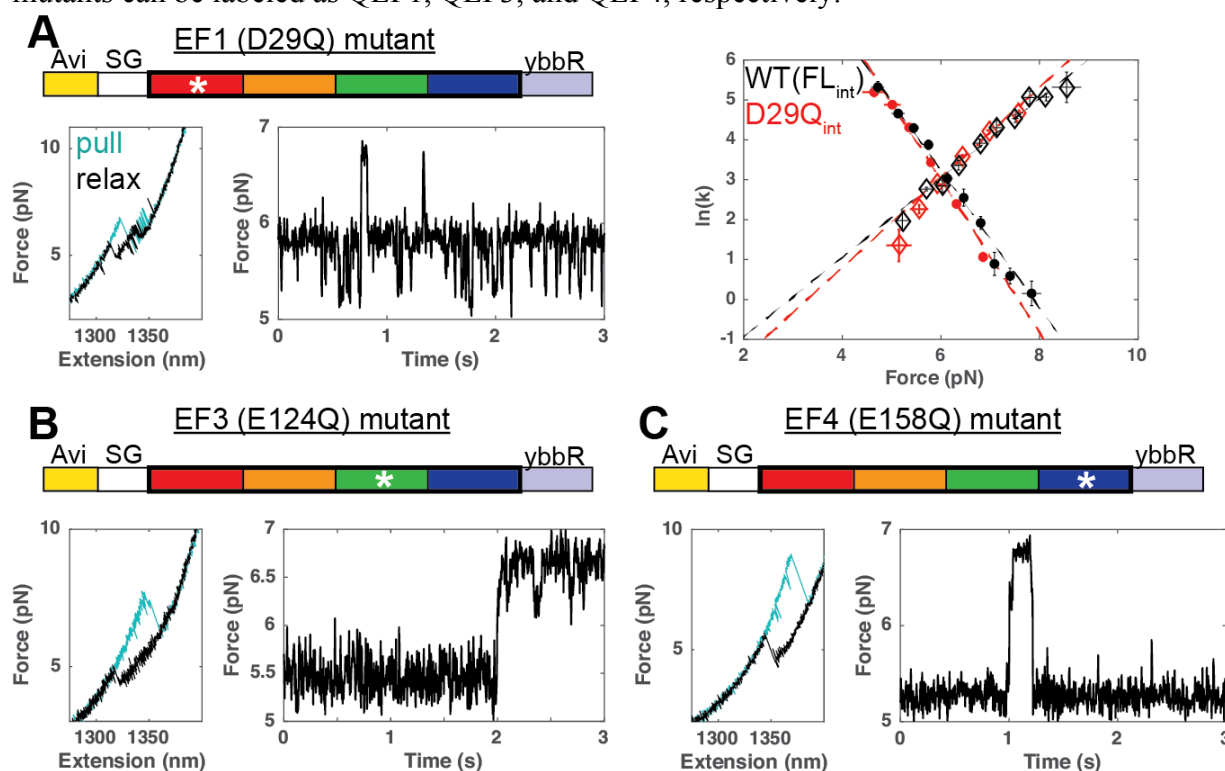


**Figure 2.12: Folding of Glutamine mutants (calcium-deficient)** A) The QEF1 mutant (D29Q) shows two-step (non-cooperative) unfolding and faster unfolding from folded state to the intermediate, but no effect on the intermediate's kinetics (compare $FL_{int}$ to $D29Q_{int}$), consistent with a C domain intermediate. Error bars are SE. B) The QEF3 mutant (E124Q) destabilizes the intermediate (folding begins at ~5 pN instead of 7 pN). C) The QEF4 mutant (E158Q) shows the same qualitative behavior as the EF3 mutant, but refolding occurs at even lower forces. A similar partial unfolding of the fully folded state is seen. HMM fitting was not performed on the QEF3 and QEF4 mutants due to the very short lifetimes of the intermediate.

The glutamine mutation in EF1 (belonging to the N-terminus domain) did not affect formation of $FL_{int}$. This intermediate folded and unfolded with the same kinetics (Figure 2.12A), meaning the D29Q mutation does not likely affect this state and $FL_{int}$ does not include that portion of the protein. In contrast, the rate of unfolding from the fully folded state was inflated– unlike in WT protein, the fully folded state was not an absorbing state in passive, and transitions could be seen in and out of this state frequently. This identifies the second transition as the N-domain folding, meaning the first transition must be the C domain. The effect of the mutation was also observed in force-extension curves, where the QEF1 mutant unfolded in two steps, the mutation having abrogated the cooperativity/stabilization between the domains.

In contrast, mutation of the critical calcium-binding residue in EF3 or EF4 does destabilize $FL_{int}$, so that it only folds at lower forces and unfolds more rapidly. Interestingly, in both QEF3 and QEF4, the fully folded state is not stable and one hand (not one domain) spontaneously unfolds (the downwards jumps from the state at 6.8 pN in Figure 2.12B or C). For QEF3, it is likely that this state corresponds to partial unfolding/undocking of EF4 due to the perturbation of the interface with EF3, not a partial unfolding in EF3. A partial unfolding of EF3 would likely also cause EF4 to unfold since it would not have a binding partner nor interaction interface, thus potentially showing a larger change in length. In QEF4 this partial unfolding is likely unfolding/undocking of the mutated EF4 hand. These assignments are speculative, however, and further work would need to be done to unambiguously assign them. It is also likely that the QEF3 or QEF4 mutants are not in a native state since the primary calcium binding sites are knocked out; the observed states may be molten-globule like. Regardless of the underlying structure, it is clear that these mutants have impaired formation of $FL_{int}$. Taken together, all three of these mutants confirm that $FL_{int}$ is indeed the C domain.

Proline mutant

Calerythrin also has an interesting feature in that it has a proline in the middle of an alpha helix in EF3, which provides a kink in the helix's structure (Figure 2.13). Prolines are rare within helices, as they distort the hydrogen bonding along the peptide backbone (*56*). Prolines are also known to have an effect on folding, with many proteins having a slow kinetic phase due to proline isomerization, where the protein can only properly fold when the proline is in either the *cis* or *trans* configuration and thus folding is delayed for a subset of the population until the proline configuration reverses.

The geometric constraint of the cyclic side chain in proline makes the normally heavily favored *trans* state less dominant than in non-cyclic amino acid side chains (*57*). Return to the *trans* isomer of a *cis* peptide bond with an adjacent proline side chain is much slower than isomerization of other peptide bonds due to the steric clash of the proline ring (Figure 2.13). Isomerization of the amide bond for non-proline residues was measured in small peptides to be between 0.86–2 $s^{-1}$ for *cis*→*trans* and (0.86–1.6)x$10^{-3}$ $s^{-1}$ for *trans*→ *cis* (depending on the side chain) (*58*). However, for small peptides with proline, the *cis*→*trans* rate was measured to be 0.05 $s^{-1}$ and 6.6x$10^{-3}$ $s^{-1}$ for *trans*→*cis* (*57*). The *cis* state lifetime is thus longer, while the *trans* state is slightly destabilized by the constrained ring.



A ... *trans* proline ... *cis* proline   B
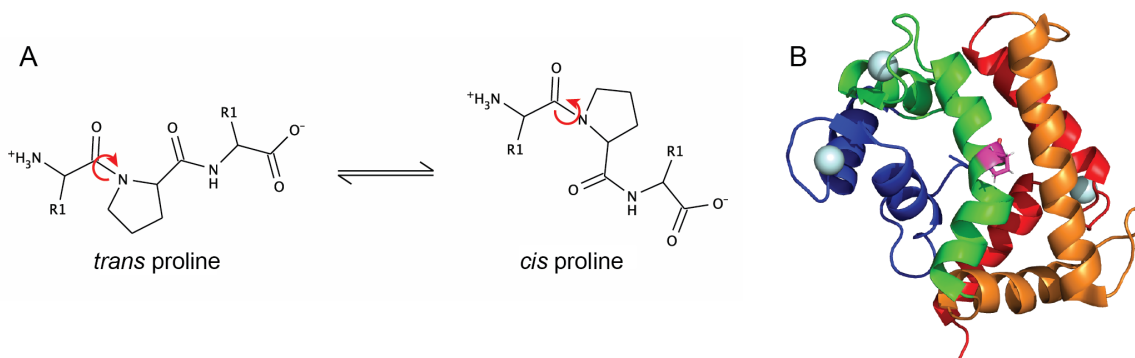
**Figure 2.13: Isomerization of a X-Pro-X tripeptide; visualization of the proline in calerythrin** A) The peptide bond before a proline can exist in 2 stable conformations which slowly interchange. Once the protein is folded, the bond will likely not interchange further. The peptide bond which undergoes rotation is marked with a red rotation arrow. B) The proline in calerythrin is in the *trans* conformation within EF3, and is highlighted in magenta.

Since synthesis generates peptides in the *trans* configuration at the peptidyl transfer center in the ribosome, calerythrin's proline is in the *trans* state, and isomerization is slow, we do not expect the proline isomerization to be a rate-limiting step for calerythrin folding except in a small subset of molecules. Still, isomerization has been reported before for optical tweezers experiments over long time scales for a known proline-switch protein (*59*). In a similar vein, we noticed in WT calerythrin that sometimes the folding of the C domain would halt temporarily and then resume tens of seconds later (Figure 2.14).
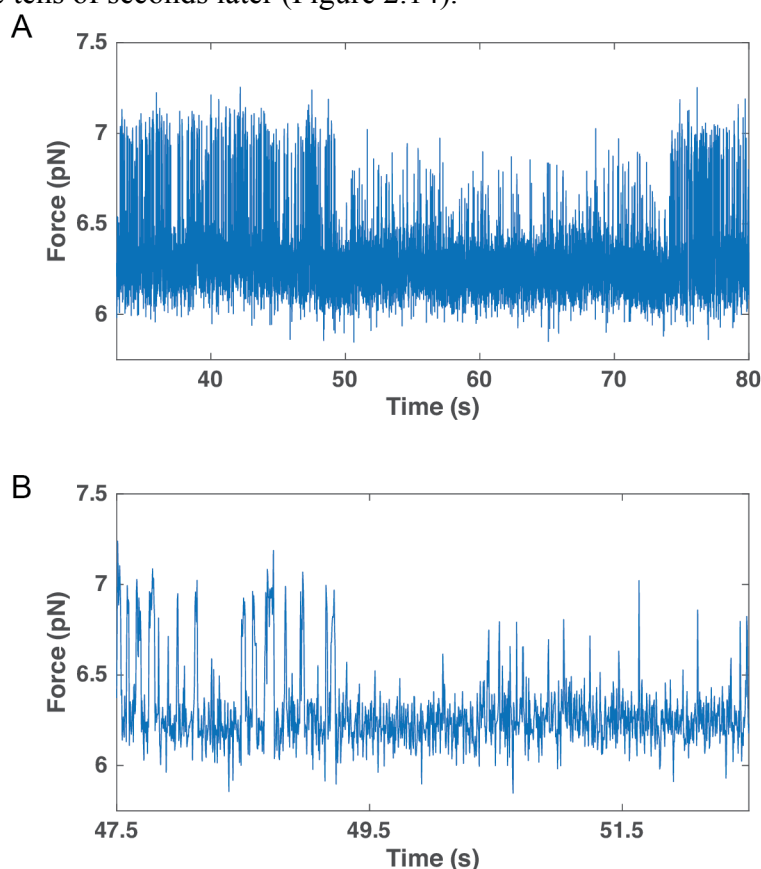


**Figure 2.14: Reversible proline isomerization in calerythrin** A) an example of calerythrin folding normally, then suddenly stopping and folding to a smaller, shorter lived state (from 49.5 seconds to 74.5 seconds). After 25 seconds, normal hopping resumes. B) A zoom into panel A that shows the transition region from *trans* to *cis*.

Because this process is reversible, it cannot be due to photodamage of the molecule. We observed the impaired state to be approximately exponentially distributed, with a mean time of 20.4 seconds, corresponding to an exit rate of 0.049 s$^{-1}$. The rate of entry from normal hopping to the impaired state was 0.009 s$^{-1}$, however this is likely an underestimate of the rate of entry

because this is estimated from the number of events observed per time of hopping, thus including the time spent in FL$_{int}$, where it likely cannot isomerize due to constraints of the backbone. The true rate is probably about two times faster, based on the crude assumption that within our resolvable hopping region it is unfolded about half the time.

We cannot report on the force-dependence of this state because the events are so rare (on average twice per molecule during data collection) that all lifetimes were grouped together to find the rate. However, it has been previously reported that isomerization is force-dependent with higher forces favoring the *trans* configuration (*59*). Additionally, we are only sampling lifetimes from 5-7 pN, a relatively small range, because at low force it is folded and at high force, when fully extended and no folding is observed, we cannot differentiate the *trans* versus *cis* states. Despite these limitations, these rates match surprisingly well with the reported rates derived from small peptide assays.

Based on the relative lifetimes of the folding-competent state and the impaired region, as well as the native fold arising from the folding-competent state, this excursion likely represents the *cis* state of the peptide. The folding that we observe during the *cis* region may be partial folding of the C domain up until the proline residue. It is unlikely to be the N domain based on the size. Any regions of isomerization were excluded from the kinetic analysis of the WT protein discussed earlier.

Although this phenomenon matches the expected timescale of isomerization, we also decided to test it by mutating the native proline to alanine (P102A mutant). The P102A mutant displayed the same qualitative folding as WT calerythrin (Figure 2.15) but was never observed to enter this impaired state (0 times in 1611 seconds; in WT protein you would expect to observe the *trans→cis* transition 14 times in this time window). Because we are not interested in the precise kinetics of folding for this protein, this construct was only analyzed in the mini-tweezers instrument. It is somewhat surprising that the proline does not seem to be required for native-like folding given that it is in an unusual helix position, although there may be a modest effect in overall stability (average force of unfolding of the full folded state from force-extension data in mini-tweezers (pulling 100nm/sec) was $14.3 \pm 0.3$ pN for WT compared to $11.7 \pm 0.3$ pN in the P102A mutant, showing $\pm$ SE).
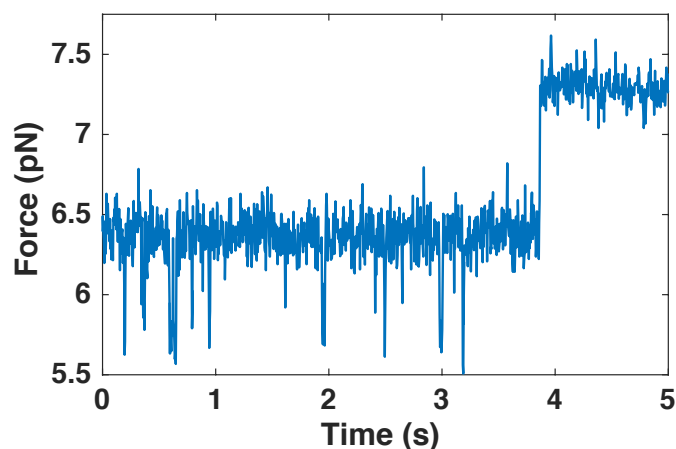
**Figure 2.15: Folding of the P102A mutant** The P102A mutant displays the same folding behavior (U→FL$_{int}$→Folded), but it never enters the impaired folding (*cis*-Pro) state.

We also sought to test if the *cis* state could be induced or shortened through the introduction of peptidyl-prolyl isomerases (PPIs). We purified cyclophilin and parvulin as possible enzymes to catalyze the interconversion. Both enzyme preparations were active based on a standard small peptide trypsin assay (*60*). However, the results in the OT were inconclusive when adding micromolar levels of protein to the buffer in the fluidic chamber. Studying the isomerization kinetics is very difficult on the OT because the isomerized state is rare and random (it is not inducible). It is hard to generate sufficient statistics to prove an effect. It is also unclear how PPIases recognize their substrate and if having the two protein termini tethered or having the protein extended under force will hinder substrate recognition. This system may be an interesting model system to pursue to study PPIases in rescuing protein folding, as compared to isomerization in proline switches as previously reported (*59*).

## 2.6 Methods

Cloning, expression, and purification of proteins

The calerythrin gene was purchased (UniProtKB – P06495, IDT gene block), amplified, digested (NotI, NheI from NEB) and inserted into a plasmid with flanking Avi and ybbR tags. Truncated constructs for released proteins were prepared via PCR with internal primers containing inserted restriction sites and subsequent digestion with the same set of enzymes followed by cloning into the same backbone vector. Site-directed mutagenesis for mutants was conducted via whole-plasmid PCR amplification with non-overlapping primers with mismatches in the forward primer, followed by DpnI (to digest the parent vector) treatment, then PNK treatment and blunt-end ligation. All constructs were verified by plasmid sequencing.

The domain isolates were tested for expression and solubility prior to purification and OT analysis due to concerns of limited stability, particularly for the N domain which does not have a canonical fold. 6mL test cultures were induced with IPTG for 2 hours at 37, then treated with lysozyme for gentle cell lysis conditions. The total cell lysate, supernatant after pelleting, and the pellet resuspended in detergent were ran on a gel (Figure 2.16). This shows that although the N domain is less soluble and a significant portion goes to the pellet, there is still a soluble fraction of protein.
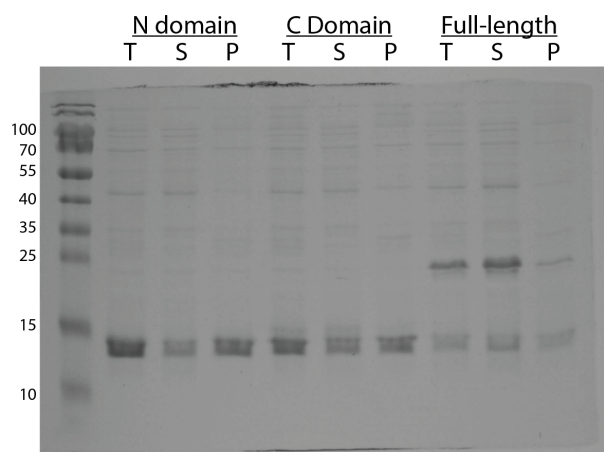
**Figure 2.16: Solubility test for domain isolates.** After a 2-hour induction, cells were gently lysed, then pelleted. T = total S = supernatant (soluble fraction) P = pellet (insoluble fraction). The N domain with tags is expected to be 15 kD, as is the C domain. The full-length protein with tags is 25 kD.

Wild-type calerythrin, calerythrin truncations, and calerythrin mutants were expressed in BL21 E. coli cells with IPTG induction (1mM for 2 hours at 37º C or 4 hours at 25º C), lysed, and biotinylated in vitro. Biotinylation of the lysate was conducted with 1 μM recombinant birA, 25 μM biotin, 5mM ATP, and 5mM $Mg(OAc)_2$ at room temperature for 40'. Excess biotin was removed by overnight dialysis in 1X PBS with 0.5 mM PMSF at 4 degrees. Biotinylated protein was purified on monomeric avidin resin using standard elution procedures (Pierce). Excess biotin after elution was removed by overnight dialysis at 4 degrees in 20 mM HEPES, 100mM KCl, and 0.5 mM $CaCl_2$. The protein sample was concentrated via Amicon centrifugal filters (Millipore, MWCO 8kD or 3.5 kD) and stored at -80.

Synthesis and purification of oligo-modified CoA

Although this method is previously reported, we have modified and optimized the purification procedure, and thus have included it here. The crosslinking chemistry uses the natural thiol within coenzyme A to attach an amine-modified oligo. GMBS is used as a dual-functional crosslinking agent according to the scheme below (Figure 2.17):
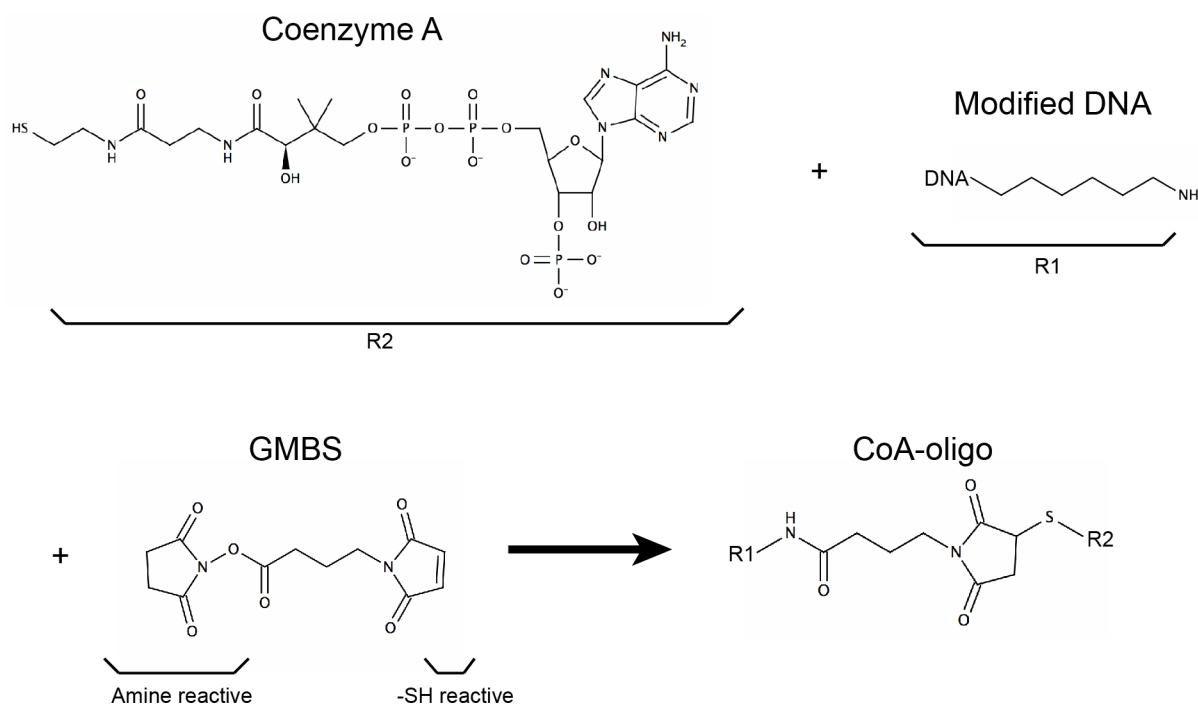
**Figure 2.17: CoA oligo reaction schematic**

The coA-oligo was prepared by reacting 40 nmol of amine-modified oligo (IDT, resuspended in water to 1mM) with 3.75 mmol GMBS (50 mM in DMSO) in PBS buffer with 1 mM EDTA for 30 minutes at room temperature. After this time, pelleted material was redissolved in additional DMSO and added back to the reaction supernatant for 15 minutes to allow for further reaction. This reaction of the amine with GMBS was quenched with excess Tris, ethanol precipitated, and resuspended in 50 mM HEPES with 2.5 mM EDTA to prepare for the next step. 1.25 mmol of coA (97 mM stock in water) was added and reacted for 1 hour. The reaction was ethanol precipitated and resuspended in formamide loading dye (95% v/v formamide, 0.025% w/v bromophenol blue 0.025% w/v xylene cyanol, 0.025% w/v SDS, 5mM EDTA).

For purification, this sample was loaded onto a 15% acrylamide, 8M urea gel (sequencing gel, approximately 20 cm long). After running, the gel was imaged via UV shadowing, which usually shows a doublet of bands. The top band is the modified oligo and the lower, weaker band is unreacted oligo. The CoA runs off the gel. The cut bands were extracted via diffusion overnight with shaking at 600-800 rpm in gel elution buffer (200 mM Tris pH 7.5, 25 mM EDTA pH 8, 300 mM NaCl, and 2% w/v SDS). The sample in diffusion buffer is quite dilute, so this was further concentrated by performing a phenol chloroform extraction, adding 70% v/v ethanol, cooling on ice 1 hour, then ethanol precipitating. The pellet from the precipitation was resuspended in 10 mM Tris with 1 mM EDTA, pH 8, and stored at -80. This procedure can provide yields of 60-65%. The oligo must be annealed to the complementary strand prior to use in the Sfp conjugation reaction. The complementary strand includes the overhang to the BsaI site, and so must be 5' phosphorylated for ligation.

Optical tweezers samples

The optical tweezers sample consists of a protein tethered between two DNA handles. One handle is biotinylated, and through a streptavidin bridge it connects to the biotinylated substrate. Because the SA can bind up to 4 biotins, large excesses of SA (>1000x) are required so there are not multiple handles or proteins bound per SA. This DNA handle was generated via PCR of a 2kb portion of the lambda genome with digoxigenin modifications in the reverse primer and biotin in the forward primer, then gel purified from 0.8% (w/v) agarose in TAE, ethanol precipitated, resuspended in 10 mM Tris, and stored frozen at -80$^{o}$C or -20$^{o}$C.

The other handle is restriction digested to provide an overhang which is complementary to the chemically crosslinked CoA-oligo group. Both handles have digoxigenin modifications to attach to anti-digoxigenin coated polystyrene beads. The DNA handle ligated directly to the sample was generated via PCR of the same 2kb region of the lambda phage genome with a digoxigenin modified reverse primer and an inserted BsaI site in the forward primer. The PCR product was digested with BsaI to produce the 4nt overhang then gel purified from 0.8% (w/v) agarose in TAE, ethanol precipitated, resuspended in 10 mM Tris, and stored frozen at -80$^{o}$C or -20$^{o}$C.

