Continuous time Markov models of the kinetics of protein folding and fluorescent protein blinking

by

Geoffrey C. Rollins

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

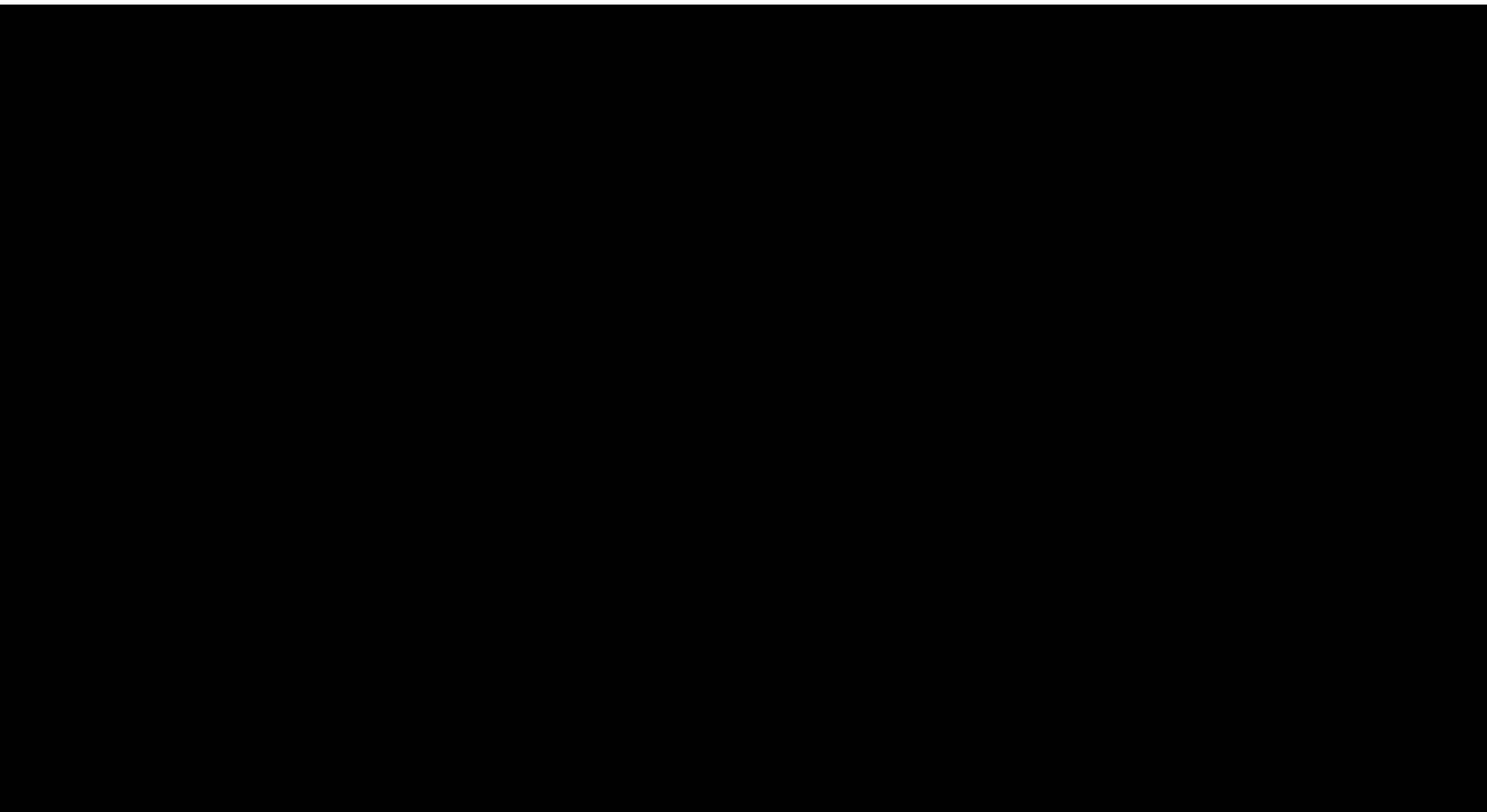
Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO



Acknowledgments

I owe a debt of thanks to many people. First and foremost, I would like to thank my family. The PhD process is like a marathon, but without them, I wouldn't have even gotten to the starting line. I would also like to thank Ken Dill for being a great advisor. His natural ability to communicate science has been an inspiration. I also want to acknowledge the former and current Dill lab members. Science is a social activity, and a significant part of the learning process has been the conversations that I've had with them along the way.

I would like to thank my thesis and orals committee members–Matt Jacobson, Susan Marqusee, Tanja Kortemme, Andrej Sali, and Jed Pitera–for their helpful feedback on my work. UCSF has been a great place to spend the past five years of my life. I've felt fortunate to be in a place that is full of so many creative and intelligent people. My classmates in the iPQB program are a great group of people, and they've been wonderful companions with whom to complete this phase of life. On the administrative side, I'd like to extend a big thank you to the program coordinators, Rebecca Brown and Julia Molla, for helping me navigate all of the administrative aspects of graduate school. Their clear and prompt communication style is something that I've greatly appreciated.

I had a lot of fun participating in UCSF's Science and Education Partnership (SEP). I would like to thank the folks who coordinate SEP, in particular Rebecca Smith, Ben Koo, and Sabine Jeske. SEP is a wonderful organization. I would also like to thank the SFUSD teachers who I worked with as part of SEP–Angie Patterson, Drew Tuomey, and Larry

Cohbra–for allowing me into their classrooms to teach and to observe. I learned a great deal about teaching from all of the interactions I had with them and with the people at SEP. Finally, I would like to thank the National Science Foundation Graduate Research Fellowship Program (NSF-GRFP) for supporting the research presented in this thesis.

- Chapter 2 of this thesis will be submitted to *Proceedings of the National Academy of Sciences USA*. K.A. Dill directed and supervised the research that forms the basis for chapter 2.

- Chapter 4 of this thesis is a manuscript in preparation for publication. S. Pressé supervised the research and helped develop the theoretical foundations for chapter 4. J.Y. Shin, S.H. Lee, and A. Lee gathered the experimental data analyzed in chapter 4, under the supervision of C. Bustamante.

Thesis advisor: Ken A. Dill Geoffrey C. Rollins

Continuous time Markov models of the kinetics of protein folding and fluorescent protein blinking

Abstract

We develop continuous time Markov models for a pair of biophysical problems. The first problem is the kinetic mechanism of protein folding. We develop a model that aims to explain the nine-order-of-magnitude dependence of folding rates on protein size and the predominance of two-state folding kinetics. In our model, secondary structures, which are intrinsically unstable in isolation, are stabilized and directed towards the native state by cooperative interactions with neighboring secondary structures along the folding routes. The model fits folding-rate data on a set of 82 proteins and can be applied to estimate the distribution of intrinsic folding rates for proteins in the proteomes of cells. The second problem is the analysis of fluorescent protein blinking in super-resolution microscopy. We develop an aggregated continuous time Markov model for quantifying the fluorescent proteins in a diffraction-limited volume. Using a maximum-likelihood approach, we apply the model to study the *in vitro* photophysics of the protein Dendra2 and to quantify the number of FliM molecules in bacterial flagellar motors. The end goal of the method is to count proteins in molecular assemblies with single molecule precision.

Contents

List of Tables

Listing of figures

Synopsis

This thesis is organized in the following way. Chapters 1 and 3 serve as introductions to the research problems addressed by chapters 2 and 4, which contain original work. Chapter 1 reviews the protein folding kinetics literature, both folding experiments and models. Chapter 2 contains a manuscript which will be submitted to *PNAS* that describes a model of the chain length dependence of folding kinetics and an application of the model to the folding kinetics of proteomes. Chapter 3 reviews aggregated Markov models for ion channel gating dynamics. Finally, chapter 4 contains a manuscript in preparation for publication that describes how the methods developed in the ion channel literature can be adapted to the analysis of fluorescent protein time traces collected by super-resolution microscopy for molecular counting in live cells.

1

Protein folding kinetics

Proteins are linear polymer chains of amino acids. Many proteins spontaneously fold into three-dimensional structures, and Anfinsen discovered that all the information required for folding is encoded in the amino acid sequence.[1] Ever since the first protein structures were solved in the late 1950s,[2,3] the folding field has sought to understand how amino acid sequences encode three-dimensional structures.

The focus of folding research over the past 50 years has been small globular proteins that are amenable to *in vitro* refolding experiments, atomistic simulations, and theoretical models. This chapter will focus on that literature, but it's important to also mention some of the additional complexities involved in *in vivo* folding, many of which have yet to be

fully elucidated.

- Over 60% of prokaryotic proteins and over 70% of eukaryotic proteins consist of multiple domains.[4,5]
- A significant fraction of proteomes are intrinsically disordered, and disorder appears to be functionally important in many cases.[6]
- Many proteins are membrane-bound. They require a membrane environment to fold correctly, and they are critical drug targets.[7,8]
- Chaperones are critical for proper folding in the crowded environment of living cells,[9,10] and the failure of proteostatic mechanisms can lead to protein aggregates that are implicated in many diseases.[11–13]

Our understanding of folding in the cell is certain to develop in the next 50 years, thanks in part to the advances in basic folding physics reviewed in this chapter.

1.1 Driving forces, pathways, and funnels

For many small proteins, folding is a thermodynamically two-state process,[14] represented by $U \rightleftarrows F$, where U is the unfolded state and F is the folded state. Under folding conditions, the folded state is a global free energy minimum for the protein chain and its solvent. The primary (non-covalent) molecular driving forces[15–17] (Fig. 1.1.1) involved in folding are:

1. **the hydrophobic effect**. The hydrophobic effect refers to the burial of nonpolar amino acids in the protein core. Near $25°$C, this burial process is favorable because it leads to an increase in solvent entropy. The hydrophobic effect is the dominant driving force of folding.
2. **hydrogen bonds**. Hydrogen bonding is a polar interaction, primarily between amide groups in the protein backbone. Proteins form regular local structures,

called secondary structures, to satisfy steric constraints and to ensure that all hydrogen bonds are satisfied in the folded state. In the unfolded state, hydrogen bonds are satisfied by interactions with water, so failing to make a hydrogen bond in the folded state would result in a significant energetic penalty.

3. **salt bridges**. Salt bridges are electrostatic and hydrogen bonding interactions between positive (Lys, Arg, His) and negative (Asp, Glu) amino acid side chains. Some of the earliest models of protein stability in the 1930s predicted that salt bridges would be the main driving force of folding, but that view was overturned by further work on hydrophobic interactions.[15]

4. **conformational entropy**. The conformational entropy of the protein chain is the main force that opposes folding.

The balance between these forces leads to a net stability of roughly 5-15 kcal/mol with a weak dependence on protein length.[18]

Folding is a disorder-to-order transition, in which a heterogeneous ensemble of unfolded conformations converges to a unique folded structure. It's more similar to a phase transition, like liquid→solid, than it is to a simple chemical reaction.[20] Folding kinetics is about how proteins accomplish this disorder-to-order transition within a biologically-relevant timescale. A key question is "how do proteins fold so quickly?" This issue was clearly articulated by Cyrus Levinthal[21] who posed the problem in the following way: if proteins folded by randomly sampling conformational space, then folding would require timescales on the order of the age of the universe. If we estimate that there are $z = 3$ rotational isomers around each peptide bond, and if there are $N = 100$ peptide bonds in a chain, then the number of conformations the protein must search is $z^N \approx 10^{50}$. Even assuming a rapid sampling rate, it would take the age of the universe to randomly search through such a vast space of conformations. Instead, proteins fold on biological timescales, so the search for the folded state must be guided,

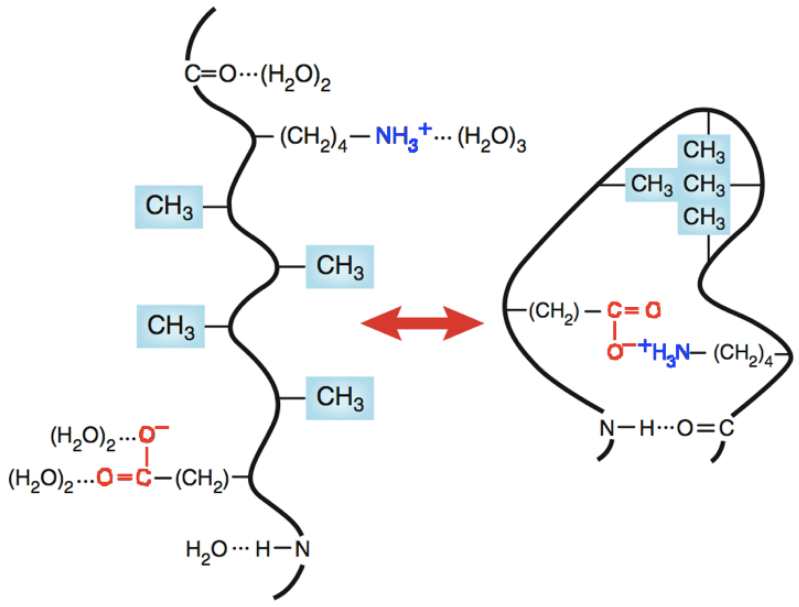


Figure 1.1.1: Driving forces in protein folding. The methyl groups represent the hydrophobic effect. The CO and NH groups represent the formation of backbone hydrogen bonds. The Asp (red) and Lys (blue) side chains represent salt bridges. Conformational entropy of the protein chain is the primary force that opposes these stabilizing forces. Adapted by permission from Macmillan Publishers Ltd: *Nature Structural and Molecular Biology* [19], copyright (2009). See page 98 for license information.

not totally random.