Released protein samples were CoA-labeled by reacting the purified protein (20 μM) with recombinant Sfp (10 μM) and CoA covalently modified with a 20bp ds-oligo with a 4nt overhang (35 μM) in 50 mM HEPES pH 7.5 and 10mM MgCl2 for 1 hour at 30 degrees. This reaction can be stored at -80 for future experiments or further diluted two to ten-fold and ligated to the DNA handles. The ligation to DNA handles was conducted in 1X T4 ligase buffer (NEB) with 6 U/uL T4 ligase at 16 degrees for 1 hour. The ligated sample was deposited on BSA-passivated α-digoxigenin coated 1 micron polystyrene beads and diluted in 1X polymix (20 mM HEPES, 95 mM KCl, 5mM MgCl$_2$, 5 mM NH$_4$Cl, 0.5 mM CaCl$_2$, 1 mM spermidine, 8 mM putrescine, 0.1 mM DTT, 0.01 mM PMSF, and 0.1 mM benzamidine) supplemented with 10 mM NaN$_3$ for use in the optical tweezers. The biotin handle was deposited on a separate aliquot of BSA-passivated α-digoxigenin coated 1 micron polystyrene beads, then saturated with a more than 1000-fold excess of streptavidin and also diluted in 1X polymix with azide.

Data Analysis

Data were collected on a dual-trap optical tweezers set-up, previously described (*61*) and set to a trap stiffness of 0.25-0.26 pN/nm. Data were collected at 2500. Data were used at full frequency for all fittings and analysis, but are displayed downsampled at 250 Hz for equilibrium data figures or 125 Hz for force-extension curves.

Data analysis of equilibrium folding kinetics was done using a Hidden Markov Model (HMM) python library (hmmlearn) and lifetime fitting in Matlab. Fits were also performed with varying numbers of states, with the best fit selected according to the lowest Bayesian Information Criterion (BIC). Rates were obtained by fitting a cumulative density function to fit lifetimes (*33*). Because the protein folds and unfolds very quickly and each molecule was tested at multiple forces, one molecule will yield approximately 10 rate measurements at 10 forces, from 1000 individual folding events. The force data was used for passive analysis with the HMM

rather than extension data due to offsets observed between bead pairs in extension, likely due to bead size variations. Each bead pair was calibrated after each experiment.

For composite binned data as shown in the Bell plots, the rate data for all molecules for each condition was binned into equally spaced bins along the force axis (x axis). The number of bins was determined by the square root of the number of rate measurements. Then, the linear fit was performed with weighting on the binned data that was $1/\sigma^2$ of each bin or $1/\sigma^2$ from the lifetime fits within the bin, whichever was a greater source of variation (usually the spread of the bin is greater, except in bins at very high or low force, which have more error in the fit and are also less populated). It is more appropriate to use a weighted linear fit because our data does not have equal variance among all data points (this trait is heteroskedasticity)– the certainty with which we can determine the rate is dependent on the number of transitions observed, and there are fewer folding or unfolding transitions at high or low forces.

The weighted linear fit fit our overall data quite well. Although more exact models exist that are non-linear (*29*), they do not converge well when applied to our data set likely because the data only covers a small range of forces (~ 2 pN) such that we do not sufficiently sample the non-linear regions. To test if values from the linear fits ($\ln(k_0)$ or $\Delta x^\dagger$) are equivalent for different experimental conditions, we used a two-tailed Welch's *t*-test for two values with unequal variances. The variances were derived from the fit parameters in Matlab. The number of binned data points in the fit and the variances determine the degrees of freedom according to the Welch-Satterthwaite equation. The null hypothesis is that they are from the same distribution. A *p* value of <0.05 indicates they are significantly different values.

To convert force data into amino acids, we used a combination of two worm-like chain models. For the DNA, we used an extensible WLC model with persistence length P = 35 nm and stretch modulus S = 1200 pN. The WLC parameters were determined by pulling DNA-only tethers in our same buffer conditions. These parameters are as expected for buffer conditions with high magnesium and polyamines (*62*). For the polypeptide contribution, we used a worm-like chain model with a persistence length of 0.6 nm. A value of 0.6 nm fit our protein tether behavior as well as fit the overall unfolding signature size in force-extension curves, and is similar to previously reported values (*42, 63*).

<u>Expected distributions for under-fit HMMs</u>

For a single kinetic step, the probability density function (PDF) is $ke^{-kt}$. This describes the distribution of times to cross a single barrier, or is an approximate solution for a scenario with one rate-limiting step (i.e. multiple kinetic steps but one step is of a much longer time scale and thus dominates). As mentioned before, the PDF, while true, requires binning the data to generate and thus is prone to distortion due to bin size effects or limited sampling. Instead, we can use the CDF, which we can easily generate from the data by making a sorted list of lifetimes and normalizing the probability of finding an event of length τ or shorter. This mathematically corresponds to the integral from 0 to τ of the PDF. For the single step, the CDF is $1 - e^{-kt}$. Since all the underlying transitions in a HMM are single steps, all the distributions of a HMM that adequately describe the underlying states would follow this form.

Although we will not solve a variety of state connectivities directly, we can consider some general features of underfit-HMMs. Consider the case for EF123, where 3 states exist with connectivity 3 $\leftrightarrow$ 1 $\leftrightarrow$ 2. In this case, the combined PDF for states 2 and 3 would be a linear combination of the PDF of state 3 and PDF of state 2, since they are independent. The probability of entering state 2 versus state 3 from state 1 depends on the branching ratio, i.e. $a = \frac{k_{12}}{k_{12}+k_{13}}$. Thus, the combined PDF would be: $a\,k_{21}e^{-k_{21}t} + (1-a)k_{31}e^{-k_{31}t}$ and the resulting CDF would be: $1 - a\,e^{-k_{21}t} - (1-a)e^{-k_{31}t}$.

If the states are not independent/parallel states (states 2 and 3 are connected) the CDF is much more complicated. If the states are connected to each other, they can transit between those states or back to state 1. First if you consider the PDF of two sequential transitions, which is the expected lifetime for two stochastic events to occur in series, you need to solve a convolution integral of the PDFs. In this case, for steps a and b (not worrying about state connectivity for now) you would have:

$$PDF = \int_0^t k_a\,e^{-k_a\tau}\,k_b\,e^{-k_b\,(t-\tau)}d\tau$$

$$PDF = \frac{k_b k_a}{k_b - k_a}[e^{-k_a t} - e^{-k_b t}]$$

This PDF looks similar to a Gamma distribution, with a minimum at zero. The CDF corresponding to this PDF is then the integral of this function, and has a sigmoidal shape. In terms of the HMM, the observed combined PDF of states 2 and 3, if all states are connected, would be a combination of transitions from 2$\rightarrow$1 (exponential), transitions from 3$\rightarrow$1 (exponential), and all possible combinations of visits between 2 and 3 before returning to 1, weighted probabilistically (Gamma-like). Depending on the values of the rate constants, this CDF could look almost exponential or more like a Gamma distribution. Thus, although an exponential CDF does not guarantee that the HMM is correctly fitting the underlying states, anything that is non-exponential (combination of exponentials or Gamma-like) is guaranteed to be under-fit. However, we also have finite sampling of lifetimes and so some distributions may look skewed from exponential due to sampling error. Thus, unless distributions across a range of forces or molecules are all underfit, a single underfit distribution is likely not evidence of additional states. This was also confirmed by simulation of a Markovian trajectory, where the transitions are randomly pulled from an exponential distribution, but resulting lifetimes can show significant skew depending on sample size (simulation length).

Tables

**Table 2.1: Fit and extrapolated parameters of $FL_{int}$, C domain, and N domain**

|  | $k_{0,\,folding}$ $(s^{-1})$ * | $p$ value compared to $FL_{int}$ | $\tau_{0,\,unfold}$ (s) * | $\Delta x^{\dagger}_{folding}$ (nm) | $p$ value compared to $FL_{int}$ |
|---|---|---|---|---|---|
| $FL_{int}$ | 7.7 (3.3, 18) x$10^5$ | n/a | 1.3 (0.5, 2.9) x$10^{-6}$ | 7.1 ± 0.6 | n/a |
| C domain | 1.8 (0.6, 5.4) x$10^6$ | n.s. (0.11) | 5.5 (1.8, 16) x$10^{-7}$ | 7.9 ± 0.8 | n.s. (0.06) |

| N domain | 660 (370, 1200) | $2.6 \times 10^{-10}$ | 1.5 (0.86, 2.7) $\times 10^{-3}$ | $5.6 \pm 0.6$ | 0.002 |
|---|---|---|---|---|---|
|  | $k_{0,\,unfolding}$ (s$^{-1}$) * |  | $\tau_{0,\,fold}$ (s) * | $\Delta x^{\dagger}_{unfolding}$ (nm) |  |
| FL$_{int}$ | 5.5 (3.2, 9.3) $\times 10^{-2}$ | n/a | 18 (11, 31) | $4.1 \pm 0.3$ | n/a |
| C domain | 3.1 (1.5, 6.1) $\times 10^{-2}$ | n.s. (0.10) | 33 (16, 65) | $4.6 \pm 0.4$ | 0.04 |
| N domain | 1.2 (0.6, 2.5) | $2.7 \times 10^{-6}$ | 0.8 (0.4, 1.7) | $3.9 \pm 0.8$ | n.s. (0.33) |

n.s.: not significant

* Note that the rates and lifetimes are derived from linear fits to ln(k), and as such the error propagation is quite large when exponentiated and in some cases unphysical. These estimates are using the 95% confidence intervals from the fits as upper and lower bounds for the exponentiated value, expressed in parentheses as (lower, upper).

# Chapter 3.
## Folding of calerythrin on the ribosome

Portions of this section are reproduced with modifications from co-authored material with permission of the other authors.

### 3.1 Introduction

Through folding, polypeptides adopt the specific three-dimensional structure required for their function. In the cell, structures begin to form cotranslationally (*18*, *64–66*), albeit with reduced stability with respect to folding off the ribosome (*12*, *13*, *67*, *68*). Nonetheless, the vectorial N-to-C terminal nature of synthesis by the ribosome is thought to increase the overall yield of active protein by promoting folding in a domain-wise manner and by preventing off-pathway misfolded states (*69*); folding of the N-terminal domain can also increase solubility and stability of the nascent chain (*70*). Given these advantages of N-terminal directed folding, what happens if the protein folds through a C-terminal intermediate, and, more generally, how does real-time synthesis affect the folding pathway? Here we compare the folding of a full-length multidomain calcium-binding protein with: non-ribosome bound truncations; stalled, ribosome-bound, nascent chains (at equilibrium); and with the nascent chain as is it being synthesized in real-time (non-equilibrium), via optical tweezers.

We show that a truncation with a partial C-terminal deletion adopts a misfolded state that is not seen in the full-length protein. This same misfolding is observed on stalled ribosomes, although the rate of misfolding is slower and the rate of unfolding is increased. This observation shows that while the ribosome can disfavor misfolding in stalled complexes that have reached equilibrium, it does not prevent it entirely. Surprisingly, however, during active translation, the attainment of the misfolded state is strongly suppressed, appearing only after a long delay (71 s, i.e. 50 times slower than on stalled ribosomes). This result indicates that as it is being synthesized by the ribosome, the growing polypeptide is not equilibrated with its ensemble of accessible conformations, as has been previously assumed. Non-equilibrium folding dynamics may play an important role in promoting productive folding by delaying it until adequate

sequence is available and/or by allowing time for chaperones to bind. Moreover, these results caution against predicting folding pathways from equilibrium experiments that may mask these effects.

## 3.2 Equilibrium folding experiments

In order to investigate the effects of the ribosome on multidomain folding events and during synthesis, we used an optical tweezers (OT) assay that allows us to apply a denaturant (force) selectively and to directly measure rates and intermediates of defined molecular species. We selected as a model the protein calerythrin, a two-domain calcium-binding protein from the prokaryote *Saccharopolyspora erythraea* (*35*, *36*), which, as previously established in Chapter 2, folds in solution through the C domain. During translation, however, the N domain is synthesized before the C domain. To test whether folding proceeds through the N domain on the ribosome, we prepared ribosome nascent chain complexes (RNCs) stalled at codons 135 to 177 in 10-11 codon increments (Figure 3.1), tethered by the N terminus of the nascent chain and an added tag on protein L17 of the ribosome (*13*). We also prepared a stalled complex with an additional 45 amino acid linker past the natural stop codon, which should extend the full sequence outside the ribosome (RNC177$_{+45}$).
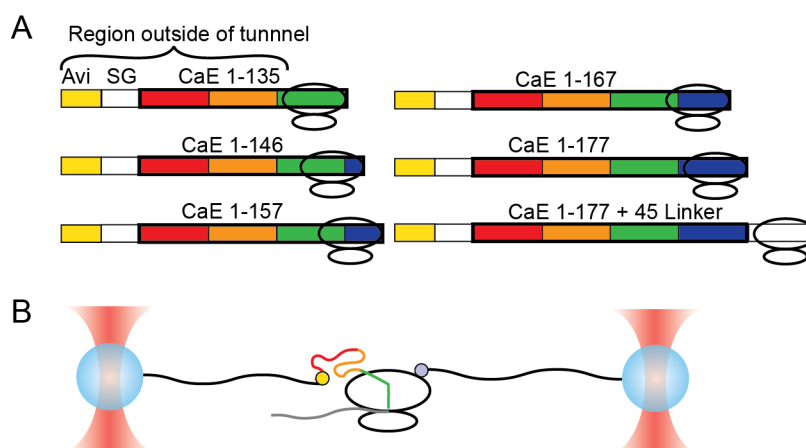


**Figure 3.1: Design for testing folding of stalled, equilibrated RNCs.** A) The mRNAs for stalled RNCs. Each mRNA is truncated at a different codon position, named by final codon added. B) The OT geometry for RNC experiments, tethered via L17 of the 50S and the N terminus of the nascent chain

At codon 135, the N domain (1-90, corresponding to EF1 and EF2) is completely outside of the exit tunnel with a 45 amino acid linker between the domain boundary and the peptidyl transfer center (*8*). Since we observed N domain folding in isolation (Figure 2.9) and other experiments have demonstrated a minimum linker length for folding of 32-39 amino acids (in cells/OT) (*71*) or 42-45 amino acids (via NMR) (*12*), we expected to see the N domain hopping, perhaps with reduced kinetics. However, we do not observe folding for this complex (RNC 135), nor for RNC146 and RNC157 (Figure 3.2).
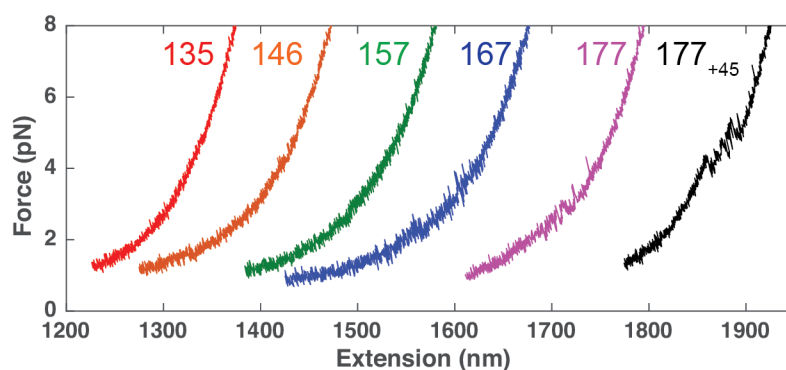
**Figure 3.2: Force-extension curves of a library of stalled RNC constructs.** Plots of refolding curves of all tested stalled RNCs. Constructs do not show folding until RNC167 and RNC177. RNC177$_{+45}$ reverts to C-terminal folding. Representative force-extension curves from n = 3, 10, 9, 5, 10, 5 molecules respectively.

Force extension curves as shown in Figure 3.2 are the best way to demonstrate a lack of folding, since it covers the full range of forces. The nonequilibrium bias of force-extension curve measurements also tends to increase the force at which unfolding occurs, making it easier to detect. However, if the rate of pulling is fast, we could conceivably miss refolding transitions. To emphasize that this is not the case and the early complexes do not fold, we can also probe a range of forces with passive data collection. Again, even while waiting at low forces for long times, we do not observe folding of RNC135, RNC146, or RNC157 (Figure 3.3).
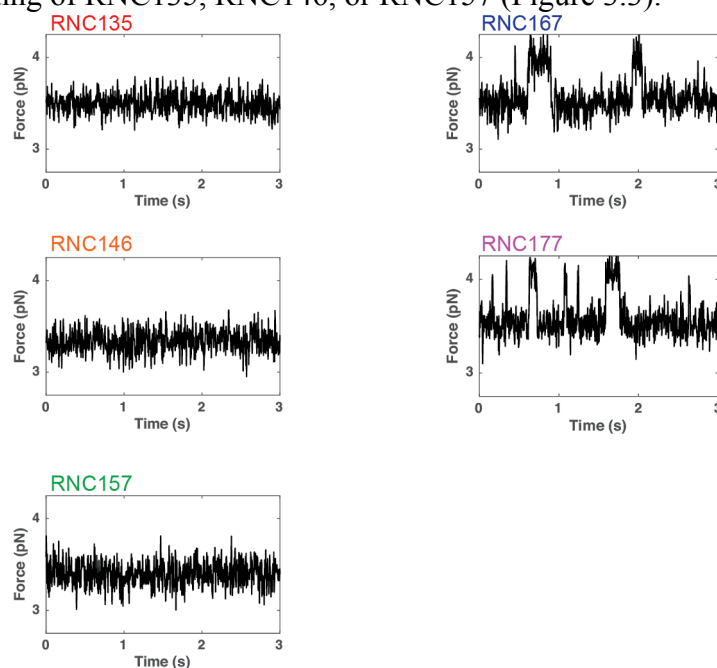


**Figure 3.3: Passive regions for the stalled RNC library.** In this case, we see no folding near 3.5 pN for RNCs 135, 146, and 157. At 3.5 pN, the N-domain transition would be to 3.8 pN and the misfolded state would be at 4.0 pN. Compare to Figure 2.9, where the N domain is observed. RNC167 and RNC177 fold to the misfolded state. Representative from n = 3, 10, 9, 5, 10 molecules respectively.

The isolated N domain off the ribosome folds at a rate of $k_0 = 660 \pm 370$ s$^{-1}$, so we infer that the ribosome destabilizes the N-domain by slowing its folding rate or increasing its unfolding rate relative to the isolated N-domain such that it is not sampled. Surprisingly, when we tested RNC167 and RNC177 (FL but partially in the tunnel), we observe a misfolded state of $130 \pm 15$ amino acids, significantly larger than either the N or C domain.

The misfolded state observed with RNC167 and RNC177 is rarely observed for the fully synthesized and released protein; however, the same state is accessible to an isolated truncated version of the protein lacking the 4th EF hand (EF123). Unlike RNC177 which only has one (mis)folded state, EF123 can fold to either this misfolded state or the N domain (see Chapter 2.4). While the full-length protein in solution usually avoids this state due to the faster folding of the C domain, this state is accessible on the surface of the ribosome when the sequence needed to form the C domain is still partly sequestered in the ribosomal exit tunnel. Significantly, although the misfolded state is sampled both on the ribosome and for the isolated EF123, the ribosome decelerates its folding kinetics by $10^4$ –fold and accelerates its unfolding rate by 90–fold (Figure 3.4, Table 3.1 at end of chapter).
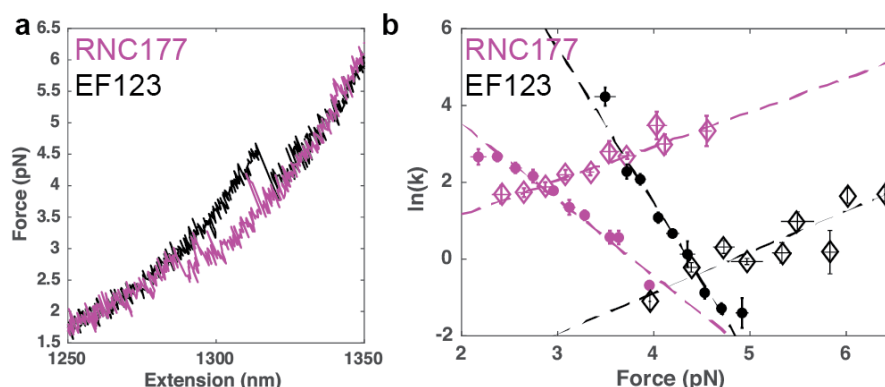


**Figure 3.4: Stalled RNC177 show altered kinetics compared to EF123.** A) Refolding curves of RNC177 (purple) and EF123 (black) show the changed behavior on the ribosome. Both constructs fold to the misfolded state, and so overlay, but RNC177 is less stable and folds at lower forces. B) The full force-dependent kinetics of RNC177 and EF123 (unfolded to misfolded state only) are shown. The values from the fits are listed in Table 3.1. Error bars are SE. RNC177: n = 10 molecules, 111 rate measurements from 9958 transitions. EF123 (unfolding): n = 14 molecules, 97 rate measurements from 2081 transitions. EF123 (folding): n = 13 molecules, 88 rate measurements from 1662 transitions. One molecule was excluded from folding kinetics due to increased baseline noise that made detection of state 2 and therefore distorted the apparent time to fold.

Both of these trends contribute to a higher probability for the RNC to avoid the off-pathway misfolded state and are beneficial for recovering productive folding. Notice, however, that the time until misfolding on the ribosome at zero force, while longer than for EF123, is still quite short ($6 \pm 3 \times 10^{-4}$ s) and the time it remains trapped in this state ($1.9 \pm 0.5$ s) is still relatively long. Therefore, it is not clear whether the effect of the ribosome, as revealed by the stalled

RNCs, is sufficient to prevent the misfolded state throughout all of polypeptide synthesis, which is on the time scale of seconds.

Once we extend the entire protein sequence outside of the exit tunnel with a 45 amino acid artificial linker (RNC177$_{+45}$), native-like folding is restored (Figure 3.5). This construct has an intermediate of $70 \pm 6$ amino acids, similar to FL$_{int}$ of the released protein ($76 \pm 6$). There are two noticeable effects, however. First, once the linker is out of the tunnel, the misfolded state is no longer observed and all folding reverts to the native-like pathway. This supports the interpretation that the C domain outcompetes the folding of the misfolded state and the two do not coexist. Second, the native-like folding is destabilized by proximity to the ribosome–the folding of the C domain (FL$_{int}$) occurs at lower forces/slower kinetics (Figure 3.6, Table 3.2 at end of chapter).
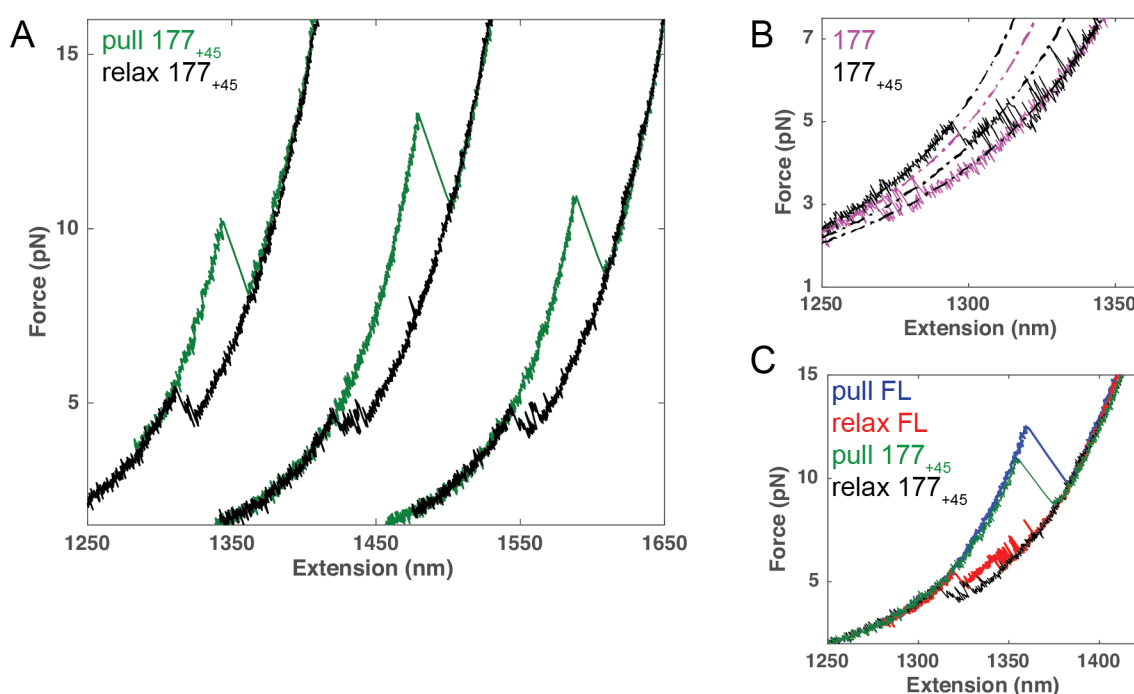


**Figure 3.5: RNC177$_{+45}$ reverts to C-terminal folding.** A) Force extension curves for RNC177$_{+45}$ show FL-like folding every time. B) A zoomed-in overlay illustrates the misfolding for RNC177 versus FL-like for RNC177$_{+45}$. Worm-like chains with 177, 134, 75, and 0 folded amino acids are shown (black dash-dot lines; 134 is magenta). C) A zoomed-in overlay illustrates the free protein (FL) folding versus RNC177$_{+45}$. Note that refolding (black curve) begins at much lower forces than in FL protein (red curve).
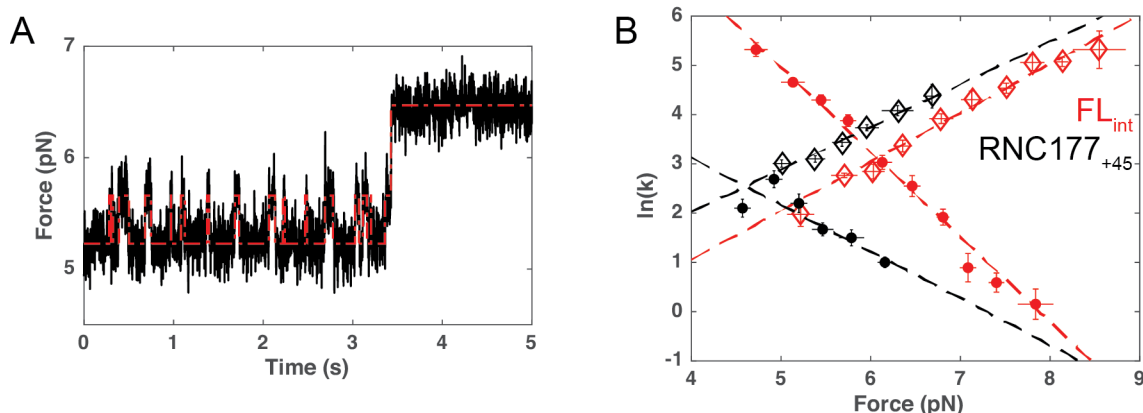
**Figure 3.6: RNC177$_{+45}$ has altered kinetics of folding on the ribosome.** A) Passive data of RNC177$_{+45}$ shows folding to one intermediate but at lower forces that FL$_{int}$ (see Chapter 2; FL$_{int}$ usually fold at 6-7 pN). B) The Bell plot of FL$_{int}$ and RNC$_{+45}$ shows that folding is decreased on the ribosome (p<0.05). Unfolding is unaffected. (for RNC177$_{+45}$ n = 5 molecules, 57 rate measurements from 3651 transitions).