Inspired by Levinthal, early models were based on the idea that proteins folded through a pathway of intermediate conformations that guided the protein chain to the folded state, and early kinetics experiments focused on searching for and characterizing these intermediates.[22–29] However, ideas about pathways are macroscopic ideas based on macroscopic experiments; they don't explain the microscopic physics of folding. They don't explain how an ensemble of unfolded chains, each following a separate trajectory through conformational space, can all converge to the same folded structure. The microscopic perspective was addressed by the "new view" of folding.[30]

The aim of the new view is to connect the microscopics of folding with macroscopic observations. It's rooted in statistical mechanical models, and it's defined in terms of folding funnels and energy landscapes, rather than pathways.[31,32] The idea of an energy

landscape is a very general one, not specific to folding.[20,33] It's simply a description of the free energy of a system with respect to its degrees of freedom. In the case of folding, the energy landscape is the free energy of the protein and its solvent with respect to the degrees of freedom of the protein chain, often represented by the backbone dihedral angles. The folding funnel idea is that, in order to overcome Levinthal's Paradox, the energy landscape should have a shape that guides the vast ensemble of unfolded conformations towards the folded state.[31] Additionally, the energy landscape should be smooth with no major kinetic traps along the way (minimal frustration).[34,35]

A common misconception in the literature is that the classical pathway view is inconsistent with or contradictory to the new view of folding (in Krishna & Englander,[36] for example). This is incorrect. The two views operate at different levels of resolution. The classical view is about what we observe in bulk-averaged experiments; the new view is about how the microscopic physics of folding gives rise to those bulk-averaged observations. Energy landscapes can certainly be shaped in such a way that they give rise to folding behavior that appears pathway-like on the macroscopic level. To say that one view is right and the other view is wrong would be akin to saying this: reading the temperature of a glass of water is right, but a model for how a macroscopic property like temperature arises from the wiggling and jiggling of many water molecules on the atomic scale is wrong. They are both valid. The one you choose simply depends on your goal: a phenomenological model of macroscopic behavior or a microscopic model that explains where the macroscopic behavior comes from. This is precisely the goal of folding kinetics research: how do we connect macroscopic data to the microscopic trajectories of individual proteins and build a general understanding of folding that is consistent with both?

1.2 How do we observe folding kinetics?

1.2.1 Relaxation experiments

The most common folding kinetics experiments perturb the folding environment and then monitor the relaxation back to equilibrium.[37] The most common perturbations are changes in temperature, pH, or concentrations of chemical denaturants, like urea or GdmCl. The perturbations can be introduced in a variety of ways. Some of the most common ways are rapid mixing, lasers, and electrostatic discharges. Another method for perturbing the folding equilibrium is force, introduced by atomic force microscopy or optical tweezers.[38] The relaxation to equilibrium is measured by optical methods, such as circular dichroism (CD), fluorescence, or infrared (IR) spectroscopy. CD and IR report on changes in secondary structure, whereas fluorescence is primarily used to measure changes in the local environment of tryptophan residues–burial in the protein core leads to a change in fluorescence. The output of such experiments is signal versus time, which is typically well fit by a function of one or two exponentials (Fig. 1.2.1A).

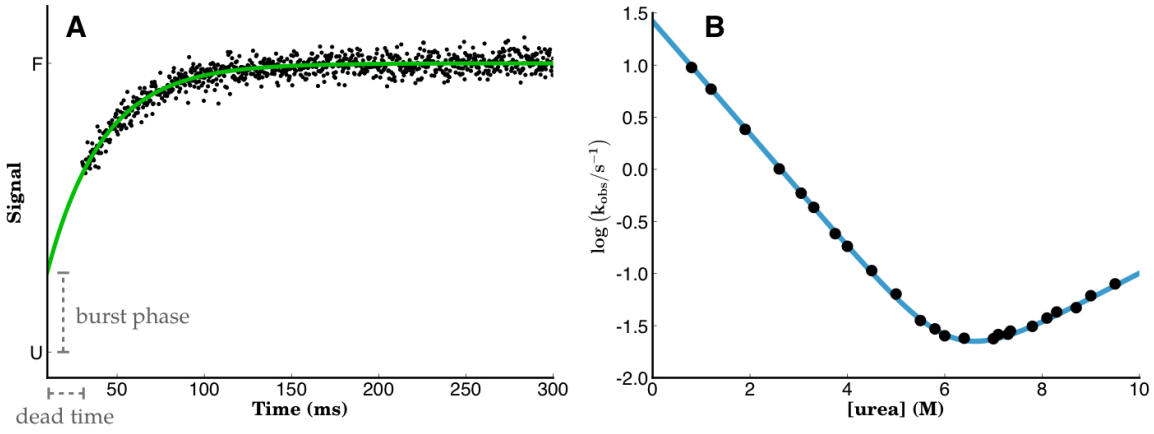


Figure 1.2.1: (A) relaxation of an optical signal over time from a refolding experiment, simulated data. The burst phase is a rapid change in signal that occurs during the dead time of the mixing instrument; (B) a chevron plot from a series of refolding mearsurements at various denaturant concentrations, based on CTL9 parameters from Maxwell et al.[39]

Folding and unfolding rates in water are determined by extrapolating observed relaxation rates to zero denaturant. This type of experiment is summarized in a chevron plot, so named because of the characteristic V-shape of the data (Fig. 1.2.1B). On the left side of the plot (low denaturant), the observed relaxation rate (k_{obs}) is dominated by the folding rate. On the right side of the plot (high denaturant), k_{obs} is dominated by the unfolding rate. Extrapolating these two linear regions of the curve back to zero denaturant gives the folding rate and unfolding rate in water. In some cases, non-linear chevrons are observed (rollover), and this is usually attributed to the presence of folding intermediates.

1.2.2 Phi-value analysis

Fersht pioneered the use of protein mutagenesis to study folding.[40–42] The common name for mutational studies of folding is "phi-value analysis". The basic experiment is to make a minimally disruptive point mutation and characterize the variant protein using the same relaxation-type experiments used to study the wild-type protein. Phi (equation 1.1) is a way of quantifying the folding behavior of the variant protein relative to the wild-type protein. It represents the change in transition state free energy over the change in net protein stability. The change in transition state free energy (the numerator) is computed from the relative folding rates of the wild type and variant proteins, and the change in net stability (the denominator) is computed from equilibrium denaturation experiments.

$$\phi = \frac{\Delta\Delta G^{\ddagger}}{\Delta\Delta G_{eq}} \tag{1.1}$$

Since the original Fersht studies, phi-values have been collected for many other proteins,[43–54] and they've been applied to protein design problems[55] and used as benchmarks for folding kinetics models.[56–59] Phi is inherently an energetic quantity, but it

has often been interpreted in structural terms. A phi-value of 0 is interpreted as disruption of interactions that are present in the folded state, but not in the transition state. A phi-value of 1 means that the disrupted interactions were present in both the folded state and the transition state. Most phi-values are fractional, which is attributed to partial interactions or interactions present in only some structures in the transition state ensemble. The correct structural interpretation of phi-values is still an open question.[60,61]

1.2.3 Hydrogen exchange experiments

Another major branch of folding kinetics experiments is hydrogen exchange (HX).[26,36,62–71] Amide hydrogens in the protein backbone exchange naturally with hydrogens in water. H-bonded structures don't exchange with solvent, so HX provides a residue-by-residue readout on structure formation. It's possible to control the rate of exchange of unprotected amides by adjusting pH and temperature.

One type of experiment is pulse-labeling.[25–27,62,70] It starts with an unfolded protein in D_2O, fully exchanged with deuterium. The sample is rapidly diluted into H_2O at low denaturant. Refolding is allowed to proceed for a short time, and then the sample is briefly mixed into high pH, which promotes rapid HX. This HX pulse selectively labels unstructured regions of the protein backbone. The exchange process is then halted by transfer into slow HX conditions and refolding goes to completion. The slow HX conditions preserve the labeling from the high pH pulse so that they can be analyzed by NMR[26,62–66] or mass spectrometry (MS).[72,73]

Another type of HX experiment is native state hydrogen exchange (NHX).[65,70] These experiments are done at equilibrium as a function of denaturant, and they reveal local unfolding transitions in addition to the global unfolding transition of the entire protein. These local regions of structure that undergo cooperative subglobal unfolding transitions have been named "foldons".[66,67,70]

1.2.4 Single molecule methods

The recent development of single molecule methods is expanding our understanding of folding.[38,74–76] One approach is single molecule FRET (smFRET), which allows distances in the 2 to 10 nm range to be probed.[74,77,78] Conformational changes in the protein result in measurable changes in energy transfer between the donor and acceptor FRET probes. One recent success of smFRET was a study that measured the mean transition path times of two small proteins.[78] Another single molecule approach is to apply force to single molecules through AFM[79] or optical tweezers.[80,81] AFM can apply high forces (up to nN), but it has a high spring constant that makes it less suitable for observing refolding. On the other hand, optical traps are well suited to refolding studies because they are sensitive in the low force regime (below 50 pN).[76] A recent optical tweezer study on the two-domain protein calmodulin discovered a complex folding network with at least six states.[82] Another recent study found that src SH3 is more resistant to force applied along the axis of its terminal β-strand, compared to force applied perpendicular the strand.[81] Optical tweezers have also been used to apply force to circular permutants[83] and molten globules.[80] It's clear that both smFRET and force-based single molecule methods are providing new avenues for probing protein folding, and they will likely be the driving force behind advances in the field for years to come.

In summary, folding has been studied by a large community of investigators using an extensive set of experimental methods and tools. We've reviewed some of the most important methods above. Now, we turn our attention to the similarly diverse and wide-ranging efforts to interpret and understand experimental data through models of folding kinetics.

1.3 Models of folding kinetics

This section will be a broad survey of models of folding kinetics. We start with the most macroscopic classes of models: descriptive models, mass action models, and linear regression models. Then, we zoom down to the most microscopic folding models, the all-atom models. Finally, we discuss models of folding at the coarse-grained level. Some coarse-grained models are lattice-based or use a simplified off-lattice chain, while others are more abstract and do away with an explicit chain representation entirely (we call these "simple models"). We shall discuss both types.

1.3.1 Descriptive models

A lot of ideas about folding mechanisms in the literature are descriptive models. Examples include nucleation-condensation,[84,85] hydrophobic collapse,[86] the framework model,[25,27] and the Foldon model.[66,67,70] Another model is diffusion-collision; it's different than the other descriptive models discussed here because it was proposed as both a description of folding and as a quantitative model.[87,88] We'll return to it in the discussion of simple models below. Descriptive models help us make sense of experimental results in a qualitative way, but alone they can't help us make quantitative predictions about future experiments.

1.3.2 Mass action models

The kinetic law of mass action relates reaction rates of elementary reactions to reactant concentrations and stoichiometries.[89] The simplest mass action model of folding is the two-state reversible reaction between U and F. For some proteins, the data is better fit by a sum of exponentials. In these cases, the corresponding mass action models include intermediate I states (Fig. 1.3.1). The intermediate states can be on-pathway or off-pathway, as in the case of proline cis-trans isomerization.[22,23,25,27] States with more

than three states have also been proposed.[36,90–92] Mass action models tell us how many macroscopic states are present, but they can't provide microscopic insights. They can tell us that U and F and sometimes I exist, but they can't tell us the structures, the underlying physics, nor how folding is affected by solution conditions like salt, pH, and temperature.

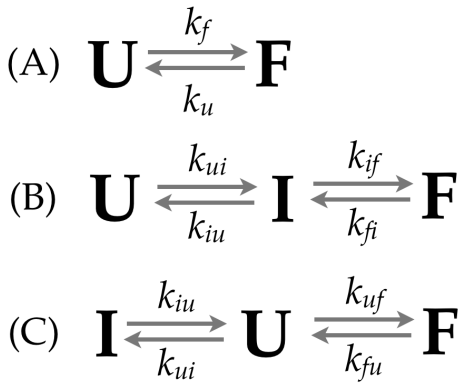


Figure 1.3.1: (A) two-state kinetics (B) three-state kinetics with an "on-pathway" intermediate, (C) three-state-kinetics with an "off-pathway" intermediate.

1.3.3 Linear regression models

A key insight in folding kinetics was the observation by Plaxco, Simons, and Baker that folding rates are correlated with contact order, a measure of protein topology.[93] This result inspired a wave of linear statistical models that fit protein metrics, like topology, chain length, surface area, and volume to folding rate data.[94–107] One of the major insights has been that chain length is a better predictor than topology alone, and chain length combined with topology is better still.[96–98] A primary concern with these statistical models is overfitting. Metrics like R^2 will only improve with the addition of more parameters, which is why they must be paired with an F-test or an information criterion. Statistical curve fit models are useful hypothesis generators, but alone they don't tell us about the microscopic physics of folding.

1.3.4 All-atom models

Unlike the classes of models discussed above, atomistic models are very detailed at the microscopic level. Atomic interactions are encoded in an energy function, also called a force field or a potential. Biomolecular simulations are based on molecular mechanics (MM) force fields, such as CHARMM, AMBER, and OPLS.[108–115] MM force fields represent atoms as rigid spheres connected by springs. The springs capture the stretching, bending, and torsional modes of covalently bonded structures. There are also non-bonded terms in the force field that capture coulombic interactions and van der Waals forces. Hydrogen bonds are not encoded by specific energy terms; they arise due to the electrostatic interactions. Water is an important part of biology, and so MM force fields always need to be paired with accurate water models.[116] Water can be modeled explicitly with individual water molecules added to the simulation box, or implicitly with an energy function that mimics the effects of water.[115,117] Explicit water models are more accurate, but they are also more expensive to compute. On the other hand, implicit models are faster to compute, but they lack the accuracy of explicit water.