This kinetic effect on refolding for RNC177$_{+45}$ is similar to the effect demonstrated on RNC177 and is also similar to what has been reported for T4L's refolding (*13*). Unlike RNC177, the unfolding of the intermediate in this case does not seem to be affected.

One unexpected effect is that in both RNC177 and RNC177$_{+45}$, the distance to the transition state for the folding intermediate is altered compared to the equivalent protein in solution, although the state that it folds to seems to be unchanged. Comparing RNC177 to EF123, the change in the distance to the transition slope (the slope of the linear fit in Bell's model) decreases by 8.1 nm (from 16.6 nm in EF123 to 8.1 nm, *p* = 3.8x10$^{-5}$). Comparing RNC177$_{+45}$ to the FL protein, this distance changes by 2.7 nm (from 7.1nm in FL$_{int}$ to 4.4 nm). Although it is interesting that they both show the same directional effect, note that in this case, unlike RNC177, the change may not be significant as the folding rate of RNC$_{+45}$ has large spread; *p* = 0.068 by Welch's t-test, not <0.05. The large spread likely results from the lower number of molecules in this condition, which was not tested very much since we were more interested in translation intermediates rather than artificial constructs.

Since the distance to the transition state for folding changes but the distance to the transition state for unfolding is not significantly changed in either case, we think it is unlikely that this represents a different pathway of folding on the ribosome. Rather, this could reflect a change in the energy minimum associated with the unfolded state, thus giving a different distance from the unfolded potential energy well to the transition state for folding. This shorter distance to the transition state would conform to previous studies of nascent chains which show a compaction of the unfolded state on the surface of the ribosome (*72, 73*).

When we look at the unfolding of the fully folded two-domain structure, we see that its overall stability is also minorly affected (Figure 3.7). In this case the overall lifetime at zero

force (extrapolating to the y-intercept) changes from 146 seconds in FL to 16 seconds for RNC177$_{+45}$.
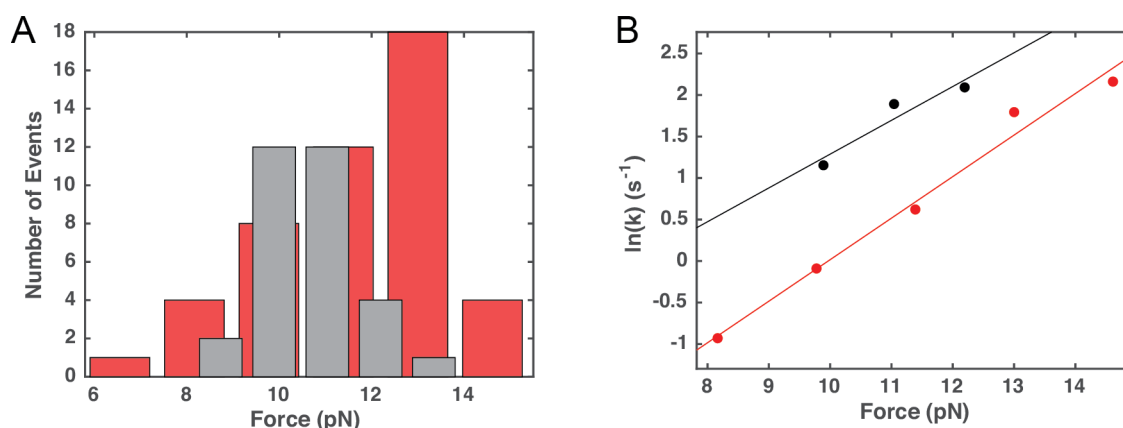


**Figure 3.7: RNC$_{+45}$ has altered kinetics of unfolding of the native structure** A) RNC177$_{+45}$ unfolds at slightly lower forces than the FL protein (11.5 pN to 10.6 pN). B) A Bell plot of lifetimes using the Dudko transformation from force histogram of A (*30*). Bins with less than 2 events were excluded due to high expected error from under-sampling issues.

Overall, our stalled RNC equilibrium experiments provide detailed the kinetics of folding at "snapshots" of positions in translation. According to these results, the nascent chain misfolds during the course of translation, but reverts to native folding once extended from the surface further or presumably once translation terminates. As we do not observe coexistence of the misfolded state and the native intermediate, it is not clear how this switch occurs or on what timescale. Based on our kinetic analysis, we expect that the misfolded state will dominate the folding outcome prior to termination despite the destabilization. These snapshots, however, do not include active elongation, which may introduce non-equilibrium effects. Since the protein is getting longer during active elongation, the polypeptide may not have time to fully explore/equilibrate to the potential energy landscape prior to further amino acid addition, and thus not all intermediates may be relevant. It is also possible that the dynamics of the active ribosome, which undergoes many structural rearrangements during each elongation cycle, can induce changes in the folding of the nascent chain that are not present for stalled, inactive ribosomes. Therefore, it is valuable to also explore the folding behavior of a RNC during real-time translation.

## 3.3 Non-equilibrium cotranslational folding experiments

To determine if this misfolded state is relevant during progressive peptide elongation, we designed a method of creating stalled ribosomes which allows for later restart in the OT chamber. Once translation restarts, we can follow in real-time the growth of calerythrin while simultaneously monitoring its co-translational folding. As previous stalled complexes were generated by truncation of the mRNA and thus cannot restart, we need to modify our design to instead stall RNCs by amino acid omission. The mRNA was engineered to contain an Avi tag, a longer serine-glycine linker to extend the Avi tag out of the tunnel, and then a stall site of 3

valines prior to the natural sequence of calerythrin. The triple valine site allows us to stall translation by leaving out valine from our stalling reaction.

There are 225 codons after the stall site, corresponding to three valine codons for stalling translation, all 177 amino acids of native calerythrin, and the same C-terminal linker used for generating the stalled RNC177$_{+45}$. We restarted translation by introducing a homemade *in vitro* translation mix (*74*) (see Table 3.4 at the end of the chapter). Since these experiments are conducted in passive mode, where the positions of the optical traps are fixed, active elongation is observed as a drop in force concomitant with the polypeptide length increase; folding, as in previous experiments corresponds to a rise in force (Figure 3.8).
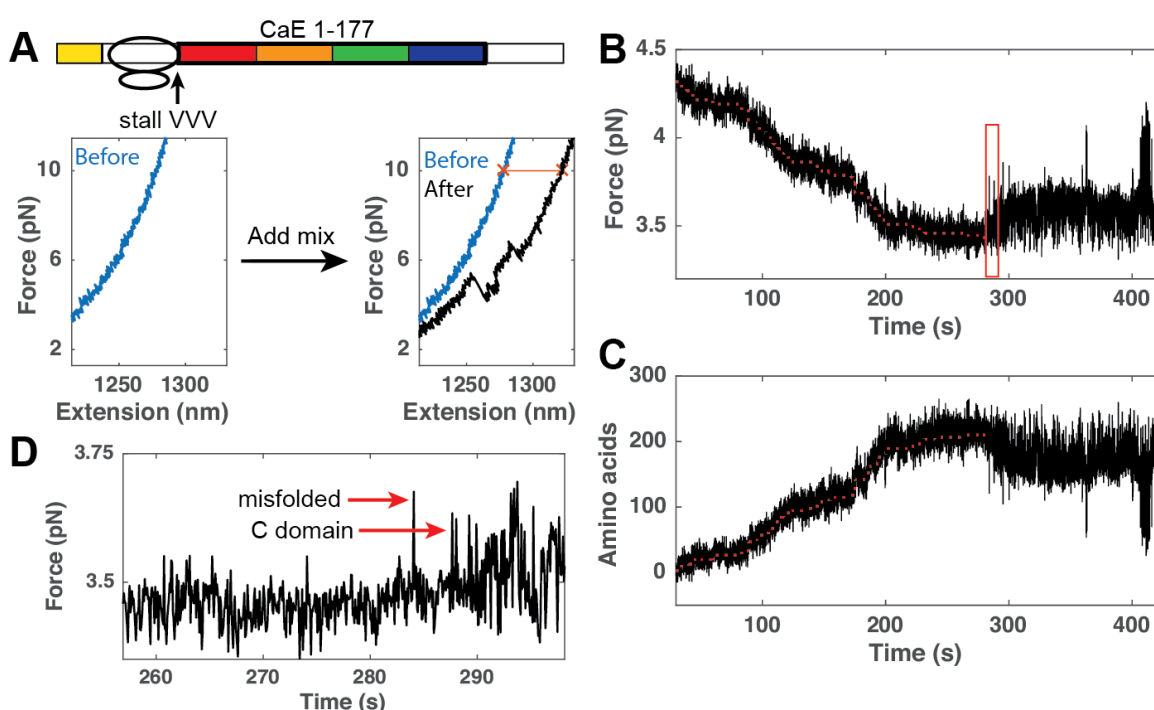


**Figure 3.8: Real-time translation by the ribosome.** A) The sequence for stalling and restart. RNCs were stalled at a 3-valine repeat and restarted with the addition of translation mix in the OT. After translation, the force extension curves show the distance change of the tether (red line) as well as folding of the full-length protein, which confirms it reaches the end of the mRNA. B) An example ribosome translating. As each amino acid is added, the force decreases as the tether gets longer. The fit is shown as a red dotted line. C) The trajectory is converted to amino acids using a worm-like-chain model. D) A zoom of the panel in (B) that shows the folding region. The red arrows show the misfolded state and folded C domain. Representative molecule from n = 12.

We converted force to amino acid position using a worm-like chain (WLC) model for the polypeptide and an extensible WLC for the DNA handles. By fitting the data with a general isotonic regression, we find an average elongation rate of $1.0 \pm 0.3$ amino acid/second, consistent with other *in vitro* translation experiments using the commercial PURExpress system (*75, 76*).

This conversion also allows us to map to the underlying sequence position and assign the ribosome to a codon on the mRNA (Figure 3.8C).

As expected from our stalled equilibrium experiments, at forces below 5 pN we observe folding transitions. Depending upon the sequence position in translation, these transitions are to the folded C domain or the misfolded state (Figure 3.8D, red arrows mark the folded C domain or misfolded state). Every molecule exhibits misfolding or C domain folding, but the N domain folding is not observed, as predicted from stalled RNC data. Thus, in a qualitative sense, the data from real-time translation matches our stalled data.

Most surprisingly, however, although the misfolded state is detected during real-time translation, it does not form readily when the ribosome has reached codon 167, as would be expected from our experiments with stalled complexes. Instead, we observe a delay until the onset of folding, after which the chain starts hopping and shows reversible folding kinetics. This delay can be quite long in some cases, such that folding does not begin until the ribosome has translated well into the linker and the C domain is accessible (Figure 3.9).
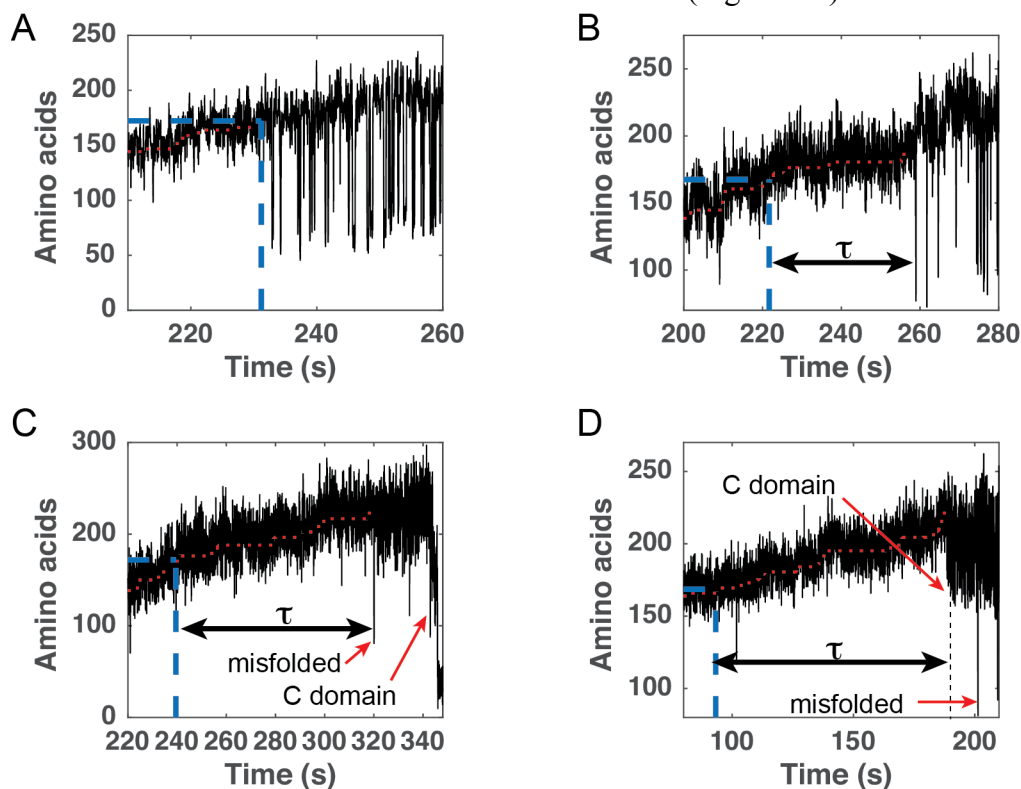


**Figure 3.9: Real-time elongation shows a delay before folding begins.** Panels showing real-time elongation data of amino acids added versus time. In this case, folding is a downward transition. In all panels, codon 167 is marked with a blue dashed line; if the nascent chain is equilibrated folding should begin with <2 seconds of this point. A) In this molecule, folding is essentially indistinguishable from stalled RNC data; folding to the misfolded state begins shortly after reaching codon 167. B) In this molecule, it shows a slight delay after 167 but still begins to fold without translating very far into the added linker (folding begins at codon 186). C) Folding is delayed until substantial portions of the linker are translated (first transition at codon 221); the

misfolded and C domain state can be seen to coexist here. Folding of the native structure is seen near the end at 348 seconds. D) Folding is delayed until the entire linker is translated. The first folding transition is to the C domain.

We measured the initial delay $\tau$ as the time elapsed between the moment when the ribosome reaches codon 167 to the time when we first observed a folding transition, either to the misfolded state (75 % of the molecules) or to the folded C domain (25 %). In 8% of molecules, the delay is essentially indistinguishable from $\tau_{unfold}$ in stall complexes, demonstrating that this outcome is observable in our assay, though less common (Figure 3.9A). The exponential distribution of $\tau$ supports the idea of a finite equilibration step that is stochastic (Figure 3.10). The characteristic delay until the onset of folding, $\tau_{delay} = 71 \pm 14$ s, is 50 times longer than the measured lifetime of the unfolded state at equilibrium in RNC167 and RNC177 ($\tau_{unfold} = 1.5 \pm 0.3$ s at 3.98 pN, the average force at which the elongating ribosomes reach codon 167).
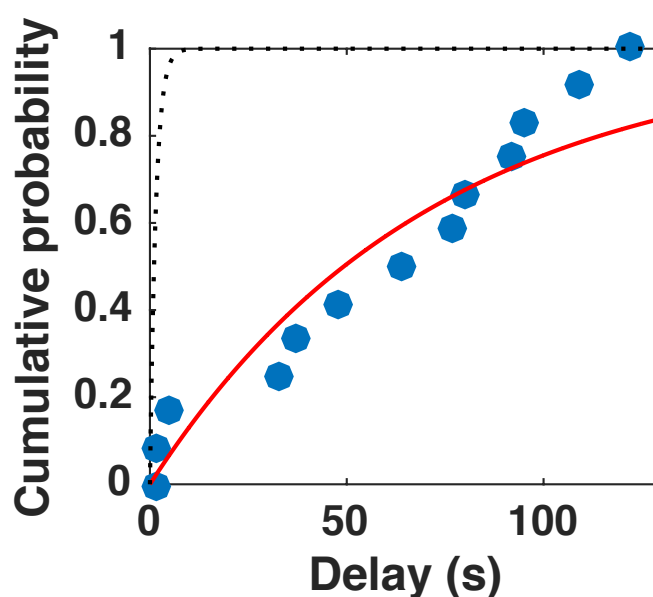


**Figure 3.10: $\tau_{delay}$ is a single-step kinetic process** The experimental data ($\tau$) are blue points sorted into an empirical cumulative distribution function and the red fit is to the relationship $1 - e^{-\tau/\tau_{delay}}$. The black dotted line shows the expected distribution for the time until folding at 167 assuming equilibrated kinetics (fromRNC167 and RNC177).

This behavior indicates that prior to the onset of folding, the newly-synthesized polypeptide resides in a "trapped" unfolded state that is not folding-competent. This state may involve a set of local interactions with the ribosome surface or exit tunnel that were established during its growth in real time or may be due to dynamics of these surfaces. Once the chain manages to escape, however, it can then (mis)fold and unfold without visiting again the trapped state and is equilibrated within its energy landscape. In support of this interpretation, we observe no subsequent delay after ramping the force up to extend the chain and returning it to forces as low as 2 pN; instead, immediate hopping between unfolded and (mis)folded states resumes, as in stalled complexes. Moreover, this delay is not due to interactions between components of the

translation mix and the nascent chain, as experiments with stalled RNCs in translation mix did not show a delay (Figure 3.11).
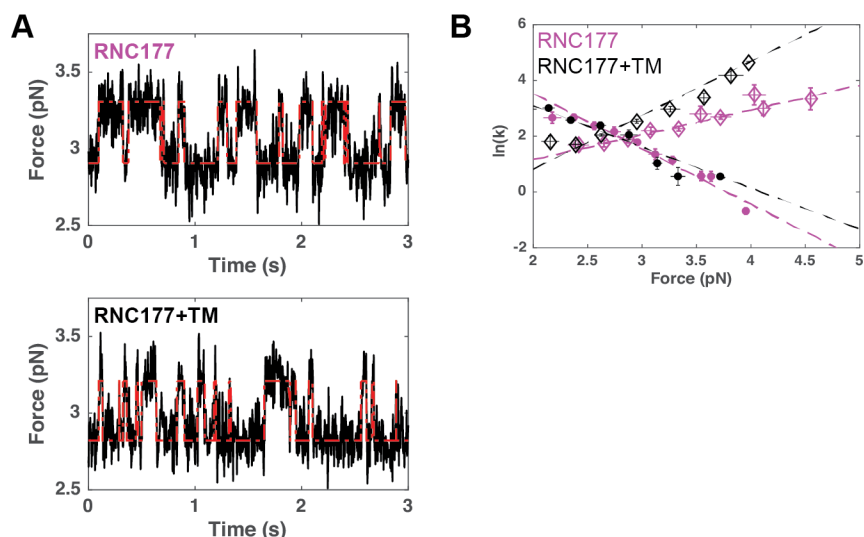


**Figure 3.11: Measurements on RNC177 in the presence of translation mix.**
A) Examples of passive show the same qualitative folding to a misfolded state. There is no delay before hopping is observed. B) The force-dependent kinetics show small changes in the presence of TM but not non-equilibrium behavior. Error bars are SE. RNC177: n = 10 molecules, 111 rate measurements from 9958 transitions; RNC177 + TM: n = 5 molecules, 68 rate measurements from 8889 transitions.

## 3.4 Discussion

Since translation is slow relative to the timescale of folding, traditionally the nascent chain has been thought to be in "quasi-equilibrium" (*72, 77, 78*) wherein the polypeptide reaches its thermodynamic minimum after each amino acid addition. Under such conditions, the kinetics of folding in real time elongation experiments are expected to match those of stalled complexes. However, despite the relatively slow rate of translation in our experiments which should even more heavily favor quasi-equilibrium, we find that this assumption is incorrect and the nascent polypeptide equilibrates only after a significant time delay, preventing the formation of an off-pathway misfolded state. Given a known elongation rate of 12–18 s$^{-1}$ *in vivo* (*4*) and a measured $k_{cat}$ of 1.5 s$^{-1}$ for peptide release (*79*), the measured delay of 71 s is long enough to avert misfolding for the remaining duration of elongation until the release of the nascent polypeptide. It could also allow time for chaperones to bind to the unfolded polypeptide and prevent the formation of the misfolded state.

Kinetic modeling has suggested that not all proteins are equilibrated on a translational timescale (*80*), however experimental testing thus far has been indirect, for example through changes in protein activity after changing the translation rate via synonymous codon mutations (*18, 64, 66*). Other *in vitro* experiments which incorporate active translation employ ensemble assays where ribosomes are not synchronized (*11*), which may make deviations from equilibrium

behavior difficult to detect. Using a real-time single molecule assay we detect a clear signature of non-equilibrium dynamics that may play an important role in the fate of the newly synthesized polypeptide. This result indicates that comparing the equilibrium-derived folding rates from stalled RNC behavior with the translation rate is not sufficient to predict co-translational folding, and instead the peptide product needs to be observed in real-time as it is being generated.

Most of the common techniques to study cotranslational folding are not time resolved. For example, one common technique (which was developed in our laboratory) uses stalling peptides like secM to stall translation. If the protein folds while near the surface of the ribosome, there is a steric occlusion force that is transduced to the peptidyl transfer center and overcomes the secM-induced stalling (*71*). This assay can be used *in vitro* and *in vivo* to test for folding by having the nascent chain of interest upstream of secM and a reporter gene (e.g. GFP) downstream of secM (*10, 71, 81*). This technique is very innovative and versatile, however it is not time-resolved. The assay will tell you of the protein *can* fold outside of the ribosome not if it *will* fold in the context of elongation. Additionally, since the stall release is an irreversible process (absorbing state), the equilibrium state will be released, even if folding is very rare. There is no data provided on how long it was stalled at secM prior to overcoming the stall site– it could be stalled for minutes prior to folding, after which stalling is released. However, during normal elongation without secM the ribosome may have reached the stop codon in less time than this, thus it would not, functionally speaking, fold cotranslationally. This is true for many techniques that probe folding state without folding kinetics, i.e. that only look at the thermodynamic equilibrium ($K_{eq}$) of folding (*12, 67, 73, 82*), regardless of possible non-equilibrium effects.

The non-equilibrium dynamics that may take place exaggerate this effect and make time resolved observation and knowledge of the kinetic information crucial. The sample preparation time involved in producing RNC stalled complexes would in almost all cases obscure the non-equilibrium delay observed here (71 seconds), regardless of methodology used to subsequently probe the nascent chain state. The real-time folding assay developed in this study will also be useful in other situations where measurements of stalled RNCs may not capture the full dynamics of folding, such as length-dependent protein conformational changes, dynamic interactions with chaperones, or folding near domain boundaries.

## 3.5 Methods

Cloning, expression, and purification of proteins

The tRNA synthetases and nucleoside diphosphate kinase were purified as described (*83, 84*). Pyrophosphatase, myokinase, and creatine kinase were purchased from Sigma. EF-G, EF-Tu, and EF-Ts were expressed with autoinducing media. EF-G was purified via a His column gradient from 20 mM to 300 mM imidazole in 1X PBS with 3mM β-mercaptoethanol over 12 column volumes (CV). The EF-G fractions were pooled, diluted three-fold, and then further purified on a Q column via gradient from 50 mM to 600 mM NaCl in 25 mM Tris pH 8 with 3 mM β-mercaptoethanol over 20 CV. EF-G was dialyzed into 25 mM Tris, 60 mM NH4Cl, 6 mM β-mercaptoethanol overnight, concentrated and stored at -80. EF-Ts was purified in a gradient

from 10 mM to 1 M imidazole in 20 mM Tris, 300 mM NaCl, and 2 mM β -mercaptoethanol over 12 CV. Fractions were pooled and dialyzed overnight into 50 mM HEPES, 100 mM KCl and 10 mM β -mercaptoethanol. The purified protein was concentrated and supplemented with buffered glycerol to 30% final concentration (buffer composition unchanged), then stored at -80. EF-Tu was purified in a gradient from 30 mM to 800 mM imidazole in 20 mM Tris, 300 mM NaCl, 0.5 mM MgCl2, 50 µm GDP, and 2 mM β -mercaptoethanol over 12 CV. Fractions were pooled and dialyzed overnight into 50 mM HEPES, 100 mM KCl, 0.5 mM MgCl2, 50 µM GDP, and 10 mM β -mercaptoethanol. The purified protein was concentrated and supplemented with buffered glycerol to 30% final concentration (buffer composition unchanged), then stored at -80. All proteins were checked for purity via SDS-PAGE.

Optical tweezers samples: stalled RNCs

All mRNA was designed without a stop codon, which will cause the ribosome to stall once the last codon is translated. Because there is no further sequence in the A-site, the ribosome cannot by-pass the stall, and normal termination factors require a stop codon in the A site. The rescue system for truncated mRNAs is not included in our in vitro reaction (PURExpress ΔRibosome kit). The mRNA generated via run-off in vitro transcription to ensure a blunt end (terminator transcription can yield heterogeneous mRNA products) (MEGAscript kit). The transcription template was generated via PCR and gel purified from an agarose TAE gel and ethanol precipitated prior to use. The resulting RNA product was also gel purified from 7M Urea/4% polyacrylamide gels before use, phenol-chloroform extracted, and then ethanol precipitated before resuspension in 1 mM EDTA pH 8.

The ribosomes were CoA-modified prior to the stalling reaction as previously described (*13*). Stalled RNC samples were generated by as described (*13*), except the buffer for the sucrose cushion and resuspension was 1X polymix with 1U/uL RNAse Out. Stalled RNCs were analyzed for a homogeneous biotinylation product via Western blot with a streptavidin-HRP probe before use (Figure 3.12).

After resuspension from the sucrose cushion pellet, the RNC was aliquoted and stored at -80. Once ready to collect data on the tweezers, the ligation of the RNC to DNA handles was conducted in 1X polymix buffer supplemented with 1 mM ATP, 1mM $MgCl_2$, 1 mM DTT, and 11 U/uL T4 ligase, then deposited on beads the same as for protein samples (see Chapter 2.5) and kept on ice.
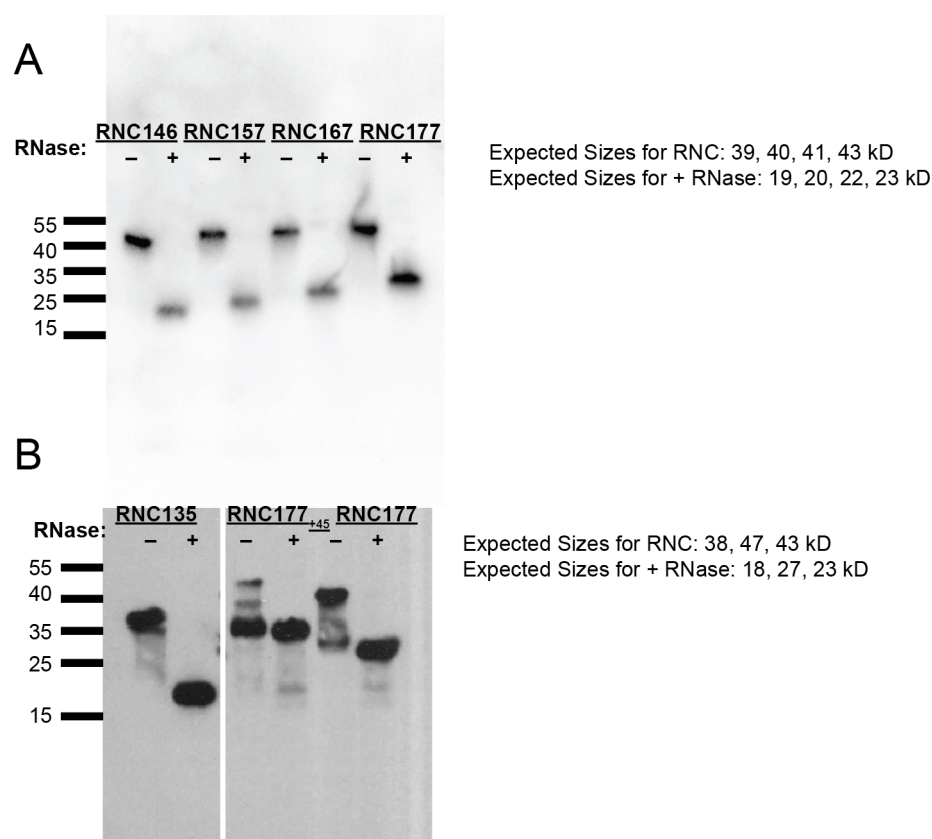
**Figure 3.12: Western blot verifies that stalled RNCs are homogeneous.** A) A Western blot of RNC146, 157, 167, and 177 shows very clean stalled reactions. B) a Western blot for RNC135, RNC177+45, and RNC177 (same reaction as in (A) ) shows clean stalling, although RNC177+45 has minor premature stalling. Both RNC177+45 and RNC177 show some hydrolysis of the stalled product, but this is likely due to gel preparation and running rather than sample quality, since the same sample in blot (A) is uniform (A is a commercial gel and B is a homemade bis-tris gel, which may also influence sample running). Even if some hydrolysis product is present, this would not be tethered in the OT since it is released from the ribosome and thus protein L17, our attachment point.

Optimization of omission-stalled complexes

Samples for real-time elongation were stalled via amino acid omission. Stalled ribosomes used for restart were prepared slightly differently than truncated mRNA RNCs due to the transience of this stalled state. Our experiments made via a $^{35}$S-Met incorporation time course indicate the stalled product is only stable for about 10 minutes before read-through if stalling at a single codon (Figure 3.13A). Presumably, after this time, the ribosome misincorporates a competing amino acid and continues translating until the end of the sequence. Thus, we added a triple codon repeat at the beginning of the natural ORF which prolongs this stalled state to >30 minutes (Figure 3.6B); we also added additional amino acids to the serine-glycine linker in between the Avi tag and the beginning of the natural sequence in order to fully expose the Avi tag outside of the ribosome exit tunnel.

A Stalling at natural cysteine

t = 0' 1' 5' 10' 20' 30' 40'

< Full length

< Stalled

< Met-tRNA

B Stalling at inserted VVV site

+Stalling mix | +all amino acids

t = 0' 5' 10' 20' 30' | 0' 5' 10' 20' 30'

< Full length

< Stalled

< Met-tRNA
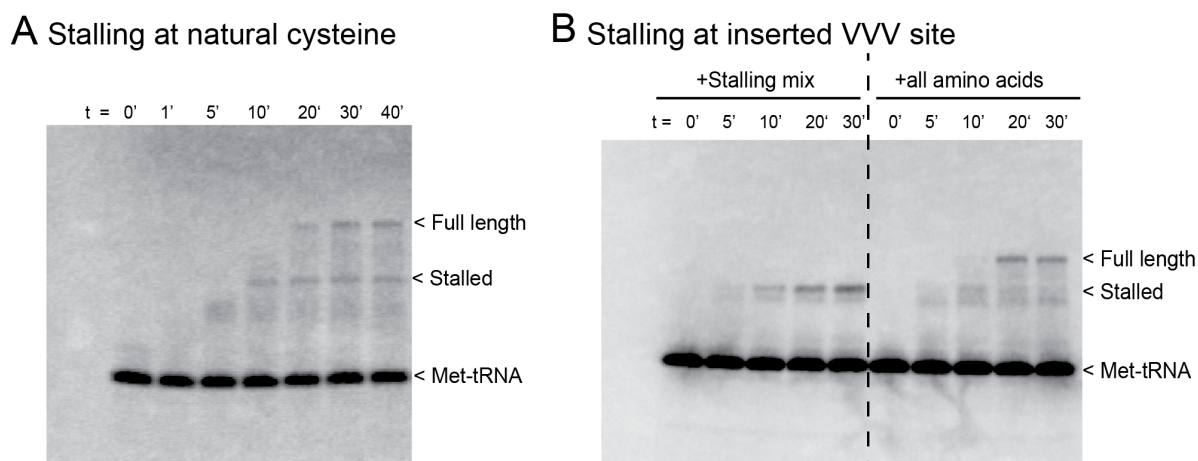
**Figure 3.13: Time course of translation using $^{35}$S-Met.** To design our stall site, we first tested the stability of stalling at a single codon, taking advantage of the single natural cysteine within calerythrin's sequence. A) The stall RNC accumulates after 10', however, by 20' a mixture of stall and read-through is present. Based on this, we decided to design a stall site of 3 sequential codons to ensure long-lived stalling. B) The complete sequence we use in our experiments, which contains 3 sequential valines, does not show read-through up to 30 minutes. Running an experiment with all 20 amino acids on the same gel shows that the read-through band is distinguishable. The faint band underneath the stalled band and in the all amino acid condition may be due to slower translation through the serine-glycine linker.

This new stalling mRNA was tested to ensure the longer SG linker and addition of the triple valines did not affect folding once the protein was synthesized. We generated stalled RNCs with the new mRNA truncated at calerythrin's codon 177 (SGV RNC177) and compared its folding with RNC177's folding (Figure 3.14)
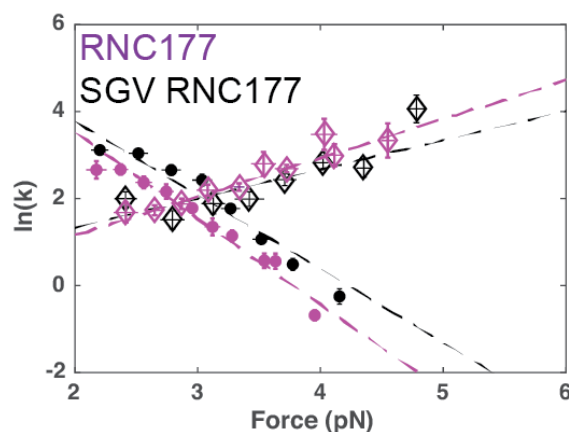


**Figure 3.14: The linker used for real-time stalling does not affect folding.** The force-dependent kinetics do not change with the addition of more linker residues at the N terminus. RNC177: n = 10 molecules, 111 rate measurements from 9958 transitions; SGV RNC177: n = 7 molecules, 79 rate measurements from 6874 transitions. $p$ values for $k_{0, folding}$ and $k_{0, unfolding}$ show changes from RNC177 are not significant (>0.05; 0.35 and 0.24 respectively).

Although the stalled RNC band accumulates significantly in the time course after 10 minutes, we could not obtain tethers in the OT when stalling for this time period. Prerequisite to tether formation, the Avi tag must come completely out of the exit tunnel of the ribosome, be recognized by the biotinylation enzyme birA, and subsequently biotinylated. To conserve material (both the commercial PURExpress mixes and the labeled ribosomes) we are working at a low concentration of ribosomes (0.2 µM). BirA is already relatively slow as an enzyme under saturating substrate conditions (biotinylation rate of 0.17 s$^{-1}$ for 20 µM substrate and 0.9 µM enzyme (46)) and so we thought that inefficient biotinylation may be the root of the issue. It is also possible that substrate recognition is hindered by the proximity of the large, negative ribosome surface. To test biotinylation efficiency, we performed a translation time course with birA and quenched the reaction and different times, then analyzed the products on a Western blot (Figure 3.15).
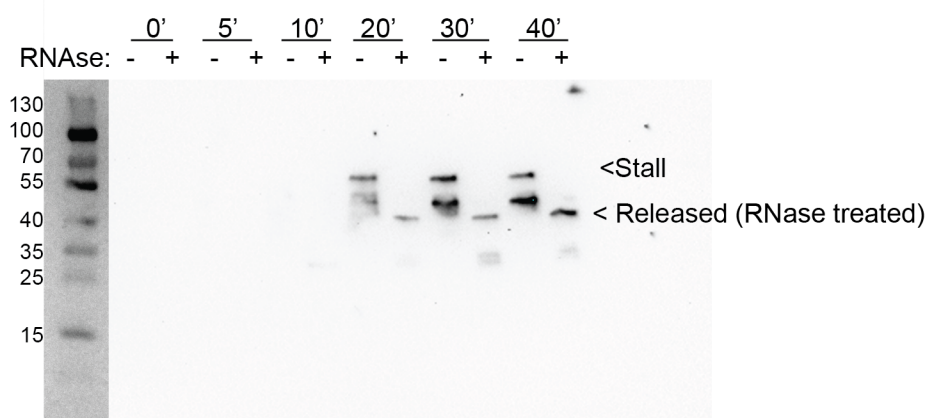


**Figure 3.15: Biotinylation time course.** At each time point, a sample of a translation reaction was taken and split in half. One half was quenched and one half was RNase treated and then quenched. Because the expected stall product for the SGV RNA is very short and does not show up efficiently on a Western blot, we used a different mRNA with a double-cysteine codon repeat stalling site in the middle of calerythrin (stable up to 30-40') that will produce a 56 kD stalled product (this is not the native sequence– it has an SG linker and added tags that increase the mass. See Figure 3.16 for a discussion of this protein design). Biotinylation is not detectable until at least 20 minutes has elapsed and continues to improve after that time. The lower band may be a similar premature stall/slow down through the SG linker as observed for the SGV mRNA.

This method will give a result which is a convolution of the rate of synthesis to expose the Avi tag and the rate of biotinylation and so does not directly report on biotinylation rate. However, this combined time period/efficiency is more relevant to our desired experiment regardless. Thus, incubation for a minimum of 20-30 minutes is required. Although it is likely not fully biotinylated at this point, we did not want to start to get read-through product and so we limited our reaction time.

After the time for sample prep was optimized, we worked on optimizing downstream steps of isolation of stalled complexes. Although sucrose cushions and resuspension of the resulting ribosome pellet works well for truncated mRNA-based RNC species, we had difficulty obtaining tethers using this method. Furthermore, the tethers we did obtain were quite heterogeneous. As we did not know the source of the heterogeneity, for testing in this case we used a tandem-

protein design of Avi-tag–T4L–CaE M110C. This construct has two cysteines in the middle of calerythrin (M110C mutant) that prolongs stalling to at least 30 minutes (established via $^{35}$S). When stalled at the cysteines, T4L is outside of the tunnel and provides a distinctive unfolding signature that can be used to verify correct tethers. This construct was only for testing, however since it is not well designed for real time translation. The dynamics of T4L can interfere with clear assignment of the calerythrin folding, and also it only translates 67 amino acids of natural sequence since it is stalled at codon 110, giving a small overall expected translation signal. As seen in Figure 3.16, preparations of this mRNA using the standard sucrose cushion protocol were mixed in length, curvature, and the presence of the signature:
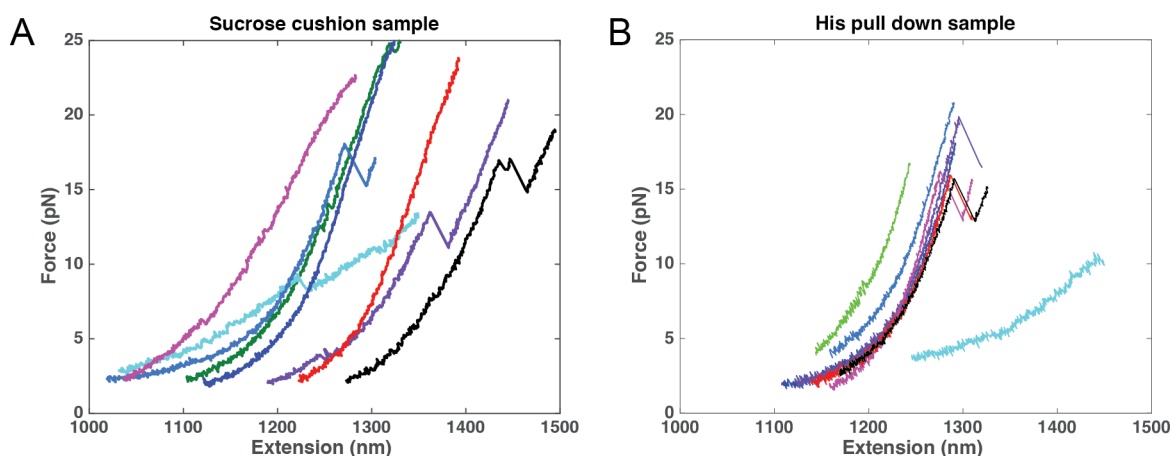


**Figure 3.16: Stalling by omission leads to greater heterogeneity in the observed tethers.** A) RNCs made using the standard protocol of stalling, pelleting through a sucrose cushion, and resuspending. The observed tethers are not uniform in length. The blue tether in the middle is an expected tether with correct curvature, length, and signature of T4L. Some other tethers also seem to have T4L but have odd length/curvature. Some meet none of the expected criteria. B) RNCs made with an alternative, gentler isolation procedure. Using this method, about half have the expected signature and the curvature/length are much more uniform.

It is interesting to note that this heterogeneity is not observed for sucrose cushion prepared RNCs stalled by truncation. In this case, the lengths and curvatures are very uniform (although even these stalled RNCs are not 100% uniform in all aspects– a minority of tethers even of correct curvature show no signature for unknown reasons, possibly early stall/misincorporation). Since there is mRNA in the A site and downstream of the ribosome, and we also require the ribosomes to resume translation after tethering, it makes sense that these complexes may be more sensitive to harsh isolation methods. It is not clear why the curvature and lengths are so variable for omission-stalled RNCs with sucrose versus either truncated RNCs with sucrose or omission-stalled RNCs without sucrose. The DNA handles are from the same batch and ligation is performed in the same way for all 3 cases. The change in curvature suggests that perhaps ribosomal rRNA or protein is partially unfolding and contributing both additional length and additional "floppy" tether components.

The purpose of sucrose cushioning is to remove translation mix components (elongation factors, release factors, amino acids, tRNAs) that would eventually cause the ribosome to either

terminate or bypass the stalling site; in theory the tethering and OT experiment can still take place in the presence of these components since the tethering is specific via ligation to the CoA oligo and biotinylation of the N terminus. However, when we did not remove protein components and simply kept stalled ribosomes on ice during experimentation (to minimize read through of stalling), it was difficult to form tethers and the beads were very sticky. It is likely that all of the protein components coated the bead surfaces and either caused beads to stick together during fishing and/or obscured the ribosome tethering sites. Therefore, it is advantageous to remove some or all of the components required for translation if possible.

In order to accomplish this without using sucrose cushioning, we can take advantage of the fact that the PURExpress kit we use for stalling contains all His-tagged proteins. The recommended protocol from NEB for isolation of the *in vitro* translated proteins is to remove the ribosomes via size exclusion or ultrafiltration, then use a nickel resin to bind translation factors, spin down the resin, and reserve the supernatant. As this method does not seek to isolate the ribosomes, we developed an alternate method of removing the components of the translation mix using nickel-functionalized magnetic beads. In this way we should be able to isolate stalled ribosomes in a mild manner (in buffer, on ice, no centrifugation steps). Note that this does not remove amino acids or tRNAs from solution.

First we tested the binding capacity of the magnetic nickel beads (His-Tag Dynabeads) via silver staining. For the equivalent amount of protein factors required for a small (10uL) PURExpress reaction, 2 uL of beads resulted in undetectable levels of protein remaining in the supernatant (Figure 3.17). This was tested with the PURExpress ΔRibosome kit solution A, which does not contain ribosomes or tRNAs. Once we added both ribosomes and tRNAs to the solution, we can see that these nucleic acid components do indeed remain in solution after exposure to the Dynabeads, as expected since they do not have His tags. Experiments with RNase allowed identification of mRNA, tRNA, and rRNA bands in the silver stain gel (which stains both proteins and nucleic acids with subnanogram sensitivity).

As shown in Figure 3.17, once ribosomes are added the same quantity of beads is no longer sufficient to bind all factors in solution. However, this is not due to biding capacity since many proteins are 100% removed rather than all factors being partially removed. Furthermore, tests with more beads did not result in better removal of the residual factors. Instead, it seems specific bands are only partially removed, suggesting that they may be interacting with the ribosome. However, even with partial depletion translation should still be halted since all factors are required for multi-round translation.
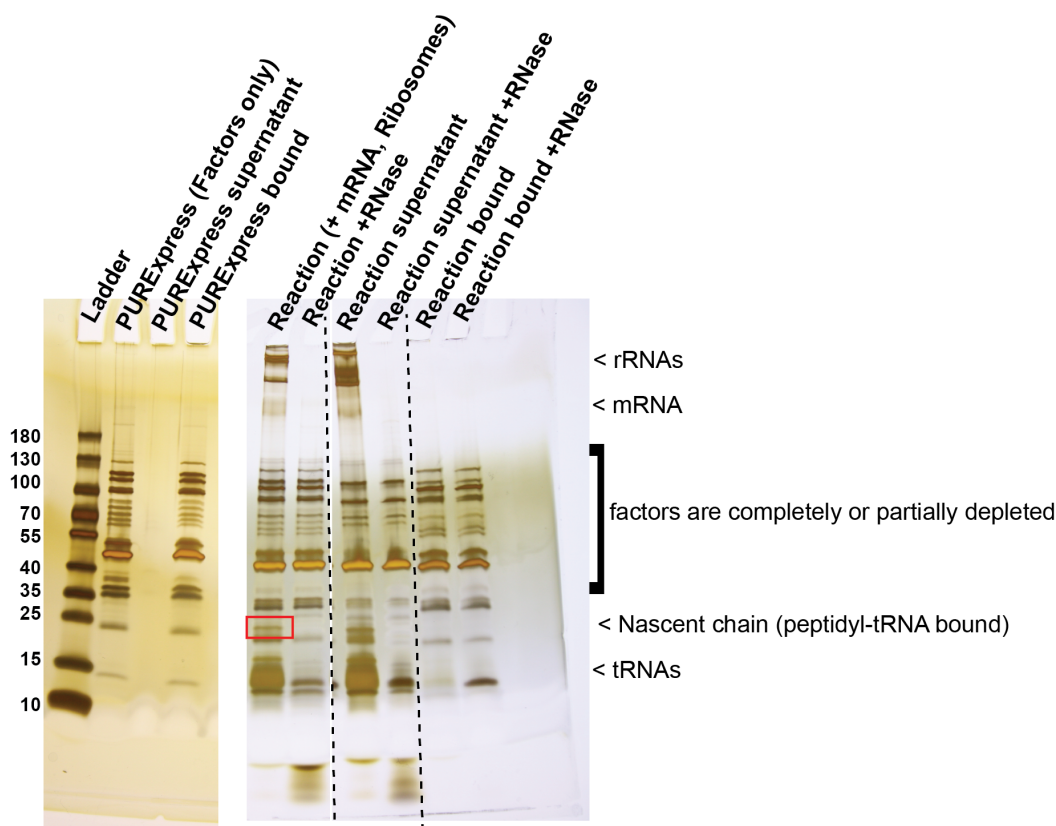
**Figure 3.17: Silver Staining of PURExpress in vitro translation reactions.** Initial experiments confirmed 100% binding of all protein factors included in the translation mix (lanes 1-3). Reaction conditions (with added ribosomes, mRNA, tRNA, etc.) show that the ribosomes do not stick non-specifically the Dyna beads at all. All protein factors are either completely removed or partially removed (lanes 6-9).

Even with our improved protocol, only about half of the tethers show the expected signature (Figure 3.16B). The remaining tethers are almost always the correct length/curvature, suggesting they are true single molecule tethers and not non-specific or multiple tethers. These may be due to: prematurely stalled translation in the SG linker such that the signature is not exposed; frame-shifted products that translate an alternate reading frame without the signature; or proteins that sustained multiple misincorporations such that folding is altered/abolished. Because our final tether design (SGV mRNA) does not include a signature this mixture of products will lower the throughput of our experiment. This also highlights the necessity for having a clear post-translation signature if lacking a clear pre-translation signature to verify tether identity.

Optical tweezers samples–real time elongation

Based on our optimization testing, the final reactions consisted of factor mix from the PURExpress Ribosome kit (mix with all protein components for initiation, elongation, and termination), solution A from the PURExpress ΔtRNA, Δaa kit (all buffer components, nucleotides, small molecules for the regenerating system and for formylating the fMet-tRNA), CoA tagged ribosomes, mRNA in a ~3-fold excess to ribosomes, birA, tRNA, amino acids,

tRNA-fMet (uncharged) and biotin. The concentrations used are listed in Table 3.3 at the end of the chapter. Translation reactions were conducted for 20 minutes at 25 degrees.

During this time, 2uL of His-Tag Dynabeads were prepared by washing 2-3 times with 1X polymix + 0.05% (w/v) TWEEN-20 and removing the supernatant. After stalling, the reactions were incubated with His-tag Dynabeads for 10 minutes on ice to remove the non-ribosomal protein components of the PURE kit. The supernatant was removed from the magnetic beads, ligated to the handle for OT experiments in the same manner as for stalled RNCs, and kept on ice until ready to use.