Molecular Dynamics (MD) simulations take a force field and the atomic coordinates of molecules and advance them forward in time by numerically integrating Newton's equations of motion.[118] In principle, if we have an accurate energy function, all-atom MD simulations have the power to connect macroscopic folding observables with detailed molecular mechanisms. However, in practice, we run into trouble because atomistic simulations are expensive to compute. The reason is that integrating Newton's equations requires a step size on par with the fastest timescale in the system. In the case of biomolecules, bond vibrations on the femtosecond timescale are the fastest motions, so a time step of 1 or 2 femtoseconds is necessary to ensure numerical stability of the calculation. Another challenge is that the force field calculation at each time step requires substantial computation: roughly one billion arithmetic operations for a system with one hundred thousand atoms.[119]

Despite these challenges, MD simulations have been developing at a faster than Moore's law pace in recent years.[119] In 2000, Duan & Kollman completed the first microsecond simulation.[120] They simulated the villin headpiece from an extended conformation to a partially folded one. Since then, the development of crowd-sourced computational efforts, like Folding@Home,[121] and powerful compute clusters–like IBM's Blue Gene[122] early on and more recently the Shaw group's ANTON cluster[123–125]–have continually pushed the boundaries of simulation since the Duan & Kollman study. Impressively, there are now several reports of millisecond folding simulations, a three order-of-magnitude jump in simulation timescale in only a decade.[125–127]

All-atom simulations are challenging to interpret. The raw data they produce is quite different from the raw data that experimentalists gather: atomic coordinates as a function of time rather than a bulk-averaged optical signal or a patterns of amide HX rates. An important goal in the MD simulation field has been to connect all-atom trajectories to experimental data. Markov State Models (MSMs) have been developed in recent years with this goal in mind.[127–134] MSMs are methods for clustering the atomic structures visited by large data sets of short trajectories into a mesoscopic description of the dynamics. The end result of MSM analysis is a set of discrete states with rates that govern hops between kinetically connected states. As an aside, the idea of a model or process being Markovian is a very general concept. It refers to the fact that the probablility of transitioning from one state to another depends only on the current state, not on previously occupied states.[135] Many of the coarse-grained models discussed below are also based on Markovian assumptions.

1.3.5 Coarse-grained polymer models

Coarse-grained polymer models use simplified energy functions and simplified chain representations to circumvent the sampling challenges associated with all-atom models. Early on, lattice models with native-centric energy functions (Gō potentials) were

developed.[136] Later, in the 80s and early 90s, hydrophobic-polar (HP) energy functions were developed, in which chain residues are either hydrophobic or polar.[137–139] HP models mimic the hydrophobic burial process through favorable HH contacts. Lattice models lack the detailed microscopic physics of all-atom models, but they have some of the key ingredients in folding: chain connectivity, flexibility, excluded volume, and sequence-dependent interactions. For that reason and for reasons of computational efficiency, lattice models have been critical for developing an understanding of the general principles of folding thermodynamics and kinetics.[34,140]

Off-lattice models with native-centric potentials have also been the focus of numerous studies.[141–149] In the literature, these are often referred to as "Gō models". Interestingly, even though Gō models are native-centric, they do not correctly capture the cooperativity of folding (quantified by the ratio of the van't Hoff enthalpy and the calorimetric enthalpy, $\Delta H_{vH}/\Delta H_{cal}$). Another problem is that experimental folding rates span nine orders of magnitude, but Gō models span only two or three orders. However, Kaya & Chan found that native-centric energy functions augmented with desolvation barriers can better reproduce the many orders of magnitude variation observed in folding rates.[147] Also, a recent statistical mechanical model of helix-bundle folding found that nonlinear coupling between tertiary interactions and helical interactions can predict correct cooperativities.[150]

1.3.6 Simple models

We use the term "simple model" here to refer to models which contain microscopic physics, but do not include an explicit representation of the protein chain. Simple models are more abstract then their chain-based counterparts, but they are useful companions to more detailed models. They can be used to help benchmark and guide the development of more detailed models. A good example comes from helix-coil theory: Zimm-Bragg[151] and Lifson-Roig[152] models not only provide an intuitive picture of helix-coil transitions,

but they are also useful for interpreting experimental data[153] and for evaluating the performance of all-atom force fields in predicting the helicities of peptides.[154] The beauty of ZB and LR models is that, even though they lack an explicit representation of chain dynamics and molecular interactions, they help us understand helix-coil transitions in a quantitative way, in terms of a small number of physically-motivated parameters. Inspired by the success of models like Zimm-Bragg theory, a number of investigators have tried to build simple models for protein folding.

One of the earliest simple models of folding was the aforementioned diffusion-collision model (DC) model of Weaver and Karplus.[87,88] The DC model represents folding as the diffusion of spherical microdomains connected by flexible linkers, which collide and coalesce. The probability of collision is determined by geometric parameters that describe the sizes of the folding units, and the probability of a pair of folding units sticking together upon collision is determined by helical propensities. The DC model has been successful in predicting the folding rates of several small proteins to within the correct order of magnitude,[155–158] but it's reliance on geometric parameters and helical propensities, which aren't always known, has prevented it from being applied more broadly.

Zwanzig, Szabo, and Bagchi (ZSB) pioneered the development of Ising-like models during the 90's.[159,160] The original ZSB model was focused on solving Levinthal's Paradox, and they found that a small energy bias towards native-like interactions, of the order of a few kT, can reduce the folding time from an astronomical timescale to a biological timescale. Extensions of the ZSB model sought more detailed representations of the energetics of folding based on (1) contact formation,[57,161–164] (2) burial of surface area and Rosetta scoring,[56,59] or (3) loop entropy.[58,165] These models have been used to predict folding rates and also phi-values. More recently, Muñoz and collaboraters developed a mean-field 1D free energy surface model,[166,167] based on a continuous analog of the Zwanzig nativeness reaction coordinate. Their model partitions the

energetics of folding into local and non-local components. They used their model to predict folding and unfolding rates based on structural class and chain length.

DC and Ising-like models are similar in the sense that they both have folding units (residues or secondary structures) that are represented in a binary way: each unit can be either folded or unfolded (originally described as "correct" or "incorrect" by Zwanzig et al.[159]). The kinetic representations are also similar: both models are described by Markovian processes on a finite number of discrete states. But this is not the only way that folding kinetics has been modeled. The aforementioned Muñoz mean-field model is one example. Another example, is the model of Bicout and Szabo. They represent folding as the diffusion of a particle within a sphere.[168] The center of the sphere represents the folded state, and the particle's search for the center of the sphere mirrors a protein's search for its folded state. The search is guided by a spherically symmetric potential that favors the folded state.

Another alternative approach is to study Gaussian chain models that have analytical solutions. One example is the topomer search model of Makarov and Plaxco.[169] In their model, the rate-limiting step to folding is the Gaussian chain's search for a native-like conformer, from which native contacts can rapidly nucleate. The topomer search model has been called into question, though, by Wallin & Chan who found that an unbiased search for the native topomer would require timescales that are much longer than the actual timescales of folding, essentially the Levinthal problem.[170] In a newer Gaussian chain model, Rustad and Ghosh found that adding springs between native contacts in the transition state leads to folding rate predictions that agree well with experimental folding rates.[106]

A final class of simple models are those whose folding units are residue-level contacts. A model by Weikl and Dill describes folding in terms of clusters of native contacts on a contact map.[171,172] They applied their contact model to phi-values[60] and circular permutant proteins.[173] A contact-based model published last year from Lane & Pande is

notable for its inclusion of non-native contacts.[174] They found hub-like behavior and folding rates that scale linearly with chain length.

1.4 Summary and outlook

Ultimately, we want models that can be compared directly with experimental data, like phi-values, chevron plots, and HX patterns. But we also want microscopic insights about folding. To accomplish both of these goals, we will need a multi-scale perspective, a toolbox of models that spans the length and timescales of folding. Atomistic models tell us about the microscopic details, but they don't directly give us macroscopic insights. Mass action models give us a macroscopic fit to data, but they lack microscopic insights. Connecting the macro world to the micro world is a long-standing challenge that continues to this day.

What we're still missing is a general folding mechanism. A general folding mechanism would be a universal statement about how proteins fold that accounts for their speed, the dependence of speed on protein length, and the dependence of speed on environmental variables, such as temperature and denaturants. It would be a comparative concept that explains why one protein folds differently than another. It would explain differences and similarities of the folding routes and rates of different proteins in advance of experimental data. It would be quantitative, rather than qualitative. It wouldn't just be about a specific protein, it would tell us something about whole proteomes.

In the next chapter, we present a simple model of the chain length dependence of folding, which builds on the Zwanzig model discussed above.[159,160] The model is defined in terms of secondary structures that are intrinsically unstable in isolation, but are stabilized and directed towards the folded state by cooperative interactions with their neighbors. The model is simple enough that it can be applied to efforts to characterize whole proteomes.

2

The Staircase-Landscape mechanism of the folding kinetics of proteomes

This chapter contains a manuscript in preparation for publication.

Abstract

We develop a model for the kinetic mechanism of protein folding. It aims to explain the nine-order-of-magnitude dependence of folding rates on protein size and the predominance of two-state folding kinetics. In our model, secondary structures, which are intrinsically unstable in isolation, are stabilized and directed towards the native state by cooperative interactions with neighboring secondary structures along the folding routes. The model energy landscape is shaped like an Up-Staircase with a cliff at the end, so the dynamics has *nested transition states:* each higher-order structure is a transition state for the preceding structure. The model fits folding-rate data on a set of 93 proteins. A main endpoint of this work is to estimate the distribution of intrinsic folding rates for proteins in the proteomes of cells. The resultant distribution shows that most proteins in E coli fold over a time scale of 10 msec - 10 sec.

2.1 Introduction

There has been much interest in the kinetic mechanism of how proteins fold. One motivation has been that if the kinetic routes of folding were known, the resultant insights might inform algorithms that aim to predict protein structures from their amino acid sequences, possibly speeding up drug discovery. Our interest here, however, is based on a different motivation. We want to be able to compute folding rates across whole proteomes of cells, for the purpose of understanding how folding and aggregation equilibria and kinetics mediate cellular health and diseases. By mechanism, we mean a general description of the kinetic process of folding across different proteins, not just the sequences of folding events in any one particular protein. We want a physical model that accounts for why small proteins tend to fold so much faster than larger proteins do. An ultimate folding mechanism should explain both the universal features of how folding rates depend on protein size and also the particular features of how folding rates depend on amino acid sequences. The present work is aimed only at the former.

Our modeling draws its basic insights from a large body of prior work. First, it is known that proteins fold kinetically through the rapid formation and assembly of secondary structures.[25,27,68,70,87,88,175,176] Second, Plaxco et al, had the pioneering insight that a protein's folding rate depends on properties that are evident from its native structure.[93] They found that helical proteins tend to fold faster than β-sheet proteins, and in general, that local structures tend to form faster than nonlocal ones. Such folding rate data has also been fitted by a number of other statistical models,[94–107] particularly showing a strong dependence of folding rate on a protein's chain length.[96–98,177] In a more detailed discussion in SI, we describe the current consensus[101] that folding rates are now better correlated with a protein's chain length L and with absolute contact order (ACO) than they are with other metrics, like relative contact order (RCO), that only consider the topology of a protein's native structure (Fig. S1, Table S1). Third, Zwanzig, Szabo &

Bagchi (ZSB) pioneered the development of a simple Ising-like model showing how the funnel shapes of energy landscapes lead to fast folding.[159,160] Muñoz, Eaton, Baker, Finkelstein, and others[56–59,162–165,167,178–181] have further developed and applied the ZSB approach, adding more detailed residue-level information in the form of contacts, hydrogen bonds, buried surface area, and loop entropies. Fourth, previous work has shown that equilibrium protein folding cooperativities can be explained as a combination of weak propensities of peptide chains to form secondary structures and stronger propensities of tertiary interactions to stabilize the secondary structures.[150]

Here, we combine the threads above in order to explain the experimental folding rates across proteins in terms of secondary structure assembly. Our model is an adaptation of Zwanzig's Ising-like model,[159,160] but we model folding at the secondary structure level, rather the amino acid level. Like Karplus and Weaver's diffusion-collision model,[87,88,155–158] our model allows for marginally stable secondary structures, but like the original Zwanzig model, our aim has been to keep the physics as simple as possible. Our model does not require prior knowledge of native topologies, structural propensities, or native geometric details.

2.2 Model

We express a protein's folding equilibrium and kinetics adapted from the Ising-like approach of Zwanzig.[159,160] We represent a chain of N secondary structures as a 1-D string of symbols

`fffuffuufuffff` ...

where the `f`'s indicate that a particular secondary structure is in its folded native-like conformation, and the `u`'s indicate that a particular secondary structure is in an unfolded non-native conformation. Let c represent the number of `f`'s or "correct" secondary structures in the string. $c = N$ represents the folded native state; $c = N - 1$ describes the

state in which the protein is native in all but one of its secondary structures, so there is one u somewhere in the string; and $c = 0$ means that the molecule has no native pieces of structure. c represents a simple 1-D "reaction coordinate" for folding. In this model, secondary structures form concomitantly with tertiary structures. When a secondary structure switches from `u` $\rightarrow$ `f`, it forms tertiary interactions with any other `f`'s already in place. We do not restrict the growth of folded segments to one or two contiguous regions (the single or double sequence approximation[57]). For example, after the first secondary structure forms at one of N locations in the protein, the next secondary structure may form at any of the other $N - 1$ locations (Fig. 2.3.2).

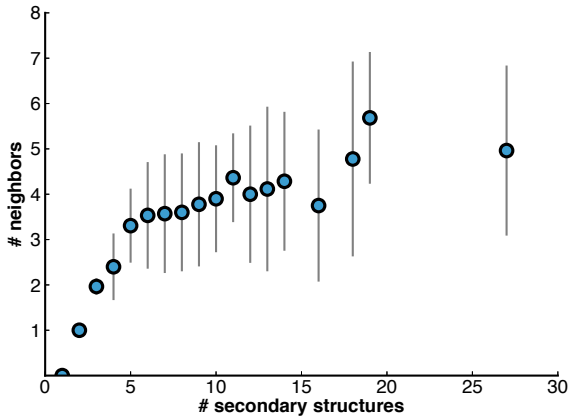


Figure 2.2.1: Nearest-neighbors of a secondary structure in a native protein vs. the total number of secondary structures in that protein. A pair of secondary structures was taken to be neighbors if they had at least 1 residue-residue contact. Residue contacts were determined from a centroid for each residue with a cutoff of 8 Å. The circles represent mean values, the error bars represent standard deviations. The plot is based on the 93 proteins in our data set.