A table of real-time elongation conditions used in the optical tweezers is listed in Table 3.4. Plasmids for expression of all 20 tRNA synthetases (RSs) were a gift from Susan Marqusee, and the majority of the synthetases were purified and tested by Dr. Liang Meng Wee. The required concentration of each synthetase was determined by measuring charging activity with radiolabeled $[\alpha\text{-}^{32}P]$-ATP; an example is shown in Figure 3.18:
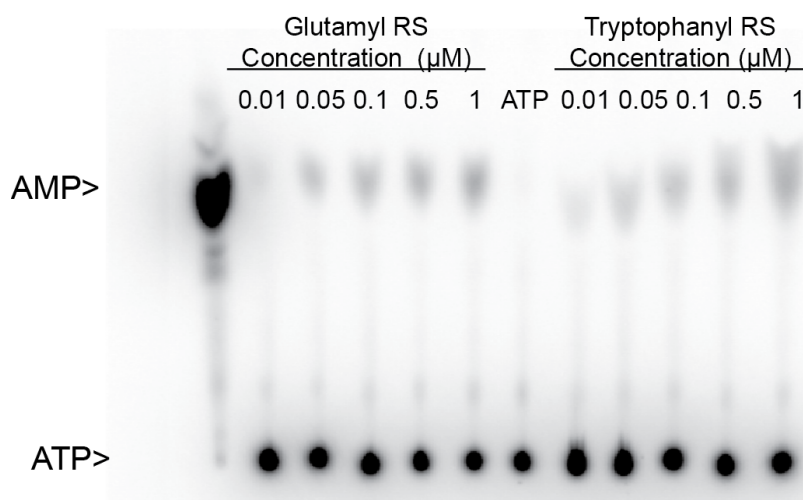


**Figure 3.18: Testing synthetase activity via $[\alpha\text{-}^{32}P]$-ATP hydrolysis** Testing for synthetase activity was conducted by reacting 1mM amino acid, 12µM ATP+0.03µM radioactive $[\alpha\text{-}^{32}P]$-ATP, 140µM tRNA with a given concentration of RS for 5' at 25 degrees (*85*). The sample was quenched with acidified sodium acetate and visualized via TLC. If the synthetase is active, it will charge the tRNA in an ATP-dependent reaction, generating AMP. In this example, GluRS and TrpRS are tested for activity. The spots can be quantified and fit to find saturating conditions. Further tests were conducted of each RS as a time course and at 25 vs. 37 degrees (not shown). Final concentration was 0.5µM of GluRS and 1 µM TrpRS.

Total tRNA (Roche) was pre-charged with S100 and all amino acids, purified via phenol-chloroform extraction and subsequent ethanol precipitation, then resuspended in 1 mM KOAc pH 5.3. These tRNAs were then added to in vitro translation mix with purified RSs to maintain the equilibrium population of charged tRNAs. Although this step should not be necessary, it removes any lag time while waiting for initial charging to occur and places less stress on the regenerating system since synthetases are ATP dependent. The synthetases then act to recharge any deacylated tRNA, either due to activity of the ribosome or spontaneous deacylation.

<u>Data analysis</u>

Data were collected on either of two optical tweezers set-ups, both previously described(*61, 86*) and set to the same trap stiffness. Data were collected at 2500 Hz for equilibrium measurements and 1333 Hz for elongation measurements, which occur on a longer timescale. Data were used at full frequency for all fittings and analysis, but are displayed downsampled at 250 Hz for equilibrium data figures, 125 Hz for force-extension curves, or 26.67 Hz for real-time data figures.

The HMM fitting and Bell's model fitting were performed as described in Chapter 2. For conversion of folding transition sizes on the ribosome, the same parameters were used as for the free protein. This conversion assumes the amount of peptide within the tunnel stays approximately constant throughout elongation, and thus the amount of linker peptide is constant. Experimentally, values for the number of amino acids within the tunnel vary between 30-35 amino acids (*71, 87*). For real-time data where we do not have a fixed protein length, we used the end point of translation, which has the clear hopping of the C domain and a well-defined length, as an internal reference for the length and back calculated the amino acid addition from the end. In a subset of molecules, we observed occasional small shifts in force in real time data, which could be either positive or negative, that we ascribed to inhomogeneity in the flow. These shifts were removed for the conversion to amino acids.

Real-time data was collected in passive mode to avoid artifacts of missed folding transitions and depressed apparent folding rates (same as for the other data described) (*32*). In passive, one individual amino acid is below our resolution (calculated step size for +1 amino acid is -0.006 pN at an initial force 4.5 pN and no added amino acids except linkers; calculated step size for +1 amino acid to a chain of 224 amino acids at 3.5 pN is -0.003 pN). Empirically, our noise at 4 pN is 0.25 pN at 1333 Hz. Assuming a desired signal:noise of 2-2.5, this would be equivalent to 10 steps at 13.33 Hz. Thus, although we cannot say where steps occurred, we can conclude that the chain has elongated by 10 amino acids within a certain window.

Because we cannot identify 1 amino acid steps and assigning 10 amino acid steps is unphysical, we instead used a basic isotonic regression which assumes that all the behavior is unidirectional but does not enforce a uniform step size. The fitting merely assigns a most probable trajectory–some "steps" will be larger or smaller than 1 amino acid. Individual steps of the algorithm are not interpreted– the reported translation rate is from the median value of the time to cross a 10-amino acid window, while the error is the standard deviation. For determining $\tau_{delay}$, the first point of the fit after the specified codon position is used as the time of arrival to that codon. The fit to $\tau_{delay}$ is to the cumulative distribution function (CDF), the integral of the probability distribution function (PDF). The simplest model would correspond to an exponential fit to $1 - e^{-\tau/\tau_{delay}}$. This is a single step stochastic process. For a multi-step process, the function would be a Gamma distribution which would have a sigmoidal shape. A sigmoidal distribution fits our data poorly and so we have stayed with the exponential fit, even though there is some spread ($R^2 = 0.91$).

Tables

**Table 3.1: Fit and extrapolated parameters of EF123 and RNC177**

| | $k_{0, folding}$ $(s^{-1})$ * | $p$ value compared to RNC177 | $\tau_{0, unfold}$ (s) * | $\Delta x^{\dagger}_{folding}$ (nm) | $p$ value compared to RNC177 |
|---|---|---|---|---|---|
| RNC177 | 1.8 (1.1, 2.9) x10$^3$ | n/a | 5.6 (3.4, 9.3) x10$^{-4}$ | 8.1 ± 0.7 | n/a |
| EF123 (to misfolded) | 4.4 (1.2, 16) x10$^7$ | 1.4x10$^{-5}$ | 2.3 (0.6, 8.3) x10$^{-8}$ | 16.6 ± 1.3 | 3.8x10$^{-5}$ |
| | $k_{0, unfolding}$ (s$^{-1}$) * | $p$ value compared to RNC177 | $\tau_{0, fold}$ (s) * | $\Delta x^{\dagger}_{unfolding}$ (nm) | $p$ value compared to RNC177 |
| RNC177 | 0.5 (0.4, 0.7) | n/a | 1.9 (1.4, 2.5) | 3.7 ± 0.4 | n/a |
| EF123 (to misfolded) | 5.9 (0.8, 43) x10$^{-3}$ | 9.1x10$^{-4}$ | 170 (23, 1200) | 4.4 ± 1.6 | n.s. (0.22) |

* Note that the rates and lifetimes are derived from linear fits to ln(k), and as such the error propagation is quite large when exponentiated and in some cases unphysical. These estimates are using the 95% confidence intervals from the fits as upper and lower bounds for the exponentiated value, expressed in parentheses as (lower, upper).

**Table 3.2: Fit and extrapolated parameters of FL$_{int}$ and RNC177$_{+45}$**

| | $k_{0, folding}$ $(s^{-1})$ * | $p$ value compared to FL$_{int}$ | $\tau_{0, unfold}$ (s) * | $\Delta x^{\dagger}_{folding}$ (nm) | $p$ value compared to FL$_{int}$ |
|---|---|---|---|---|---|
| FL$_{int}$ | 7.7 (3.3, 18) x10$^5$ | n/a | 1.3 (0.5, 2.9) x10$^{-6}$ | 7.1 ± 0.6 | n/a |
| RNC177$_{+45}$ | 2.1 (0.03, 150) x10$^3$ | 0.018 | 4.7 (0.07, 340) x10$^{-4}$ | 4 ± 3 | n.s. (0.068) |
| | $k_{0, unfolding}$ (s$^{-1}$) * | $p$ value compared to FL$_{int}$ | $\tau_{0, fold}$ (s) * | $\Delta x^{\dagger}_{unfolding}$ (nm) | $p$ value compared to FL$_{int}$ |
| FL$_{int}$ | 5.5 (3.2, 9.3) x10$^{-2}$ | n/a | 18 (11, 31) | 4.1 ± 0.3 | n/a |
| RNC177$_{+45}$ | 0.19 (0.04, 0.89) | n.s. (0.082) | 5.1 (1.1, 24) | 4 ± 1 | n.s. (0.27) |

* Note that the rates and lifetimes are derived from linear fits to ln(k), and as such the error propagation is quite large when exponentiated and in some cases unphysical. These estimates are using the 95% confidence intervals from the fits as upper and lower bounds for the exponentiated value, expressed in parentheses as (lower, upper).

**Table 3.3: Conditions for preparing stalled RNCs for real-time elongation**

| | [Stock] | [Final] | Volume |
|---|---|---|---|
| Solution A ΔtRNA, Δaa | 5X | 1X | 2.00 |
| Solution F ΔRibosome | 8.3333X | 1X | 1.20 |
| Total tRNA | 1800 μM | 100 μM | 0.6 |
| biotin | 1000 μM | 75 μM | 0.75 |

| birA | 25 μM | 1 μM | 0.4 |
|---|---|---|---|
| fMet tRNA | 65 μM | 4 μM | 0.6 |
| Amino acid mix (all except RCPYV) | 1000 μM | 300 μM | 3 |
| mRNA | 12.4 μM | 0.6 μM | 0.6 |
| Ribosome | 2.7 μM | 0.2 μM | 0.8 |

**Table 3.4: Concentrations for real-time elongation in tweezers**

| | Concentration |
|---|---|
| HEPES pH 7.5 | 20 mM |
| KCl | 95 mM |
| $MgCl_2$ | 7.4 mM (5 mM "free") |
| $NH_4Cl$ | 5 mM |
| $CaCl_2$ | 0.5 mM |
| Spermidine | 1 mM |
| Putrescine | 8 mM |
| $NaN_3$ | 10 mM |
| DTT | 1 mM |
| ATP | 1 mM |
| GTP | 1.4 mM |
| S100 charged tRNA | 30 μM |
| Uncharged tRNA | 35 μM |
| EF-Ts | 1.5 μM |
| EF-Tu | 10 μM |
| EF-G | 10 μM |
| Creatine Kinase | 3.9 μg/mL |
| Myokinase | 3 μg/mL |
| NDPK | 0.8 μM |
| Pyrophosphatase | 1 μg/mL |
| RNase Out inhibitor | 0.15 U/uL |
| Alanyl tRNA synthetase | 0.4 μM |
| Arginyl tRNA synthetase | 0.5 μM |
| Asparaginyl tRNA synthetase | 0.2 μM |
| Aspartate tRNA synthetase | 0.3 μM |
| Cysteinyl tRNA synthetase | 0.5 μM |
| Glutaminyl tRNA synthetase | 0.05 μM |
| Glutamyl tRNA synthetase | 0.5 μM |
| Glycyl tRNA synthetase | 1 μM |
| Histidyl tRNA synthetase | 0.2 μM |
| Isoleucyl tRNA synthetase | 0.2 μM |
| Leucyl tRNA synthetase | 0.5 μM |
| Lysyl tRNA synthetase | 0.11 μM |
| Methionyl tRNA synthetase | 0.05 μM |
| Phenylalanyl tRNA synthetase | 0.8 μM |
| Prolyl tRNA synthetase | 0.6 μM |
| Seryl tRNA synthetase | 0.075 μM |
| Threonyl tRNA synthetase | 0.75 μM |
| Tryptophanyl tRNA synthetase | 1 μM |
| Tyrosyl tRNA synthetase | 0.2 μM |
| Valyl tRNA synthetase | 0.1 μM |

# Chapter 4.

**Extending optical tweezers capabilities to probe protein folding with FRET**

### 4.1 Introduction

Optical tweezers gives a read out of molecular length between the tethering points, best suited to changes of more than 50 amino acids. However, there are many interesting dynamics which could be occurring at a smaller scale than this, both during folding and while the protein is active and undergoing conformational changes (binding, catalysis, etc.). We are interested in adding a parallel detection mode to follow an independent readout of folding along another reaction coordinate. To understand why this would be useful, consider the potential folding pathways shown in Figure 4.1:
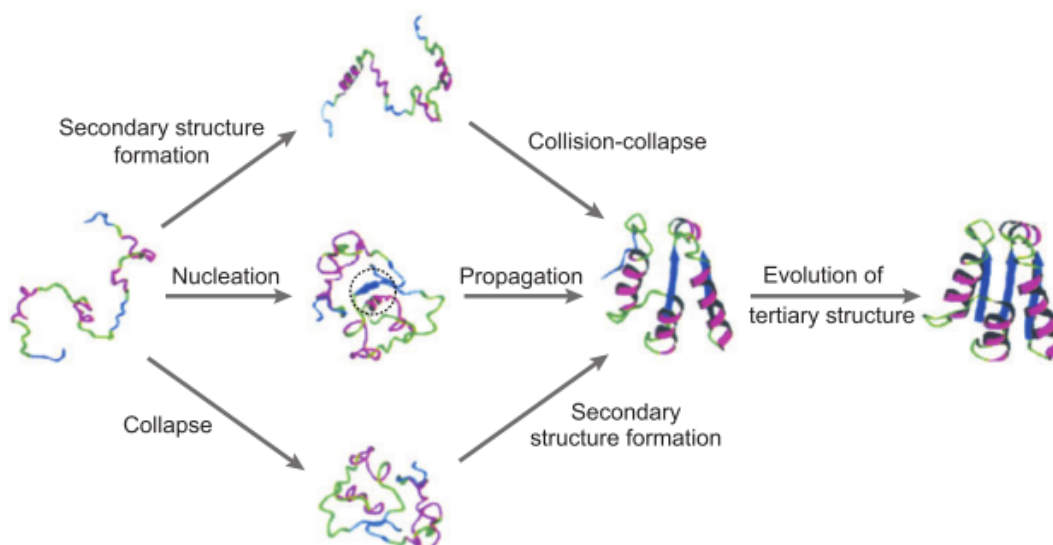


**Figure 4.1: Potential mechanisms of folding**. From reference (*88*). An unfolded protein could start forming local secondary structure (top pathway) which then condenses to a stable tertiary

structure. In the middle pathway, native contacts from disparate parts of the sequence form, which then nucleates further structure formation. In the bottom pathway, a hydrophobic core is formed, and then secondary structures form around this core.

In the OT, the pathways shown in Figure 4.1 could all potentially give the same end-to-end extension. If secondary structure forms first, it would not cause a large change in distance, thus we would only observe folding after secondary structures condense. For a collapse model, we would immediately readout this change in length, but then would be blind to rearrangements within the structure. All of these pathways, however, could be differentiated by the distance between local side chains, especially if we could also correlate this distance with the end-to-end distance as well. Since we want to time correlate two signals, they need to be in a well-defined synchronized population, making single molecule based techniques an obvious choice.

FRET (Förster resonance energy transfer) based single molecule assays have the sensitivity to measure these local distances, and are compatible with OT designs (Figure 4.2) (*86, 89, 90*). The addition of a FRET channel into an optical tweezers set-up allows for the ability to probe both local and global changes to the structure. FRET allows the measurement of states that do not correspond to changes between the tethering points, are smaller, or distinguishes states that are indistinguishable via tweezers.
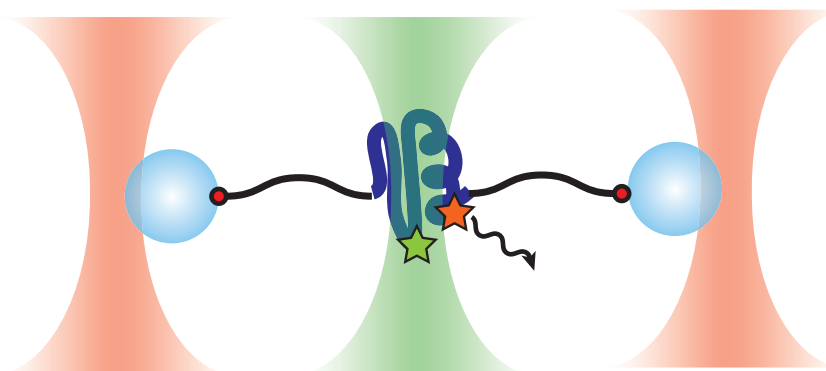


**Figure 4.2: Geometry for dual trap optical tweezers with confocal fluorescence** A dual trap OT set-up is used to extend a single protein. Between the two trapping lasers (red beams) there is an excitation beam (green) that will excite molecules within a diffraction limited confocal volume. By incorporating a FRET pair (green and red stars) onto the protein of interest and exciting the donor fluorophore (green), we can detect emission (FRET, see section 4.2) from the acceptor fluorophore (red).

Specifically, we are interested in looking at local secondary structure formation (alpha helices) compared to the signal from the tweezers. Although we know the tweezers signal corresponds to a certain number of amino acids, it cannot resolve if it is in an "active" or "native" conformation, and the tweezers signal is often cooperative for entire domains. Does this mean the domain folds all at once, or does protein rearrange after hydrophobic collapse into stable helices and secondary structure? Does secondary structure precede the tweezers signal and help "seed" tertiary formation? By looking at the temporal correlation between FRET and optical tweezers signals we can address these questions.

## 4.2 FRET background

FRET is the process of non-radiative energy transfer from one fluorophore to another. Because of this energy transfer, one fluorophore can be excited (the donor), but the detected fluorescence emission will be from the second fluorophore (the acceptor). The excitation and emission spectra of the common Cy3/Cy5 FRET pair are shown in Figure 4.3 to illustrate this point. In this case, excitation of Cy3 at a low wavelength where there is little to no Cy5 excitation (<550 nm) would normally result in observing Cy3 fluorescence emission (peaked at 570nm). If transfer to Cy5 occurs, however, the emitted light comes from Cy5 and follows its emission profile, i.e. peaked near 680nm.
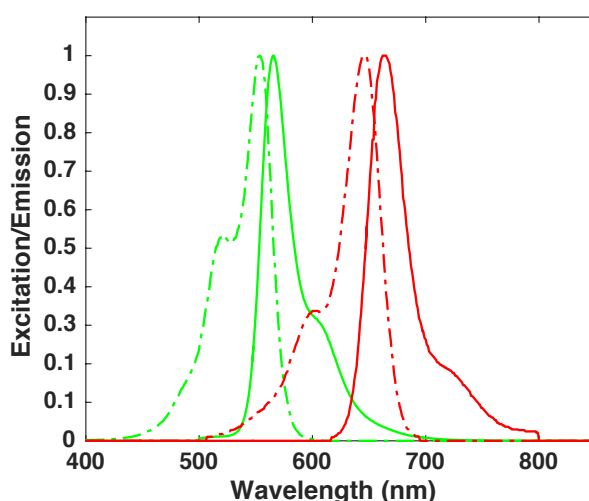


**Figure 4.3: Excitation/emission spectra for a pair of FRET fluorophores (Cy3/Cy5)**
Excitation spectra are shown as dash-dot lines; emission spectra are solid lines. The spectral overlap is defined in part by the area under the green solid line and the red dash-dot line, although solving the full integral is slightly different than the visual overlap.

This efficiency of this transfer process depends the distance between the two fluorophores with $1/r^6$ dependence according to the equation:

$$E = \frac{1}{1 + (\frac{R}{R_0})^6} = \frac{I_{Acceptor}}{I_{Donor} + I_{Acceptor}}$$

Because of the high-order R-dependence, the transfer falls off very rapidly, so it is most sensitive for small distance changes. Experimentally the efficiency can also be defined by the ratio of acceptor emissions to donor emissions by detecting fluorescence at two separate wavelengths. The other contributing factors besides distance are accounted for within the constant $R_0$, the Förster radius, the characteristic distance for two given fluorophores where half of the energy is transferred to the acceptor. $R_0$ is given by (*91*):

$$R_0 = C_0 (\frac{\kappa^2 J Q_D}{n^4})^{1/6}$$

where $C_0$ is 0.211 to express $R_0$ in Å, $J$ is the spectral overlap between the donor fluorophore's emission spectra and the acceptor fluorophore's excitation spectra, $Q_D$ is the quantum yield of the donor fluorophore, $n$ is the index of refraction of the medium, and $\kappa^2$ is the orientation factor that describes the directions of the fluorophores' dipoles (2/3 for freely rotating fluorophores).

Common $R_0$ values are in the 4-6 nm range, making FRET most sensitive for applications detecting distances of a few nanometers with an absolute distance that is 1-10 nm apart. This scale is convenient for biological applications, where one protein is often only a few nanometers wide and large complexes are tens of nanometers.

Single molecule FRET (smFRET) has been used to study protein folding in the context of chemical denaturants. For time-resolved results, the protein needs to be immobilized in some way; a single molecule will diffuse across a confocal volume too quickly. FRET trajectories as a function of denaturant can be collected, although in this case fluorophore lifetime may be limiting compared to the timescale of equilibrium folding. For example, Rhoades *et al.* measured the folding of adenylate kinase encapsulated in vesicles which were tethered to a surface (*92*) (Figure 4.4):
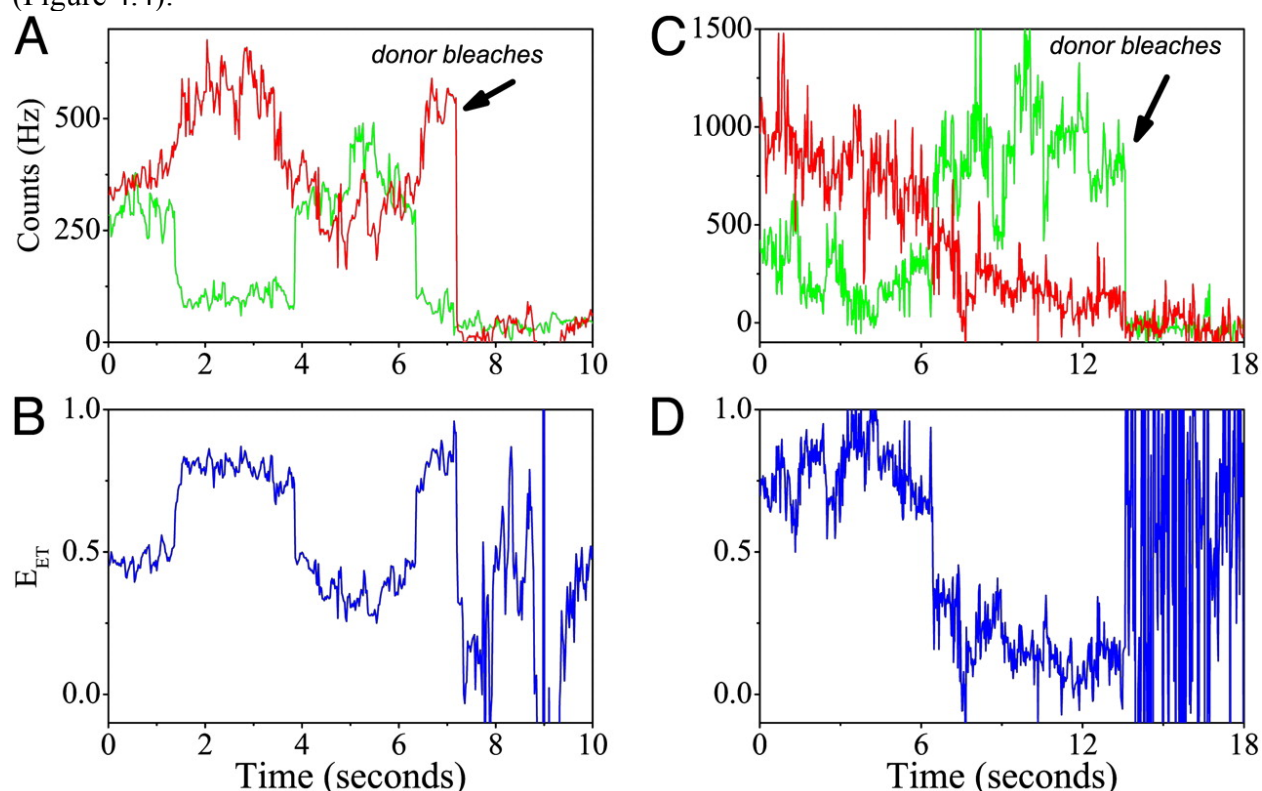


**Figure 4.4: Using smFRET to follow conformational changes during folding.** From reference (*92*). Single molecule trajectories are shown in (A) and (C). (B) and (D) show the respective FRET efficiency plots. Conducted at 0.4 M GdCl.

In this study, proteins in the folded conformation have a FRET efficiency value centered at 0.8 and unfolded proteins centered at 0.14. Adenylate kinase can fold through one intermediate state. What is interesting in this case is the timescale of transitions, which are more gradual than

folding observed in OT. It is not clear if this is a general feature of these folding transitions that is only revealed through FRET or if it is due to the presence of denaturant affecting chain dynamics. In either case the results use force, rather than chemical, denaturant, as well as the parallel readout of construct length, will provide interesting, multi-faceted data on folding.

## 4.3 Combining confocal fluorescence and optical tweezers measurements

The optical tweezers set-up by design holds one molecule suspended, away from the bead surface and other molecules that may be attached to those beads also. When even a relative low force is applied the DNA handles are extended to a large fraction of the contour length and therefore this molecule should be stationary within a relatively small volume as well. Thus, it should be amenable to confocal fluorescence microscopy with relatively low background. In practice, however, there are a number of technical limitations which must be overcome.

The first issue is the optical trapping beam. Most optical traps are made with IR light to minimize biological damage and heating, as well as the ready availability of strong lasers in this range. However, IR light is highly reactive with the excited state of many fluorophores (*93*) and can also cause two-photon excitation of ground state fluorophores (*94*). Even though the molecule is away from the trap, scattering of the trapping beam off of the beads can still impact the fluorophore. Early designs of OT and fluorescence saw fast bleaching by the trapping beam that severely curtailed the fluorescence lifetime (*95*). Thus, the trapping beam and the excitation beam cannot be on at the same time. In order to circumvent this issue, the trapping set-up needs to be modified to use an acousto-optic modulator (AOM) instead of simply splitting the trapping laser to create two traps. The AOM uses frequency modulation through a crystal to deflect the beam to different positions (*86*). By changing the frequency very quickly, the AOM can be used to generate two traps from one laser beam, although only one trap will be "on" at a time. Provided the cycling is sufficiently fast (MHz) the bead that was caught in a trap during its "on" cycle will not have time to relax and diffuse away during the "off" cycle, before the laser is on again and the bead is again caught in the trap potential, resulting in no change to the trap stiffness (*89*). Adding a third step to the cycle for fluorescence excitation allows for switching between the two optical traps and excitation without bleaching the fluorophore as quickly (*86, 96*).

The second issue is also related to damaged due to the trapping beam, although not from direct laser-induced reactivity. The IR laser can also react with oxygen to produce oxygen radicals, particularly with polystyrene beads acting as a sensitizer (*51*). This was previously discussed as a source of photodamage during folding, but these reactive oxygen species (ROS) are even more damaging to the excited state of a fluorophore. This effect can be mitigated by oxygen scavenging systems (GODCAT, PCA/PCD) or by triplet state or radical quenchers (azide, imidazole, trolox, COT). These radicals are also generated at the surface of the bead and usually react quickly before they can diffuse very far, and so their effect diminishes with distance (longer handles). Larger beads, which have less of their surface in the confocal spot and thus less of the sample in the confocal spot, will also produce less damage. Thus, longer handles, larger beads, and more complex buffer compositions can all help extend the lifetime of the fluorophore.

However, even with these improvements, the lifetime of a fluorophore in a combined fluorescence tweezers set-up is still shorter than in traditional fluorescence measurements.

While the instrument design discussed above has been published, the only work that has been done so far has been the fluorescence detection to detect binding of proteins from out of solution (*97*), rather than the fluorescent signal of an integral component of the tether. Experiments that use fluorophores within the tether have been reported for a single trap system detecting fluorescence at the surface, but this instrument design has low force resolution (*98*). For most FRET designs and for protein folding experiments in particular, these limitations must be overcome.

One of the primary challenges of using a fluorophore in the tether directly, rather than in solution, is that the sample is continually exposed to the trapping beam during experiment setup prior to forming a tether, a process which can take about a minute or even longer. In order to minimize the photodamage due to the beam and from the beam-derived ROS, we can lower the laser power while fishing to reduce the photon intensity, only raising the power to operation level (to ensure high temporal resolution) after the molecule has been stretch away from the surface. We also want the time in between catching a bead and beginning an experiment must be as short as possible, suggesting we want to work on the higher end of possible deposition concentrations.

The drawback with increasing deposition is with regards to the background. Although single molecule experiments are often described only in terms of the molecule of interest tethered between two beads, in reality there are many molecules on the surface of the bead, only one of which is tethered and measured. For FRET measurements where the fluorophore is incorporated into the tether, that means that one of the beads will have many fluorescent molecules on it and can contribute to the background signal due to stray excitation light from the excitation beam or dual-excitation from the trapping beam. Even uncoated beads will also contribute to the background through scattering or autofluorescence. In order to minimize these affects, the handles for fluorescence (and thus the distance from the confocal volume to the bead) needs to be relatively long (empirically, at least 2 kb on each side for 1 micron beads or 3 kb on each side for 2 micron beads).

When we have a FRET signal we expect the signal to be split between the donor and acceptor channels, making the absolute signal to noise ratio relatively low (lower donor intensity than if only performing confocal fluorescence). Fortunately, the FRET signal is anti-correlated between the channels, which will provide greater sensitivity in the signals. With all of these factors contributing to the signal and background, it is difficult to state exactly what our time resolution will be or what averaging will be required to generate sufficient S:N in fluorescence (at the expense of time resolution). The limits of temporal resolution will need to be empirically determined but are likely 50-100 Hz.

## 4.4 Initial experiments with calmodulin

Our model system in this case is calmodulin (CaM), a eukaryotic relative of calerythrin. Calmodulin also has 4 EF hands, like calerythrin, but the overall tertiary structure is an elongated "dumbbell" geometry rather than a globule shape (Figure 4.5).
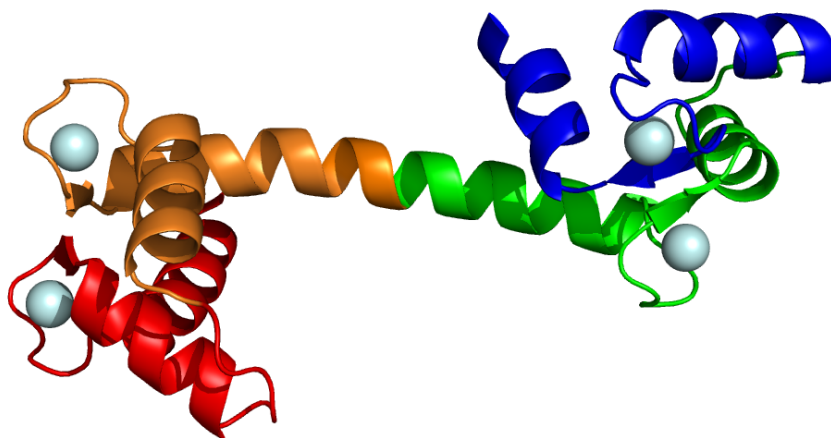


**Figure 4.5: Structure of calmodulin** From PDB: 3CLN (*99*). Here EF1, 2, 3, and 4 are color-coded red, orange, green, and blue, respectively. The two domains are separated by a long helix which connects EF2 and EF3. This is in the holo- conformation (calcium shown as light blue spheres), but it can undergo large structural rearrangements upon binding target proteins or upon loss of calcium (*39*).

Unlike calerythrin which binds calcium with nanomolar affinity and likely acts as a calcium buffer, calmodulin is a signaling protein and detects changes in calcium levels in the cell, binding with micromolar affinity (*39*). As such, it exists in the cell in both its apo- and holo-forms; the apo form is a more compact, "closed" structure where the EF hands are not at right angles to each other and the central helix is not continuous (*100*). In our experiments we will be describing folding of the holo-form, although the set-up could be easily extended to study apo-conformational changes.

Calmodulin has been extensively studied in optical tweezers (*42*, *48*) and thus we know it is well suited to the technical requirements of this study. The folding pathway has six possible states that are accessible: unfolded, C, N, fully folded, and two off-pathway states EF23 and EF123. A schematic of this pathway and an example of data in optical tweezers in shown in Figure 4.6.
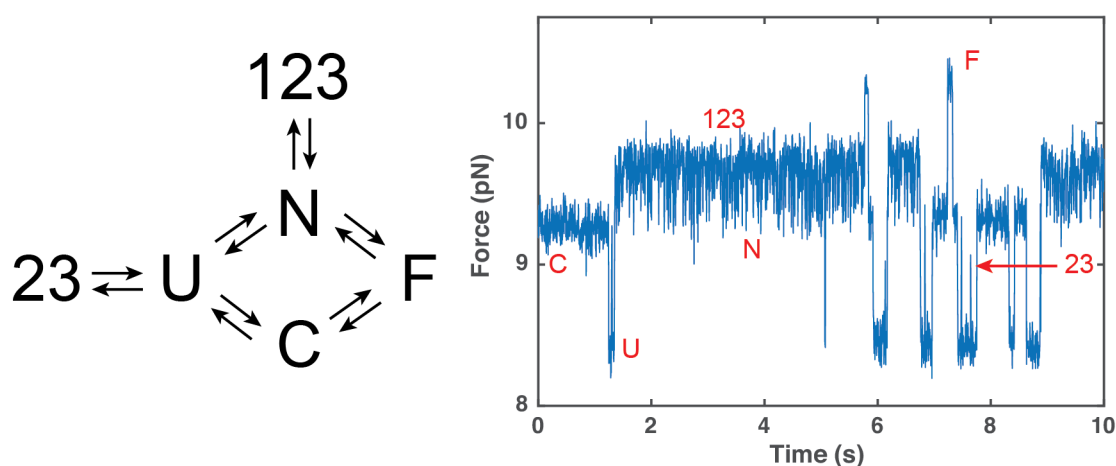
**Figure 4.6: Folding pathway of calmodulin.** The folding pathway of calmodulin has been characterized to have 6 possible states (*42*). Either the N or C domain can fold to the folly folded (F) state. Our data with WT CaM reproduces this pathway. In this panel, all 6 states can be seen. The N to 123 transition is very rapid and reversible.

In this case we would be most interested in the FRET of the C domain compared to the unfolded or folded state. For transitions between N and EF123, these are likely so fast that we will only record and average FRET value between the two states. Calmodulin is ideal because its folding is at very high forces, near 10 pN, which will provide the greatest S:N in the FRET channel (higher would be even better, but few proteins fold at higher than 10 pN). If we assume fluorophores which are 25 residues apart in a random coil, we can calculate the expected FRET efficiency as a function of force; because polypeptides are so flexible, the fluorophores will actually be quite close together at low forces, essentially indistinguishable from an ordered state (Figure 4.7). Only at or above 8 pN is the expected ΔFRET at a detectable limit.
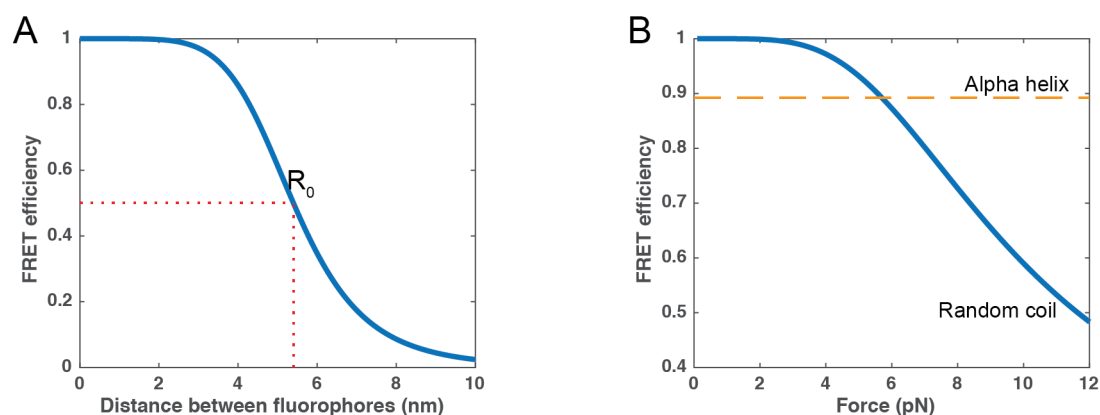


**Figure 4.7: Calculated FRET signal for fluorophores 25 amino acids apart in a helix versus random coil state.** A) The standard plot showing FRET efficiency versus distance, using a fluorophore pair with a $R_0$ of 5.6 nm (applicable for the Cy3 and Cy5 pair) (dotted red line). B) In the optical tweezers, the distance between the fluorophores is modulated by the force. Assuming a standard alpha helix that does not deform under force (a naïve assumption), this would give a constant FRET value near 0.9 (yellow dashed line). Once the polypeptide is in a

random coil conformation, it is very susceptible to changes in force according to its WLC behavior (blue line). We can use the WLC conversion to calculate the expected FRET efficiency as a function of force. This calculation is an overestimate of FRET since it does not include fluorophore linkers, which depend upon the attachment chemistry chosen; fortunately this means that the actual FRET change will be closer to $R_0$ and thus larger.

Calmodulin's folding kinetics are also on a time scale that is well-suited for FRET studies. Because we know the fluorescence lifetime is likely to be quite short (<10 seconds), we need to make sure we have multiple transitions in this time window (Figure 4.6 is a 10 second window, and multiple transitions are seen). However, since the fluorescence temporal resolution is lower than the OT, limited to < 100 Hz, we want transitions to be hundreds of milliseconds long. Calmodulin is perfect for this timescale (calerythrin, for example, is faster than this rate, even though it also folds at relatively high forces).

Calmodulin is also unique due to the bridging helix between the two domains. This helix presents a good labeling target because it is mostly solvent exposed and can change its configurations in different conditions (different buffers, adding target peptide, etc.). It will also allow us to observe what, if any, coordination exists between the folding of the two domains through this bridging helix. Thus, we selected sites at the opposite ends of the bridging helix that were relatively solvent exposed and not obviously involved in EF hand pairing to insert our fluorophore labels: P67 and V92 (Figure 4.8).
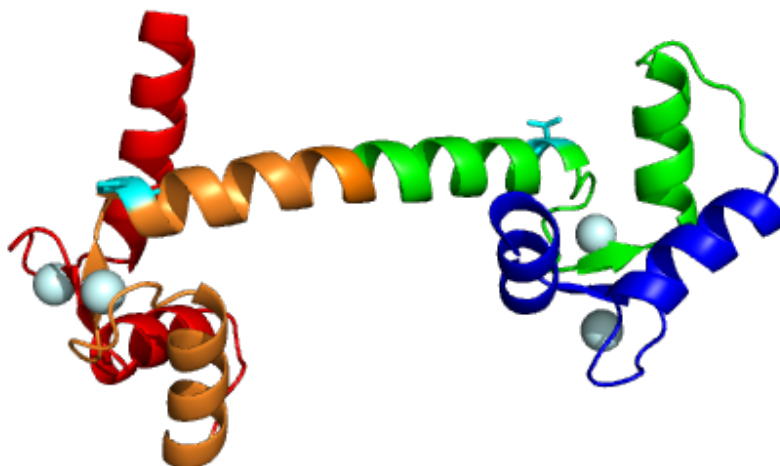


**Figure 4.8: fluorophore labeling positions in Calmodulin.** Labeling positions (cyan, showing side chains) were chosen to optimize FRET signal without disrupting expected structure. These positions are 4.1 Å apart (FRET = 0.87)

We verified that these labeling positions do not affect the overall folding pathway prior to labeling by purifying the P67C/V92C mutant. All of the 6 states are still observable in the OT data (Figure 4.9). A full kinetic characterization has not been completed, although this may help see more subtle affects on folding. For the questions we are interested in, however, the kinetic effects are not relevant as long as we can produce state assignments, which can be done visually for this protein at most forces.
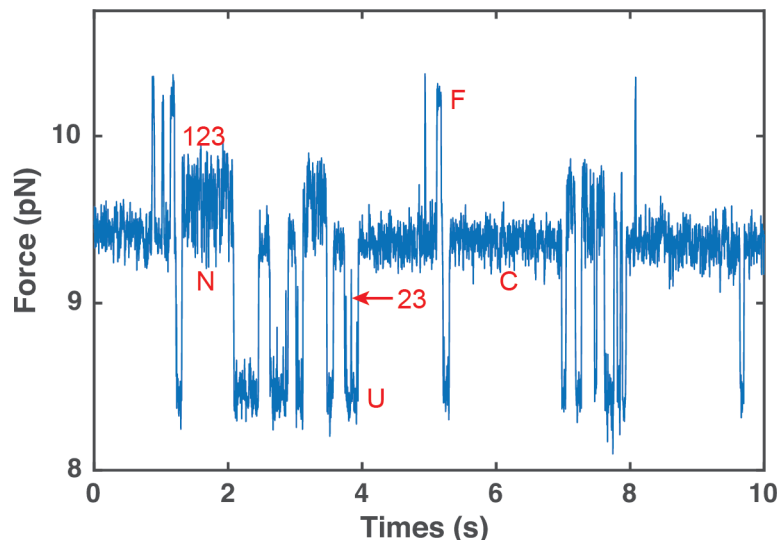
**Figure 4.9: The P67C/V92C mutant can still fold to the native structure.** The mutant folding pathway does not change. Compare to Figure 4.6. Possibly the N domain is slightly less stable relative to the WT, but still accessible on a similar timescale.

We tested the folding of P67C/V92C because our initial strategy was to use this double cysteine mutant of calmodulin to attach Cy3 and Cy5 fluorophores via a maleimide linkage. This strategy will result in a statistical mixture of labeled species: some proteins will have two Cy3 (25%), some two Cy5 (25%), and some will have one Cy3 and one Cy5 (50%, but mixed positions). Although the labeled protein is expected to be heterogeneous, the expression and purification are straightforward and the labeling can be done in one step. Moreover, the labeled species may be separable on HPLC (*101*). This strategy did not work well, however. Our HPLC was not as easily interpretable as the published results (Figure 4.10).
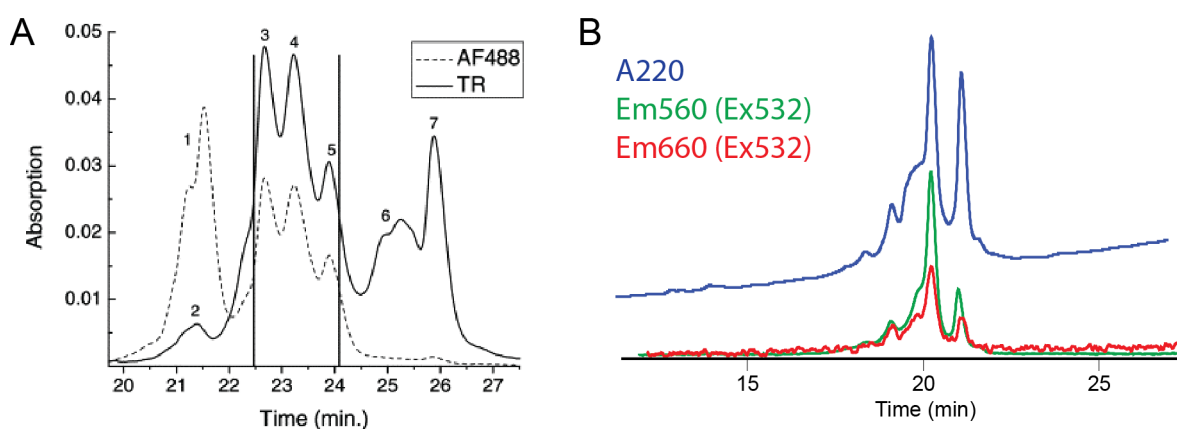


**Figure 4.10: HPLC profile of labeled P67C/V92C CaM** A) The published HPLC profile of a dual-labeled CaM for FRET (using different labeling positions and fluorophores) (*101*). In this case, peaks 1 and 2 were donor labeled, peaks 3-5 were dual-labeled, and peaks 6 and 7 were acceptor labeled. B) Our HPLC profile, measuring absorbance at 220 and fluorescence emission and 560 and 660 (only one excitation wavelength can be used on this instrument, so we cannot

directly excite Cy5). The intensities are scaled to show on one display; unscaled the Cy3 emission is about 6-7 times higher. We see that all of the protein peaks show both Cy3 and Cy5 FRET. It is unclear which peaks correspond to which species.

We also had very low overall labeling efficiency (81% labeled for peak 1; should be 200% labeled if both sites are saturated). Clearly, this column and solvent gradient combination cannot separate the labeled and unlabeled species since most of the protein is unlabeled or single labeled but these peaks are not distinguishable. Rather than optimize the HPLC conditions, we decided to forego HPLC purification and instead focus on maximizing labeling efficiency and in parallel designing an orthogonal labeling scheme.

Our first labeling used "standard" procedures for labeling as recommended by manufacturers. In this case, the protein is reduced with a 100-fold excess of TCEP and reacted with maleimides overnight. TCEP is a strong reducing agent towards thiols but should be non-reactive towards maleimides, thus does not need to be removed prior to the fluorophore reaction, in theory. However, there have been reports of undesired TCEP-maleimide reactivity (*102*), which may have contributed to our overall low reactivity since TCEP was present at much higher levels. Therefore, we attempted reduction with DTT instead of TCEP. However, DTT needs to be removed from reaction for example through size exclusion spin columns or dialysis. Both of these options are slow and may allow reoxidation to occur.

An interesting alternative to TCEP or the above methods of removing DTT is to reduce the protein, pellet it while DTT stays in the supernatant, and wash the pellet. Oxidation should be slow in this case because centrifugation can be performed quickly with fewer pipetting steps than a spin column and the thiols may be less accessible to $O_2$ in a solid form than in solution. We followed a modified labeling procedure that precipitates the protein in ammonium sulfate to remove DTT (*103*). Ammonium sulfate is a "gentle" precipitation but could still affect protein structure; calmodulin however is a robust folder so we are less concerned about the detractions of this method in this particular case. The results of this labeling showed 38 μM Cy3, 17 μM Cy5, and 15 μM protein suggesting perhaps complete labeling but incomplete fluorophore removal. This labeling preparation was characterized through a variety of means, summarized in Figure 4.11.

**Figure 4.11: Labeling efficiency of CaM P67C/V92C mutant.** A) UV/vis absorbance spectrum of the labeled protein. This allowed us to calculate the efficiency of labeling. B) A bulk fluorimetry measurement of the labeled protein. Excitation of 532 will give a signal from dual Cy3 species and a fraction of Cy3 emission from Cy3/Cy5 (depending on FRET efficiency). Emission at 633nm will only excite Cy5 labeled, both dualCy5 and Cy3/Cy5. C) The mass spectrometry spectrum shows the presence of each labeled species directly.

This characterization confirmed that the labeling is near completion and close to a statistical mixture, perhaps with a slight excess of dual Cy5 labeling. Comparing the mass spec result to the fluorimetry result, it seems likely that the FRET of the folded state is less than the predicted upper bound of 0.87. Even under ideal conditions, this protein will only give 50% usable sample in the OT due to dual Cy3 or dual Cy5 species. It would be advantageous to not have sample heterogeneity be one of the bottlenecks to data collection; this means we need different attachment chemistries for Cy3 and Cy5 that are orthogonal to each other. Common attachment chemistries with natural amino acids are through thiols, as we did, or amines. Amines are not site-specific enough since lysines are much more common than cysteines and also the N-terminus is reactive. Therefore we need to use a non-natural amino acid, such as azidophenylalanine. Azidophenylalanine can be incorporated by recoding the amber stop codon (see Methods section 4.6) and subsequently reacted with a DCBO- fluorophore.