2.2.1 Thermodynamics of the model

The Boltzmann weight of any protein configuration having c correct secondary structures is given by

$$w(c) = \frac{N!}{c!(N-c)!} K_2^c K_3^{n_c} \tag{2.1}$$

The combinatorial factor in equation 2.1 counts the number of ways that c `f`'s and $N - c$ u's can be arranged in a 1-D string. All the possible combinations for $N = 4$ are shown in Fig. 2.3.2. K_2 is the equilibrium constant for forming a secondary structure element from an unformed chain and K_3 is the equilibrium constant for forming a bundle and corresponding contacts between two isolated secondary structures. The microscopic basis for K_2 is the same as in traditional helix-coil theory: hydrogen bonds stabilize secondary structures and local chain entropy opposes them. Similarly K_3 arises from contact interactions among pairs of secondary structures, and includes hydrophobic, steric and hydrogen bonding interactions. As we discuss below, we find that the best fits to data are when $K_2 < 1$ (it is net unfavorable to form isolated secondary structures), and when $K_3 > 1$ (tertiary interactions are net favorable). The tertiary interactions help stabilize the secondary structures.

We count each pair of interacting secondary structures as one tertiary interaction, and n_c represents the total number of tertiary interactions. It is defined as a discrete function of c:

$$n_c = \begin{cases} 0 & \text{if } c = 0, 1 \\ 1 & \text{if } c = 2 \\ 3 & \text{if } c = 3 \\ 4c - 10 & \text{if } c > 3 \end{cases}$$

This definition of n_c derives from observing in native protein structures that the number of nearest neighbors per secondary structure grows with protein size and saturates at a maximum of about 4-5; see Fig. 2.2.1.

n_c is defined by the combinatorics of nearest-neighbor interactions. When there are zero or one secondary structures ($c = 0$ or 1), there can be no tertiary interactions. When there are two secondary structures ($c = 2$), they have $n_c = 1$ tertiary interaction between

them. When there are three secondary structures ($c = 3$), say A, B and C, there are $n_c = 3$ pairwise tertiary interactions between them, AB, AC and BC. For $c > 3$, each additional secondary structure gains four tertiary neighbor interactions upon folding because this is approximately the maximum that is sterically possible; see Fig. 2.2.1.

The equilibrium probability that a protein has c correct secondary structures is given by

$$p_{eq}(c) = \frac{w(c)}{\sum_{i=0}^{N} w(i)} = \frac{w(c)}{Q} \tag{2.2}$$

Q is the partition function, the sum of the Boltzmann weights of all states.

$$\begin{aligned} Q = 1 + NK_2 + \binom{N}{2} K_2^2 K_3 + \binom{N}{3} K_2^3 K_3^3 + \ldots \\ + NK_2^{N-1} K_3^{n_{N-1}} + K_2^N K_3^{n_N} K_f \end{aligned} \tag{2.3}$$

The weight 1 accounts for the fully unfolded protein, the weight NK_2 accounts for the formation of any of the N individual secondary structures, the term $NK_2^{N-1}K_3^{n_{N-1}}$ accounts for the formation of all but one of the secondary structures, and the weight $K_2^N K_3^{n_N} K_f$ accounts for the fully folded protein. We split the total partition function into unfolded and folded partition functions:

$$Q = Q_U + Q_F \tag{2.4}$$

Q_U is the partition function for the unfolded macrostate. It is the the sum over the statistical weights of all the microstates in the unfolded ensemble. Likewise, Q_F is the partition function of the folded state. In this model, the folded state consists of only one

microstate.

$$Q_U = 1 + NK_2 + \binom{N}{2} K_2^2 K_3 + \binom{N}{3} K_2^3 K_3^3 + \ldots + NK_2^{N-1} K_3^{n_{N-1}} \tag{2.5}$$

$$Q_F = K_2^N K_3^{n_N} K_f \tag{2.6}$$

K_f is an equilibrium constant that contributes to a "stability gap" between the folded state and the $c = N - 1$ state. $K_f > 1$, representing final favorable interactions that stabilize the folded state. A possible physical origin of K_f is a final desolvation step that expels water from the protein core, similar to the dry molten globule phase discussed recently by Baldwin and Rose[182,183] and earlier by Shaknovich.[184] It's also known that desolvation barriers are important in native-centric models to correctly reproduce the wide variation observed in experimental folding rates.[185] The equilibrium populations of the "first excited" state $(c = N - 1)$ and the folded state $(c = N)$ are

$$p_{N-1}(eq) = \frac{NK_2^{N-1} K_3^{n_{N-1}}}{Q} \tag{2.7}$$

$$p_N(eq) = \frac{K_2^N K_3^{n_N} K_f}{Q} \tag{2.8}$$

In Fig. 2.2.2A, we plot free energy with respect to c for a four-helix bundle (also depicted in Fig. 2.3.2). Free energy is computed from equation 2.9. The free energy reaches a maximum at $c = N - 1$.

$$\Delta G(c) = -RT \ln[p_c(eq)] \tag{2.9}$$

2.2.2 Kinetics of the model: folding and unfolding rates

The folding and unfolding dynamics of the model are described by a continuous time Markov process (for details, see SI). If the highest barrier is at $c = N - 1$, the folding and unfolding rates are well-approximated by:

$$k_f = k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_U} \tag{2.10}$$

$$k_u = k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_F} \tag{2.11}$$

where k_1 is a rate constant for the folding of a single secondary structure. In the SI, we show that these analytical expressions capture with neglibible error, for appropriate ranges of parameter values, the results of the full master equation as computed by numerical integration (Fig. 2.2.2B) and as found by eigen-decomposition of the rate matrix. It shows, for example, that the highest barrier is indeed the last step, as we have assumed. In the SI, we also show that there is a gap in the eigenvalue spectrum, which means that the model predicts a single dominant slowest exponential relaxation time, characteristic of two-state kinetics, the general behavior seen for the folding of small globular proteins.