For this mutant, we mutated P67 to the amber stop, that way any proteins that undesirably terminate at the amber stop will not yet have reached the cysteine and thus cannot compete for reactivity when we add the maleimide. We first labeled the azidophenylalanine with DBCO-Cy5, then tested labeling the cysteine with maleimide-Cy3 use both the precipitation method and spin columns after reduction with DTT. The precipitation method gave much better overall results and is the only case that will be discussed here. Overall the concentration of Cy3 was 8 µM, Cy5 was 15µM, and protein was 15µM, suggesting full Cy5 labeling and 53% Cy3 labeling (assuming all free fluorophore is removed). The protein preparation was characterized by a variety of means, summarized in Figure 4.12. Further testing of the mass spectrometry peaks needs to be conducted as we currently cannot assign all of the peaks.
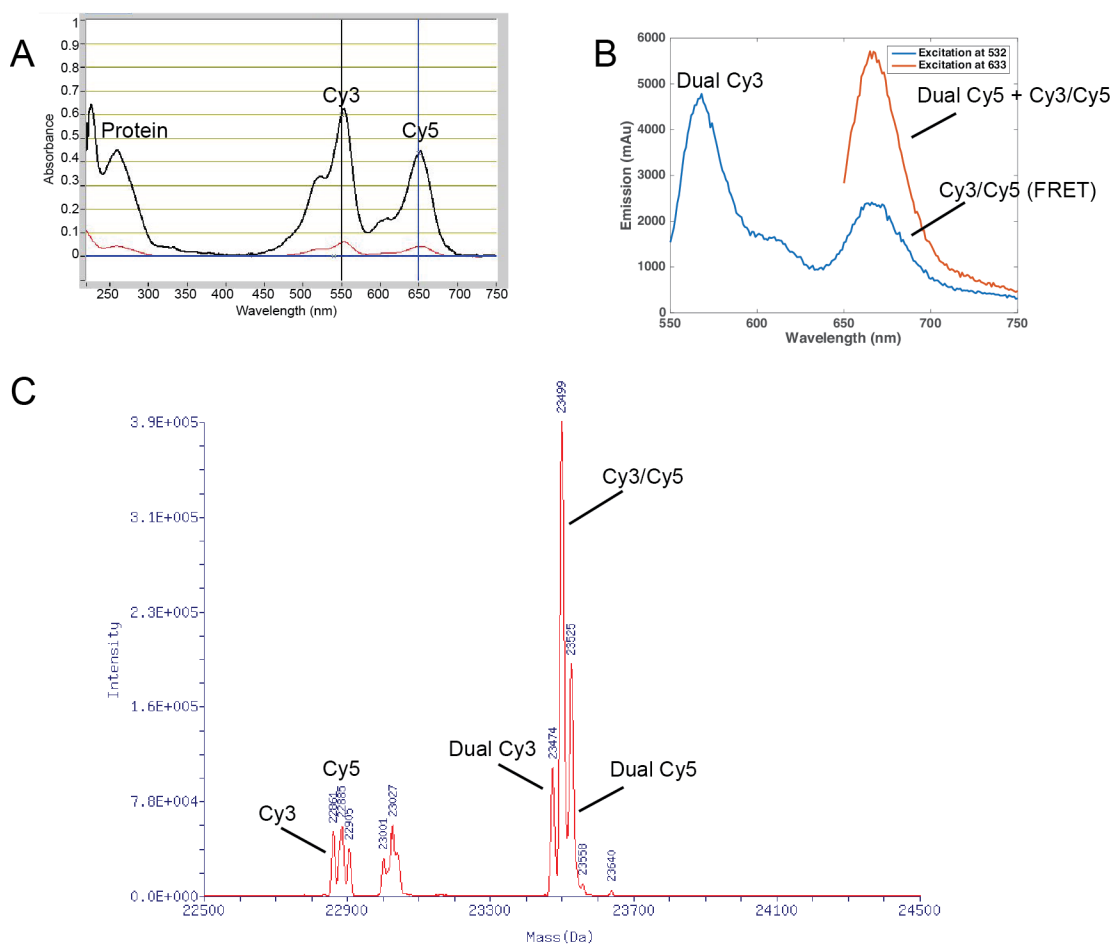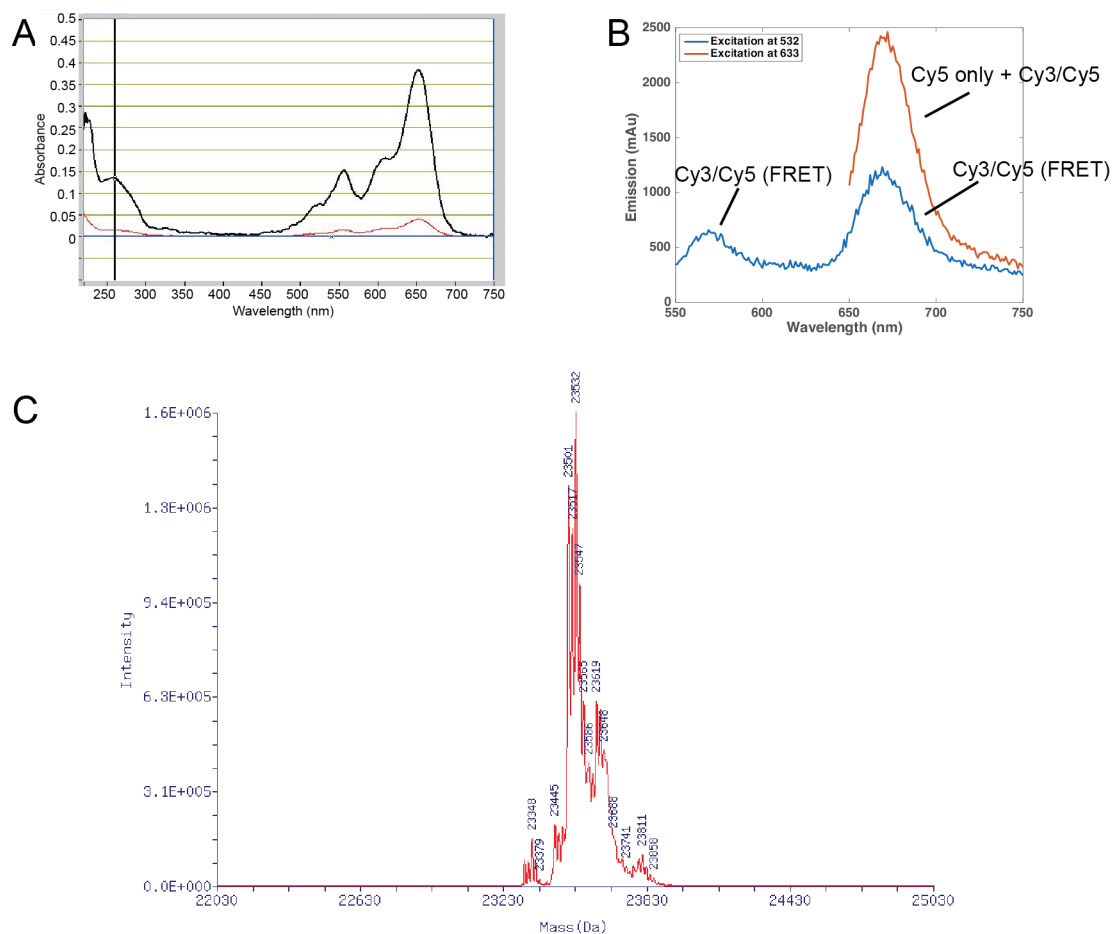


**Figure 4.12: Labeling efficiency of CaM P67AzF/V92C mutant.** A) UV/vis absorbance spectrum of the labeled protein. This allowed us to calculate the efficiency of labeling, which shows full Cy5 labeling but only 53% Cy3 labeling. B) A bulk fluorimetry measurement of the labeled protein. This is consistent with a FRET value of about 0.65 mixed with roughly 30% mono-labeled Cy5 protein. C) The mass spectrometry spectrum in this case is difficult to interpret and further experiments need to be done. It could be a mixture of species or represent contaminants/adducts that will not affect the tweezers.

Both of the aforementioned labeled proteins will give only about 50% throughput in the tweezers– the maleimide is fully labeled but a mixture and the azidoPhe/Cys mutant does not have full maleimide labeling. Therefore we continued testing both constructs at the single molecule level. The results are limited in throughput and need further testing; all numbers are preliminary. We do not have a laser in the wavelength to excite Cy5, so if we have a single Cy5 labeled protein or a Cy5/Cy5 labeled protein they would both appear dark. Both Cy3 and Cy5 are sensitive to the IR trapping beam can could be damaged prior to the experiment.

Considering 10 typical molecules, the P67C/V92C mutant had 5 without fluorescence (unlabeled, damaged, Cy5, or Cy5/Cy5), 2 with only Cy3 (Cy3 only or Cy3/Cy5 where Cy5 gets damaged), 1 Cy3/Cy3 and 2 Cy3/Cy5. We would expect to have 5 Cy3/Cy5 in this sample, so photodamage prior to experimentation is certainly happening. Worryingly, the lifetimes of Cy5 for these molecules are very limited. Two example molecules are shown in Figure 4.13:

**Figure 4.13: Simultaneous force and fluorescence measurements on P67C/V92C labeled protein.** For both molecules, the excitation laser was turned on at 5 seconds. A) The molecule was folded in the C domain, but both Cy3 and Cy5 bleached before transitions occurred. The molecule was in an intermediate FRET state. B) The molecule was in the N domain state and transitioned to the folded state near the time of bleaching but the Cy5 lifetime is too short to determine if a FRET change occurred. For both molecules Cy5 bleaches within about 1 second. Cy3's lifetime is only slightly longer, although we observed over 20 seconds in some tethers.

Since both constructs use Cy3/Cy5 fluorophores, we would expect the limitations to be similar. However, there are some indications that the type of attachment chemistry used can affect fluorophore lifetime (data from Cristhian Canari, unpublished). Therefore, the second labeling design may show improved lifetimes.

We tested the P67AzF/V92C labeled mutant in the optical tweezers under the same conditions. Considering 10 typical molecules, the P67AzF/V92C mutant had 8 without fluorescence (unlabeled, damaged, or Cy5 only) and 2 with Cy3/Cy5. Since we expect that our labeling is about 50% Cy3/Cy5, most of the "dark" molecules likely had Cy5 attached. Again, bleaching prior to experimentation is an issue despite optimizations in minimizing time prior to

data collection. However, we can see some fluorescence. Two example molecules are shown in Figure 4.14:



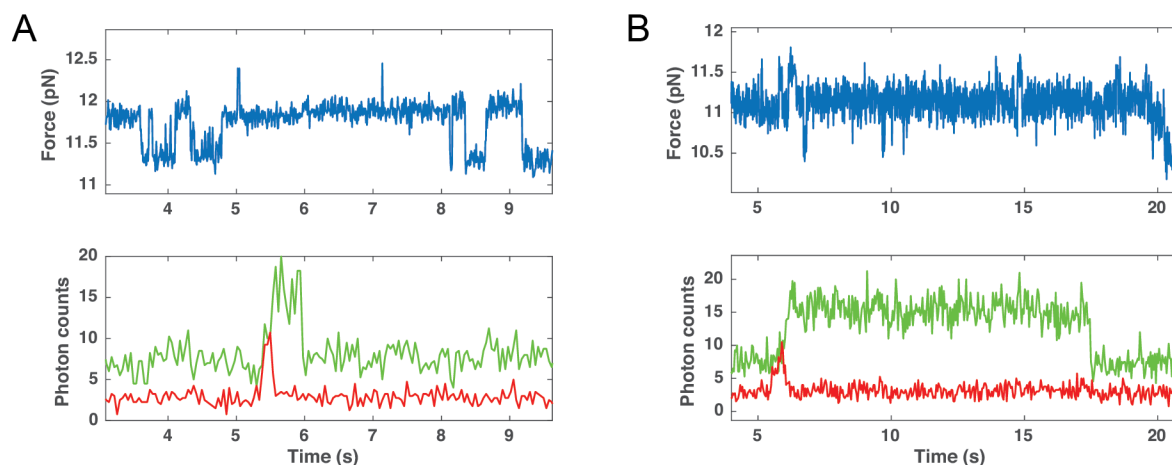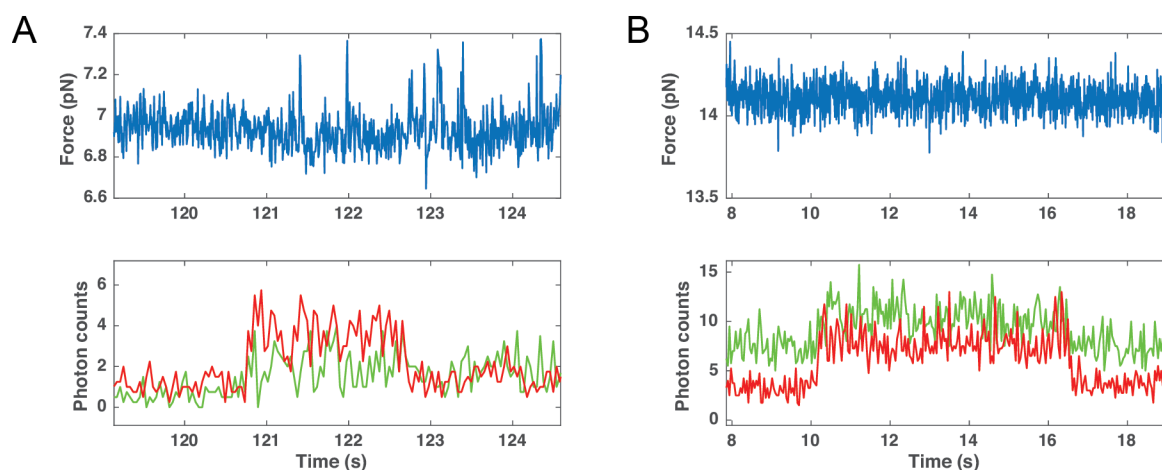**Figure 4.14: Simultaneous force and fluorescence measurements on P67AzF/V92C labeled protein.** A) For molecule A, the excitation laser was not turned on for at least 2 minutes due to difficulties finding a good force for hopping. In this case, the molecule was not hopping and was folded (high FRET). Cy3 bleached after 2 seconds, cutting short data collection. B) In this case we also could not find a good hopping region and the protein is unfolded (low FRET). Again the Cy3 bleaches first, after 6 seconds.

In this condition, we now seem to be limited by Cy3 lifetime, as evidenced by the simultaneous drop in Cy3 and Cy5, which indicates Cy3 bleaching first, rather than a drop in Cy5 and increase in Cy3 which indicates Cy5 bleaching first. This anecdotally confirms the importance of linker chemistry to dye lifetime. Further testing needs to be done to characterize this effect as well as determine Cy3 and Cy5 lifetimes. Unfortunately this cannot be optimized in bulk very easily since the main effect we are combating is that of the IR trapping beam.

As alluded to in Figure 4.14, the P67AzF/V92C mutant is difficult to find the "hopping" region. This detracts from the data quality in that it takes longer to set up the experiment. Moreover, as it is straightforward to find hopping in WT CaM and the P67C/V92C mutant, this also suggests that the folding may be altered in some way. Since we had tested the folding of the P67C/V92C mutant and found it unaffected, we neglected to test the folding of P67AzF/V92C since they were the same labeling sites. However, an azidophenylalanine mutation is much bulkier than a cysteine, and so this control should also be done. Fortunately, we need not do additional experiments because molecules that show no fluorescence or molecules after the fluorescence has bleached can be used to test folding with no more time constraints on speed of experiment. Using this data, we can compare the folding of WT to P67Az/V92C (Figure 4.15).
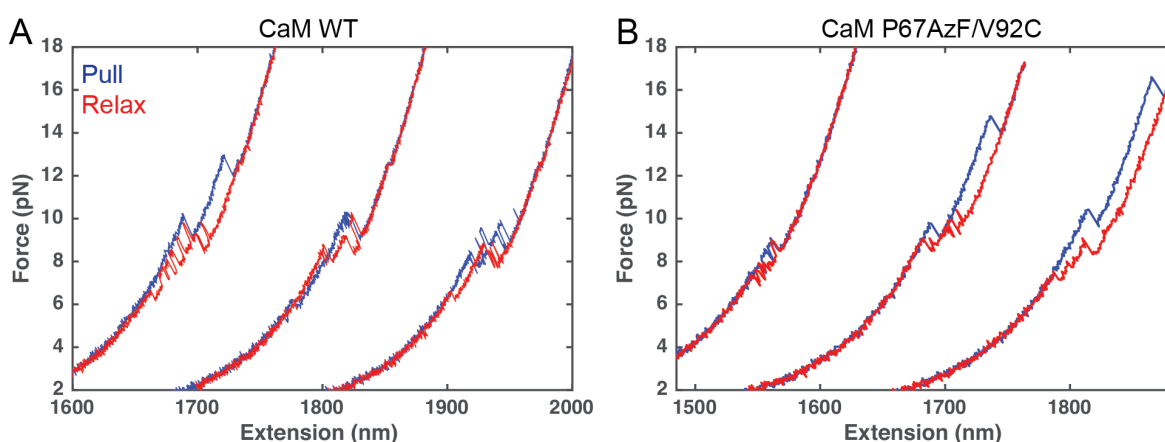
**Figure 4.15: Comparing the folding of P67AzF/V92C to WT CaM.** Force extension curves of can provide a broad overview of folding particularly if there is a large shift in forces. A) Force extension curves of WT CaM show two equal sized steps to folding around 10 pN; in some cases there are unequal steps corresponding to EF123 folding (see second blue curve and third red curve). B) The P67AzF/V92C mutant shows one domain folding/unfolding at 10 pN and one domain unfolding much higher, near 16 pN. In the first pull the second transition does not occur at all up to 18 pN tested. The EF123 state is much more rare.

This demonstrates why hopping is "hard" to find in the OT– the two transitions are no longer cotemporal and do not exist at the same equilibrium force. Only one transition can be observed at a time. Since the P67C/V92C mutant did not have changes to its folding and the V92C mutation and fluorophore linker is the same in both mutants, it is likely that the higher transition is the N domain. The addition of azidophenylalanine may stabilize the N domain and destabilize the EF123 state, although further characterization is necessary to confirm this. We also do not know if the N domain is in a native-like conformation once azidophenylalanine is incorporated, even if it is more stable.

## 4.5 Discussion

Although we have not yet achieved simultaneous measurements of folding in OT and single molecule FRET, there are very clear paths forward and our data this far is very exciting. Since the Cy5 maleimide lifetime seems to be severely limiting, it is better to pursue the azidophenylalanine/cysteine strategy. Given the effect on folding we have observed, the mutation site will need to be altered. Assuming the same methodology of incorporation and labeling, we would like to mutate a more hydrophobic residue but still near the surface. Figure 4.16 shows potential labeling sites.

A

B

```
         65      *          75
  •••••DFPEFLTMMARK•••••
           ︵_____︶
        Potential labeling sites
```

**Figure 4.16: Potential new labeling sites for azidophenylalanine incorporation.** A) The structure of calmodulin showing F66, F69, and V92 as cyan residues with side chains. B) The sequence surrounding the P67 labeling site. We would ideally like to label within ±5 amino acids of P67 so that we do not compromise the FRET sensitivity.
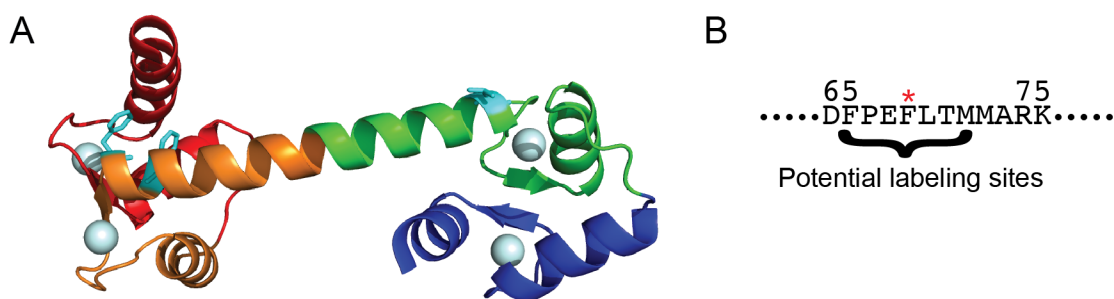
The sequence shows that there are already two natural phenylalanines in CaM near our desired labeling position. This may be related to why our labeling distorts the folding– the additional phenylalanine can interact with these nearby hydrophobic residues. However, we could simply replace a natural phenylalanine with an azidophenylalanine. F66 is very close to the calcium binding site and likely forms interactions with EF1; F69 is a more promising candidate. In either case, the addition of the fluorophore may still interfere with the structure, where phenylalanine side chain interactions may be critical. The linker to the DCBO dye is quite long, longer than for maleimide attachment, and this flexibility may help minimize disruption. However, until the mutants are tested it is purely hypothetical.

If these phenylalanines are required for structure and cannot tolerate the addition of a dye linker, we can also look into other incorporation sites, either of azidophenylalanine further away from the end of the helix or a different unnatural amino acid. In this case, we are limited by the availability of evolved orthogonal tRNA synthetases that will recognize an amino acid with useful reactivity. Since we are not labeling *in vivo*, we can also used a copper-catalyzed click chemistry reaction (rather than the catalyst-free strained octyne DBCO) and react with a terminal alkyne. This allows us to put the alkyne in the protein and the azido group in the fluorophore, switching the functional groups. There are some reported synthetases which can recognize alkyne modified amino acids– one which attaches an alkyne-modified phenylalanine (*104*) and one which recognizes an alkyne-modified lysine with a carbamate linkage (*105*). In this case T71 or K76 would be good candidate amino acids; K76 is of course a lysine already, but it is only 16 amino acids away from V92C and may be too close for best FRET sensitivity.

We can also try swapping in the labeling positions and using P67C, as in the cysteine double mutant, and then incorporating the unnatural amino acid into the C domain, where it may be more tolerated. As mentioned before, this would have the undesired side effect proteins of allowing proteins which terminate via RF1 at the amber stop to compete with the maleimide. Since we purify via the N-terminal Avi tag, both full length and truncated products will be biotinylated and bind the column. If we wanted to swap the labeling sites we may also want to swap the N and C terminal tags to purify via the C terminus instead and allow greatly selectivity to the desired protein. This may be beneficial in any case to have a better quantification of the

protein concentration that is not inflated by truncated species, but we have not seen significant amounts of truncated products in PAGE gels.

These optimizations are purely related to the biochemistry of folding and ensuring mutant protein stability after mutagenesis and labeling. However, there are also many variables to be optimized on the experimental and instrumentation side. First, a systematic investigation of bead size effects on fluorescence lifetime and efficiency of forming fluorescent tethers should be carried out. Although one micron beads have higher resolution, two micron beads may have less photodamage, which may considerably outweigh the loss of resolution considering the number of dark molecules which we observed. Also, our resolution will be limited by FRET, not in the tweezers, and so this slight loss from bead size may not affect overall sensitivity between the two channels.

We can also make further improvements to fluorophore lifetime through optimizing the buffer components (reducing agents, triplet state quenchers, etc.) and through modifying the oxygen scavenging system. Currently we are using glucose oxidase/catalase (GODCAT) with ascorbic acid, imidazole (*106*), and Trolox. We can try PCD instead of glucose oxidase (*107*), which has a lower resting $O_2$ concentration. Although ascorbic acid helps with fluorescence in bulk setups, it may actually reduce lifetimes in the presence of the IR laser (*93*). Finally, we can test additional additives such as cyclooctatetraene (COT).

If we are interested in making estimates on the inter-fluorophore distance, we will also need to perform calibration experiments to determine our detector-specific baselines that will allow for true FRET efficiencies and thus quantitative FRET values (*108*). Initial experiments with just high-low FRET changes are sufficient to detect folding transitions; this process just allows for more precise estimation of distance.

Once the conditions are optimized for this robust data collection, other systems could also be investigated. Of particular interest would be to observe the nascent chain using this combination methodology. Incorporating fluorophores into the nascent chain without perturbing the ribosome is a challenging task, but has been reported using an *in vitro* system with AUG (methionine) recoding for the labeling sites (*109*). Since the nascent chain may have different dynamics on the ribosome (*14*) or may be more compact, it would be interesting to monitor the FRET of the unfolded state in stalled complexes as well as associated FRET changes with folding. Once knowledge of equilibrium nascent chain dynamics is measured, the next logical step would be non-equilibrium measurements to see if the delay in folding also correlates to changes in the unfolded state distribution. Real-time elongation measurements would be very exciting, however, they are even more severely limited by fluorophore lifetime since translation is relatively slow in our conditions.

## 4.6 Methods

Double cysteine double mutant (P67C/V92C)

The calmodulin gene was purchased and inserted into the same backbone vector discussed in Chapter 2 via digestion with NheI and NotI restriction enzymes and standard transformation protocols. Site-directed mutagenesis was performed via PCR with non-overlapping, end-to-end primers that amplified the whole plasmid, with mismatches in the forward primer. The PCR product was gel purified if additional bands were seen, then treated with DpnI, PNK, and T4L ligase to yield circular for transformation. All constructs were verified by plasmid sequencing. Purification was performed as discussed in Chapter 2. Concentrations of CaM and mutants were determined via Bradford assay since the extinction coefficient is low and absorbance in the UV is peaked at 260nm, not 280nm due to calcium binding spectral shifts.

The first attempt at labeling reacted 36.2 nmol of P67C/V92C CaM with a 100-fold molar excess of TCEP for 10' at room temperature, flushed with nitrogen. After this time, a 10-fold molar excess of Cy3 maleimide and Cy5 maleimide (20 mM stocks in DMSO) was added, flushed with nitrogen again and reacted at room temperature for 2 hours then overnight at 4 degrees while wrapped in foil. Excess fluorophore was removed via two passes through 10DG columns (BioRad). The HPLC run used a C5 reverse-phase column with a gradient of water-acetonitrile (with added +0.1%TFA to both phases). After HPLC the samples were lyophilized and resuspended in 20 mM HEPES, 100 mM KCl, and 10 mM $CaCl_2$, then stored at -80.

The second labeling used 5 nmol of protein. Reduction used DTT at 10 mM concentration for 2 hours at 4 degrees. At this point saturated $(NH4)_2SO_4$ (called "100% s. $(NH4)_2SO_4$", s. for saturating) was added to a final concentration of 70% s. Mix gently for ten minutes; the protein should precipitate. To remove excess DTT, the sample was spun at 13000xg for 3' at 4 degrees. The supernatant was removed and the pellet was washed with 100 mM $Na_2HPO_4$ (pH 7.3) with 200 mM NaCl, 1 mM EDTA, and 70% s. $(NH4)_2SO_4$. Repeat centrifugation, then resuspend the pellet in 100 mM $Na_2HPO_4$ (pH 7.3) with 200 mM NaCl, 1 mM EDTA with Cy3 and Cy5 maleimide at 5-fold molar excess. This was reacted for 30' at room temperature then quenched with 72 mM β-mercaptoethanol for 10 minutes. To remove excess fluorophore, the sample was ran through two microbiospin P6 columns (BioRad) that were buffer-exchanged to 20 mM HEPES, 100 mM KCl, and 10 mM $CaCl_2$, then stored at -80.

Azidophenylalanine/cysteine double mutant

Expression of the calmodulin with the amber stop codon mutation was performed differently than for cysteine mutants to allow for incorporation of a non-natural amino acid at this site. Briefly, BL21 cells were cotransformed with our plasmid (under control of a lac operon and with and ampicillin resistance gene) and a plasmid that encodes for an amber codon tRNA as well as an evolved tRNA synthetase (*110*) that will recognize the amber tRNA and charge it with azidophenylalanine, under an inducible arabinose dependent promoter and with chloramphenicol resistance (*111*). After transformation, 2L of cells were grown to an OD of 0.6 in LB and gently pelleted. The cell pellets were resuspended in 670 mL of TB media and 2 mM of azidophenylalanine was added (this is to conserve material). Growth was allowed to continue for 30 minutes at 30 degrees to allow uptake of the artificial amino acid. At this point expression was induced with IPTG and arabinose, and growth continued for 5 hours at 30 degrees then

overnight at 16 degrees. After expression protein purification was carried out the same as described for calerythrin (Chapter 2).

Labeling of the maleimide functionality requires that all cysteines are completely reduced and available for reaction. However, the azido moiety can be sensitive to reducing agents. Therefore, labeling of the azido group must be conducted prior to labeling the thiol. 7.75 nmol of the P67AzF/V92C CaM was reacted directly with a 10-fold molar excess of DBCO-cy5 (20 mM in DMSO) in 20 mM HEPES, 100 mM KCl, and 10 mM $CaCl_2$ overnight at room temperature. DTT was added to 20 mM concentration and reacted for 1 hour at room temperature. At this point, the reaction was split in half; half was precipitated and half used two P6 spin columns to remove DTT. For both reactions, the maleimide was added to 20 mM HEPES, 100 mM KCl, and 10 mM $CaCl_2$ at a 5-fold molar excess. React 3 hours at room temperature then 3 hours on ice. Both samples had excess dye removed at the end via 2 P6 spin columns. Note that the precipitation protocol does not remove the Cy5DBCO prior to adding the Cy3 maleimide; this could allow more time for Cy5 to react but also could allow for undesired cross reactions once the reducing agents are added.

Optical tweezers samples

For OT samples, the labeled protein was reacted with Sfp and CoA oligo as previously described in Chapter 2. Handles were prepared in the same manner (PCR from lambda genome) except as either 2.5 kb DNA or 4 kb DNA. The ligation is the same. For experiments with 1-micron beads, 2.5 kb handles were used on each side for a total tether length of 5 kb. For 2-micron beads, 4 kb handles were used for a total tether length of 8 kb.

The buffer for optical tweezers experiments is 50 mM Tris pH 8, 100 mM KCl, 50 mM imidazole, 10 mM $CaCl_2$, 12.5 mM ascorbic acid, 0.8% glucose (w/v), 0.04 mg/mL catalase, 0.03 mg/mL glucose oxidase, and 2 mM Trolox with 0.8 U/mL RNase out. The buffer must be changed every 2 hours due to pH changes from GODCAT activity. Only the central channel needs the full suite of photostabilizing cocktail. The beads are resuspended in simply 50 mM Tris pH 8, 100 mM KCl, 50 mM imidazole, and 10 mM $CaCl_2$. Azide cannot be used in this case since it reacts with the catalytic site of catalase.

# Chapter 5.
## Conclusions

Through this work, we have demonstrated the importance of the ribosome as an effector of nascent chain dynamics, both at equilibrium conditions and while translating in real-time. In our stalled complexes, we demonstrated ribosomal effects on both the folding rate and the unfolding rate, acting to preferentially destabilize a misfolded structure during synthesis that would be a non-productive kinetic trap.

In real time, we documented the unexpected presence of a non-equilibrium relaxation time. Despite the slow rate of synthesis relative to folding, the polypeptide is not equilibrated on its energy landscape. This surprising finding, accessible only through monitoring the growth of individual nascent chains in real time, suggests that many common techniques for measuring ribosome-dependent folding may be obscuring important non-equilibrium effects. Hence, purified stalled ribosome-bound complexes (as tested here) or measurements of folding of non-synchronized populations of nascent chains would be blind to these non-equilibrium dynamics.

The non-equilibrium delay prior to folding may have important implications for multidomain folding and chaperone assisted folding, where the prolonged unfolded state lifetime would prevent kinetic trapping in off-pathway states or provide more time for chaperone recruitment to the ribosome. It may also be important in the interplay of elongation rates and folding, where transient pause sites could provide time for this equilibration to occur. This result also highlights the importance of ribosome pausing and synonymous codon effects wherein longer pauses may be necessary when the kinetics of folding are slower or delayed. Finally, our results stress the necessity of studying folding in a time-resolved, co-translational manner to better understand the dynamics of folding in the cell and highlight the potential perils of studying folding of full-length, isolated proteins.

It would be interesting to test the magnitude of this effect in other protein systems, as well as investigate the interplay of this relaxation with chaperone recruitment. We do not know the

sequence-dependence, if any, of this non-equilibrium behavior. This effect may work in tandem with rare codon usage to slow translation at key points in translation.

We have also demonstrated cotemporal measurement of optical tweezers signals and single molecule FRET. Although we are currently limited by fluorophore lifetime, our preliminary results show this experimental design is feasible. This opens up a wide variety of questions both for folding processes and for protein conformational changes associated with activity or regulation. This extension of optical tweezers can document early events in folding pathways and distinguish models in a way that single mode detection with either technique cannot.

# Bibliography:

1.  I. Tinoco, H.-K. Kim, S. Yan, Frameshifting Dynamics. *Biopolymers* (2013).
2.  A. Marintchev, G. Wagner, Translation initiation: structures, mechanisms and evolution. *Q. Rev. Biophys.* **37**, 197–284 (2004).
3.  K. B. Gromadski, M. V. Rodnina, Kinetic determinants of high-fidelity tRNA discrimination on the ribosome. *Mol. Cell.* **13**, 191–200 (2004).
4.  M. Zhu, X. Dai, Y.-P. Wang, *Nucleic Acids Res.*, in press, doi:10.1093/nar/gkw698.
5.  F. U. Hartl, M. Hayer-hartl, Converging concepts of protein folding in vitro and in vivo. *Nat. Struct. Mol. Biol.* **16**, 574–581 (2009).
6.  E. J. Guinn, B. Jagannathan, S. Marqusee, Single-molecule chemo-mechanical unfolding reveals multiple transition state barriers in a small single-domain protein. *Nat. Commun.* **6**, 6861 (2015).
7.  A. Borgia *et al.*, Transient misfolding dominates multidomain protein folding. *Nat. Commun.* **6**, 8861 (2015).
8.  N. R. Voss, M. Gerstein, T. a Steitz, P. B. Moore, The geometry of the ribosomal polypeptide exit tunnel. *J. Mol. Biol.* **360**, 893–906 (2006).
9.  R. B. Jacopo Marino, Gunnar von Heijne, Small domains fold inside ribosome exit tunnel. *FEBS Lett.* **590**, 655–660 (2016).
10. O. B. Nilsson *et al.*, Cotranslational Protein Folding inside the Ribosome Exit Tunnel. *Cell Rep.* **12**, 1533–1540 (2015).
11. E. Mercier, M. V. Rodnina, Co-translational Folding Trajectory of the HemK Helical Domain. *Biochemistry* (2018), doi:10.1021/acs.biochem.8b00293.
12. L. D. Cabrita *et al.*, A structural ensemble of a ribosome-nascent chain complex during in vivo co-translational protein folding. *Nat. Struct. Mol. Biol.* (2016), doi:10.1038/nsmb.3182.
13. C. M. Kaiser, D. H. Goldman, J. D. Chodera, I. Tinoco, C. Bustamante, The ribosome modulates nascent protein folding. *Science*. **334**, 1723–7 (2011).
14. A. M. Knight *et al.*, Electrostatic effect of the ribosomal surface on nascent polypeptide dynamics. *ACS Chem. Biol.* **8**, 1195–204 (2013).
15. H. X. Zhou, K. A. Dill, Stabilization of proteins in confined spaces. *Biochemistry*. **40**,

11289–11293 (2001).

16. P. L. Clark, J. King, A Newly Synthesized, Ribosome-bound Polypeptide Chain Adopts Conformations Dissimilar from Early in Vitro Refolding Intermediates. *J. Biol. Chem.* **276**, 25411–25420 (2001).

17. P. L. Clark, Protein folding in the cell: reshaping the folding funnel. *Trends Biochem. Sci.* **29**, 527–534 (2004).

18. S. J. Kim *et al.*, Translational tuning optimizes nascent protein folding in cells. *Science (80-. ).* **348**, 444–448 (2015).

19. Y. Kimura, P. R. Bianco, Single molecule studies of DNA binding proteins using optical tweezers. *Analyst.* **131**, 868–874 (2006).

20. F. Gittes, C. F. Schmidt, Thermal noise limitations on micromechanical experiments. *Eur. Biophys. J.* **27**, 75–81 (1998).

21. S. B. Smith, Y. Cui, C. Bustamante, Overstretching B-DNA : The Elastic Response of Individual Double-Stranded and Single- Stranded DNA Molecules. *Science (80-. ).* **271**, 795–799 (1996).

22. K. Neupane *et al.*, Direct observation of transition paths during the folding of proteins and nucleic acids. *Science (80-. ).* **352**, 239–242 (2016).

23. J. C. M. Gebhardt, T. Bornschlögl, M. Rief, Full distance-resolved folding energy landscape of one single protein molecule. *Proc. Natl. Acad. Sci.* **107**, 2013–2018 (2010).

24. J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, C. Bustamante, Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's equality. *Science.* **296**, 1832–5 (2002).

25. C. Bustamante, J. F. Marko, E. D. Siggia, S. B. Smith, Entropic Elasticity of lambda Phage DNA. *Science (80-. ).* **265**, 1599–1600 (1994).

26. M. D. Wang, H. Yin, R. Landick, J. Gelles, S. M. Block, Stretching DNA with optical tweezers. *Biophys. J.* **72**, 1335–1346 (1997).

27. B. Jagannathan, S. Marqusee, Protein Folding and Unfolding under force. *Biopolymers.* **99**, 860–869 (2013).

28. P. J. Elms, J. D. Chodera, C. Bustamante, S. Marqusee, The molten globule state is unusually deformable under mechanical force. *Proc. Natl. Acad. Sci.* **109**, 3796–3801 (2012).

29. O. K. Dudko, G. Hummer, A. Szabo, Intrinsic rates and activation free energies from single-molecule pulling experiments. *Phys. Rev. Lett.* **96**, 1–4 (2006).

30. O. K. Dudko, G. Hummer, A. Szabo, Theory, analysis, and interpretation of single-molecule force spectroscopy experiments. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 15755–60 (2008).

31. Y. Zhang, O. K. Dudko, A transformation for the mechanical fingerprints of complex biomolecular interactions. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 16432–7 (2013).

32. P. J. Elms, J. D. Chodera, C. Bustamante, S. Marqusee, Limitations of constant-force-feedback experiments. *Biophys. J.* **103**, 1490–1499 (2012).

33. J. Stigler, M. Rief, Hidden Markov Analysis of Trajectories in Single-Molecule Experiments and the Effects of Missed Events. *ChemPhysChem.* **13**, 1079–1086 (2012).

34. S. A. Mckinney, C. Joo, T. Ha, Analysis of Single-Molecule FRET Trajectories Using Hidden Markov Modeling. *Biophys. J.* **91**, 1941–1951 (2006).

35. D. Swan, R. Hale, N. Dhillon, P. Leadlay, A bacterial calcium-binding protein

homologous to calmodulin. *Nature*. **329**, 84–85 (1987).

36. H. Tossavainen, P. Permi, A. Annila, I. Kilpelainen, T. Drakenberg, NMR solution structure of calerythrin, an EF-hand calcium-binding protein from Saccharopolyspora erythraea. *Eur. J. Biochem.* **270**, 2505–2512 (2003).

37. Z. Grabarek, Structural basis for diversity of the EF-hand calcium-binding proteins. *J. Mol. Biol.* **359**, 509–25 (2006).

38. N. Bylsma, T. Drakenberg, I. Andersson, P. F. Leadlay, S. Forsen, Prokaryotic calcium-binding protein of the calmodulin superfamily: Calcium binding to a Saccharopolyspora erythraea 20kDa protein. *FEBS Lett.* **299**, 44–47 (1992).

39. D. Chin, A. R. Means, Calmodulin: a prototypical calcium sensor. *Trends Cell Biol.* **10**, 322–8 (2000).

40. J. A. Cox, A. Bairoch, Sequence similarities in calcium-binding proteins. *Nature*. **331** (1988), p. 491.

41. J. Michiels, C. Xi, J. Verhaert, J. Vanderleyden, The functions of Ca 2+ in bacteria: a role for EF-hand proteins? *TRENDS Microbiol.* **10**, 87–93 (2002).

42. J. Stigler, F. Ziegler, A. Gieseke, J. C. M. Gebhardt, M. Rief, The complex folding network of single calmodulin molecules. *Science*. **334**, 512–6 (2011).

43. P. O. Heidarsson *et al.*, Single-molecule folding mechanism of an EF-hand neuronal calcium sensor. *Cell*. **21**, 1812–21 (2013).

44. P. O. Heidarsson *et al.*, Direct single-molecule observation of calcium-dependent misfolding in human neuronal calcium sensor-1. *Proc. Natl. Acad. Sci.* **111**, 13069–13074 (2014).

45. H. Aitio *et al.*, Characterization of apo and partially saturated states of calerythrin, an EF-hand protein from S. erythraea: a molten globule when deprived of Ca(2+). *Protein Sci.* **10**, 74–82 (2001).

46. D. Beckett, E. Kovaleva, P. J. Schatz, A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. *Protein Sci.* **8**, 921–929 (2008).

47. J. Yin *et al.*, Genetically encoded short peptide tag for versatile protein labeling by Sfp phosphopantetheinyl transferase. *Proc. Natl. Acad. Sci.* **102**, 15815–15820 (2005).

48. J. Stigler, M. Rief, Calcium-dependent folding of single calmodulin molecules. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17814–9 (2012).

49. B. Jagannathan, P. J. Elms, C. Bustamante, S. Marqusee, Direct observation of a force-induced switch in the anisotropic mechanical unfolding pathway of a protein. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17820–17825 (2012).

50. C. Cecconi, E. a Shank, C. Bustamante, S. Marqusee, Direct observation of the three-state folding of a single protein molecule. *Science*. **309**, 2057–60 (2005).

51. M. P. Landry, P. M. McCall, Z. Qi, Y. R. Chemla, Characterization of photoactivated singlet oxygen damage in single-molecule optical trap experiments. *Biophys. J.* **97**, 2128–2136 (2009).

52. M. J. Davies, Reactive species formed on proteins exposed to singlet oxygen. *Photochem. Photobiol. Sci.* **3**, 17–25 (2004).

53. H. Aitio *et al.*, NMR assignments, secondary structure, and global fold of calerythrin, an EF-hand calcium-binding protein from Saccharopolyspora erythrea. *Protein Sci.*, 2580–2588 (1999).

54. T. M. Lakowski *et al.*, Calcium-induced folding of a fragment of calmodulin composed of

Calcium-induced folding of a fragment of calmodulin composed of EF-hands 2 and 3. *Protein Sci.* **3**, 1119–1132 (2007).

55. J. F. Maune, C. B. Klee, K. Beckingham, Ca2+ binding and conformational change in two series of point mutations to the individual Ca2+-binding sites of calmodulin. *J. Biol. Chem.* **267**, 5286–5295 (1992).

56. C. N. Pace, J. M. Scholtz, A helix propensity scale based on experimental studies of peptides and proteins. *Biophys J.* **75**, 422–427 (1998).

57. T. Shi, S. Spain, D. L. Rabenstein, Unexpectedly Fast Cis / Trans Isomerization of Xaa-Pro Peptide Bonds in Disulfide-Constrained Cyclic Peptides. *JACS.* **126**, 790–796 (2004).

58. K. Nguyen, M. Iskandar, D. L. Rabenstein, Kinetics and Equilibria of Cis / Trans Isomerization of Secondary Amide Peptide Bonds in Linear and Cyclic Peptides. *J Phys Chem B.* **114**, 3387–3392 (2010).

59. L. Rognoni, T. Möst, G. Žoldák, M. Rief, Force-dependent isomerization kinetics of a highly conserved proline switch modulates the mechanosensing region of filamin. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 5568–73 (2014).

60. G. Fischer, T. Aumüller, Regulation of peptide bond cis/trans isomerization by enzyme catalysis and its implication in physiological processes. *Rev. Physiol. Biochem. Pharmacol.* **148**, 105–50 (2003).

61. J. R. Moffitt, Y. R. Chemla, D. Izhaky, C. Bustamante, Differential detection of dual traps improves the spatial resolution of optical tweezers. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9006–11 (2006).

62. C. Baumann, S. B. Smith, V. A. Bloomfield, C. Bustamante, Ionic effects on the elasticity of single DNA molecules. *Proc. Natl. Acad. Sci.* **94**, 6185–6190 (1997).

63. G. Yang *et al.*, Solid-state synthesis and mechanical unfolding of polymers of T4 lysozyme. *Proc. Natl. Acad. Sci.* **97**, 139–144 (2000).

64. T. Crombie, J. C. Swaffield, A. J. P. Brown, Protein folding within the cell is influenced by controlled rates of polypeptide elongation. *J. Mol. Biol.* **228**, 7–12 (1992).

65. A. A. Komar, T. Lesnik, C. Reiss, Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.* **462**, 387–391 (1999).

66. G. Zhang, M. Hubalewska, Z. Ignatova, Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.* **16**, 274–280 (2009).

67. A. J. Samelson, M. K. Jensen, R. A. Soto, J. H. D. Cate, S. Marqusee, Quantitative determination of ribosome nascent chain stability. *Proc. Natl. Acad. Sci.* **113**, 201610272 (2016).

68. K. Liu, J. E. Rehfus, E. Mattson, C. M. Kaiser, The ribosome destabilizes native and non-native structures in a nascent multidomain protein. *Protein Sci.* **26**, 1439–1451 (2017).

69. J. Frydman, H. Erdjument-Bromage, P. Tempst, F. U. Hartl, Co-translational domain folding as the structural basis for the rapid de novo folding of firefly luciferase. *Nat. Struct. Biol.* **6**, 697–705 (1999).

70. C. W. Kim *et al.*, N-terminal domains of native multidomain proteins have the potential to assist de novo folding of their downstream domains in vivo by acting as solubility enhancers. *Protein Sci.* **16**, 635–43 (2007).

71. D. H. Goldman *et al.*, Mechanical force releases nascent chain-mediated ribosome arrest in vitro and in vivo. *Science (80-. ).* **348**, 457–460 (2015).

72. W. Holtkamp *et al.*, Cotranslational protein folding on the ribosome monitored in real

time. *Science (80-. )*. **350**, 1104–1108 (2015).

73.  A. Khushoo, Z. Yang, A. E. Johnson, W. R. Skach, Ligand-Driven Vectorial Folding of Ribosome-Bound Human CFTR NBD1. *Mol. Cell*. **41**, 682–692 (2011).

74.  F. Wruck, A. Katranidis, K. H. Nierhaus, G. Büldt, M. Hegner, Translation and folding of single proteins in real time. *Proc. Natl. Acad. Sci.* **114**, E4399–E4407 (2017).

75.  M. C. Capece, G. U. Y. L. Kornberg, A. Petrov, J. D. Puglisi, A simple real-time assay for in vitro translation. *RNA*. **21**, 296–305 (2015).

76.  A. J. Samelson *et al.*, *Sci. Adv.*, in press, doi:10.1126/sciadv.aas9098.

77.  E. P. O'Brien, P. Ciryam, Understanding the Influence of Codon Translation Rates on Cotranslational Protein Folding. *Accounts Chem. …*, 1536–1544 (2014).

78.  E. P. O'Brien, J. Christodoulou, M. Vendruscolo, C. M. Dobson, New Scenarios of Protein Folding can Occur on the Ribosome. *J. Am. Chem. Soc.* **133**, 513–526 (2011).

79.  D. V Freistroffer, M. Kwiatkowski, R. H. Buckingham, M. Ehrenberg, The accuracy of codon recognition by polypeptide release factors. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 2046–2051 (2000).

80.  P. Ciryam, R. I. Morimoto, M. Vendruscolo, C. M. Dobson, E. P. O'Brien, In vivo translation rates can substantially delay the cotranslational folding of the Escherichia coli cytosolic proteome. *PNAS*. **110** (2012), doi:10.1073/pnas.1213624110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1213624110.

81.  O. B. Nilsson *et al.*, Cotranslational folding of spectrin domains via partially structured states. *Nat. Struct. Mol. Biol.*, 1–6 (2017).

82.  S. Bhushan *et al.*, alpha-Helical nascent polypeptide chains visualized within distinct regions of the ribosomal exit tunnel. *Nat. Struct. Mol. Biol.* **17**, 313–7 (2010).

83.  Y. Shimizu, T. Kanamori, T. Ueda, Protein synthesis by pure translation systems. *Methods*. **36**, 299–304 (2005).

84.  Y. Shimizu, T. Ueda, in *Methods in Molecular Biology* (2010), vol. 607, pp. 11–21.

85.  K. E. Splan, K. Musier-forsyth, M. T. Boniecki, S. A. Martinis, In vitro assays for the determination of aminoacyl-tRNA synthetase editing activity. *Methods*. **44**, 119–128 (2008).

86.  M. J. Comstock, T. Ha, Y. R. Chemla, Ultrahigh-resolution optical trap with single-fluorophore sensitivity. *Nat. Methods*. **8** (2011), doi:10.1038/NMETH.1574.

87.  J. Lu, C. Deutsch, Secondary structure formation of a transmembrane segment in Kv channels. *Biochemistry*. **44**, 8230–8243 (2005).

88.  J. B. Udgaonkar, Multiple routes and structural heterogeneity in protein folding. *Annu. Rev. Biophys.* **37**, 489–510 (2008).

89.  R. R. Brau, P. B. Tarsa, J. M. Ferrer, P. Lee, M. J. Lang, Interlaced Optical Force-Fluorescence Measurements for Single Molecule Biophysics. *Biophys. J.* **91** (2006), doi:10.1529/biophysj.106.082602.

90.  P. B. Tarsa *et al.*, Detecting Force-Induced Molecular Transitions with Fluorescence Resonant Energy Transfer **. *Angew. Chem. Int. Ed. Engl.*, 1999–2001 (2007).

91.  J. R. Lakowicz, in *Principles of Fluorescence Spectroscopy* (Springer US, Boston, 1999), pp. 367–394.

92.  E. Rhoades, E. Gussakovsky, G. Haran, Watching proteins fold one molecule at a time. *PNAS*. **100**, 3197–3202 (2003).

93.  M. A. Van Dijk, L. C. Kapitein, J. Van Mameren, C. F. Schmidt, E. J. G. Peterman,

Combining Optical Trapping and Single-Molecule Fluorescence Spectroscopy : Enhanced Photobleaching of Fluorophores. *J Phys Chem B*. **108**, 6479–6484 (2004).

94.   E. Florin, J. K. H. Hörber, E. H. K. Stelzer, High-resolution axial and lateral position sensing using two-photon excitation of fluorophores by a continuous-wave Nd : YAG laser. *Appl. Phys. Lett.* **69**, 446–448 (1996).

95.   M. J. Lang, P. M. Fordyce, A. M. Engh, K. C. Neuman, S. M. Block, Simultaneous , coincident optical trapping and single-molecule fluorescence. *Nat. Methods*. **1**, 1–7 (2004).

96.   K. D. Whitley, M. J. Comstock, Y. R. Chemla, in *Methods in Molecular Biology* (2017), vol. 1486.

97.   M. J. Comstock *et al.*, Direct observation of structure-function relationship in a nucleic acid – processing enzyme. *Science (80-. )*. **348**, 352–354 (2015).

98.   J. G. Yodh *et al.*, Asymmetric Unwrapping of Nucleosomes under Tension Directed by DNA Local Flexibility Article Asymmetric Unwrapping of Nucleosomes under Tension Directed by DNA Local Flexibility. *Cell*. **160**, 1135–1144 (2015).

99.   Y. S. Babu, C. E. Bugg, W. J. Cook, Structure of calmodulin refined at 2.2 A resolution. *J. Mol. Biol.* **204**, 191–204 (1988).

100.   M. Zhang, T. Tanaka, M. Ikura, Calcium-induced conformational transition revealed by the solution structure of apo calmodulin. *Nat. Struct. Mol. Biol.* **2**, 758–767 (1995).

101.   M. W. Allen *et al.*, Fluorescence labeling, purification, and immobilization of a double cysteine mutant calmodulin fusion protein for single-molecule experiments. *Anal. Biochem.* **325**, 273–284 (2004).

102.   T. Kantner, A. G. Watts, Characterization of Reactions between Water-Soluble Trialkylphosphines and Thiol Alkylating Reagents: Implications for Protein-Conjugation Reactions. *Bioconjug. Chem.* **27**, 2400–2406 (2016).

103.   Y. Kim *et al.*, Efficient Site-Specific Labeling of Proteins via Cysteines. *Bioconjug. Chem.* **19**, 786–791 (2008).

104.   A. Deiters, P. G. Schultz, In vivo incorporation of an alkyne into proteins in Escherichia coli. *Bioorg. Med. Chem. Lett.* **15**, 1521–1524 (2005).

105.   D. P. Nguyen, H. Lusic, H. Neumann, P. B. Kapadnis, A. Deiters, Genetic Encoding and Labeling of Aliphatic Azides and Alkynes in Recombinant Proteins via a Pyrrolysyl-tRNA Synthetase / tRNA CUA Pair and. *J Am Chem Soc*, 8720–8721 (2009).

106.   O. Tour, R. M. Meijer, D. A. Zacharias, S. R. Adams, R. Y. Tsien, Genetically targeted chromophore-assisted light inactivation. *Nat. Biotechnol.* **21**, 1505–1508 (2003).

107.   C. E. Aitken, R. A. Marshall, J. D. Puglisi, An Oxygen Scavenging System for Improvement of Dye Stability in Single-Molecule Fluorescence Experiments. *Biophys. J.* **94**, 1826–1835 (2008).

108.   L. L. Hildebrandt, S. Preus, V. Birkedal, Quantitative single molecule FRET efficiencies using TIRF microscopy. *Faraday Discuss.* **184**, 131–142 (2015).

109.   K. Hamadani *et al.*, An in vitro tag-and-modify protein sample generation method for single-molecule fluorescence resonance energy transfer. *J. Biol. Chem.* **292**, 15636–15648 (2017).

110.   M. Amiram *et al.*, Evolution of translation machinery in recoded bacteria enables multi-site incorporation of nonstandard amino acids. *Nat. Biotechnol.* **33**, 1272–1279 (2015).

111.   T. S. Young, I. Ahmad, J. A. Yin, P. G. Schultz, An Enhanced System for Unnatural

Amino Acid Mutagenesis in E. coli. *J. Mol. Biol.* **395**, 361–374 (2010).