2.3 Results and predictions of the model

2.3.1 Folding rate predictions

Fig. 2.3.1A compares the model predictions from equation 2.10 to experimental data on folding rates of 93 globular proteins, as a function of the number N of secondary structures in their native states. We used K_2 and K_3 as parameters to fit the whole set of folding rates. We fit K_f as a function of N using the protein stability model of Ghosh and Dill.[18,186] The values of K_f ranged from 1.75 for $N = 1$ to 19.4 for $N = 30$ (Table S2). We

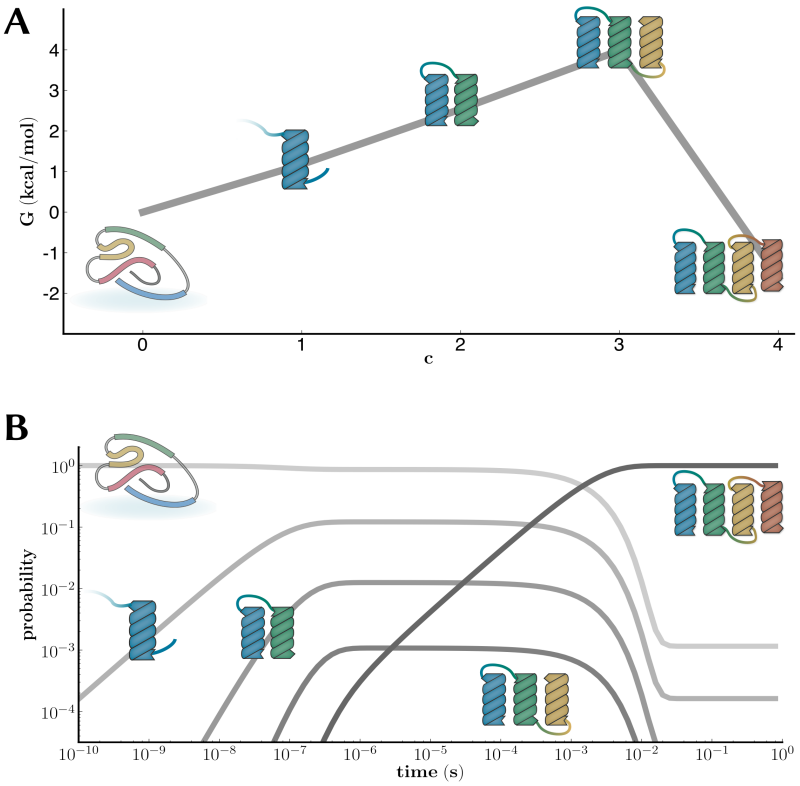


Figure 2.2.2: Landscape and dynamics of a four-helix bundle. *(A)* Free energy vs. c, the highest barrier is at $c = 3$. *(B)* Folding dynamics from numerical integration of the master equation (see SI).

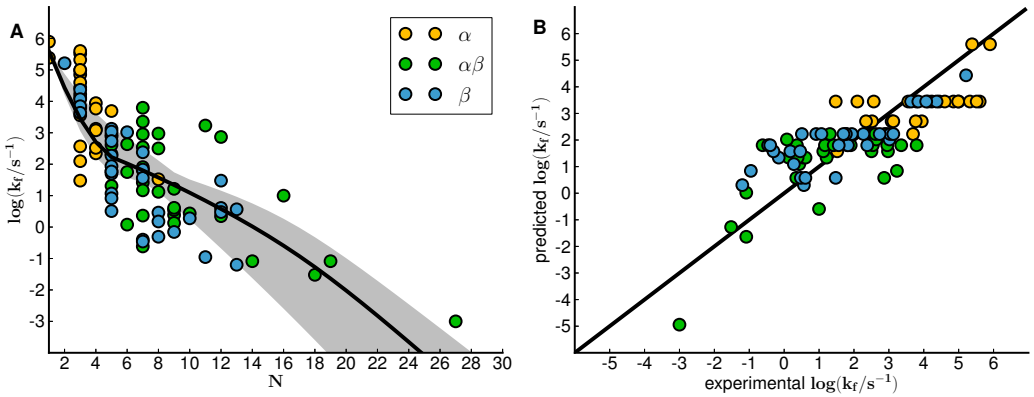


Figure 2.3.1: Folding rates predicted by model. *(A)* Folding rate vs. number of secondary structures, N. The colored points are experimental values, and they are colored by structural class. The black line is the prediction from the model, and the gray bands represent the 95% confidence interval. *(B)* Predicted folding rate vs. experimental folding rate. The colored points are the same as in the left panel. The black line represents a perfect fit to the data. Fit parameters (95% CI): $K_2 = 0.037\ (0.025, 0.058)$, $K_3 = 1.96\ (1.67, 2.23)$. We fixed $k_1 = 10^{5.6} s^{-1}$, and K_f was fitted to an equilibrium stability model, independent of the folding rate fit. Fit quality (95% CI): $R^2 = 0.63\ (0.49, 0.72)$, rms error $= 1.30\ (0.96, 1.65)$.

fixed k_1 to $10^{5.6} s^{-1}$, the mean value of the folding rates of Trp Cage and L9 helix (the two data points at $N = 1$). The proteins include both two-state and multi-state proteins. For the multi-state proteins, we fitted to the slowest folding phase. The proteins are tabulated in the supplement (Tables S3-S5). The fit line represents the predictions of our model. The best-fit values are: $K_2 = 0.037$ and $K_3 = 1.96$. Here, we've bootstrapped the data and fitted each resampled data set in order to generate a confidence interval. The 95% confidence interval bands are plotted in gray. Fig. 2.3.1B shows folding rates predicted from the model vs. experimental folding rates. $R^2 = 0.63$ for this fit.

From these comparisons, we draw a few conclusions. First, as a matter of data fitting, we compare to some other treatments of folding-rate data. While a more detailed discussion is given in SI, we note that the present fit is somewhat better than to chain length alone ($R^2 = 0.48$), to the square-root of chain length ($R^2 = 0.53$),[96–98,101,107,177] or to Absolute Contact Order (ACO) ($R^2 = 0.59$).[97,101] A disadvantage of ACO is that it requires knowledge of a protein's native tertiary structure, whereas our model does not.

Our model requires knowledge only of the number of secondary structures, which, as we note below, itself is predicted well from the chain length alone. This fact is highly advantageous for predicting folding kinetics of whole proteomes (see below). For known proteomes, protein chain lengths are fully known, whereas their native structures are not.

Second, we are interested in insights the model provides into the mechanism of protein folding. We find that $K_2 << 1$ while $K_3 > 1$. Secondary structures are unstable alone; they are stabilized by tertiary interactions. This prediction is consistent with a model of protein equilibrium cooperativity.[150] The predicted folding landscape looks like an *Up Staircase* as a function of the 1D reaction coordinate c, with a last big step down; see Fig. 2.3.2. So, the kinetics can be described in terms of *nested transition states*: from the denatured state, a rare fluctuation is required to form one of the protein's N secondary structures, then a rare fluctuation from that state further causes the protein to transition from $c = 1$ to $c = 2$, then a further rare fluctuation from state $c = 2$ gives a $c = 3$ structure, etc; see Fig. 2.2.2B. In short, each later structure is a transition state for the preceding structure. In this mechanism, the last step is the rate-limiting step, hence the kinetics are two-state.

Forming the first helix (i.e. the step from $c = 0$ to $c = 1$) is the most costly step. Forming the second helix (from $c = 1$ to $c = 2$) is less costly because the second is stabilized by the first helix. As a result, the slope of folding rate vs. N is steep at first for small N but decreases for larger values of N. The expectation that isolated secondary structures are unstable is consistent with the observation that few secondary structures remain folded if they are removed from the context of the full protein native structure to which they belong.[187]

The present model treats only how folding kinetics depends on protein size, and not how it depends on the protein's sequence. However, it is well known that sequence effects can be large. This can be seen from the broad scatter around the fit line in Fig. 2.3.1A. Some structurally similar proteins (identical N) have folding rates that can

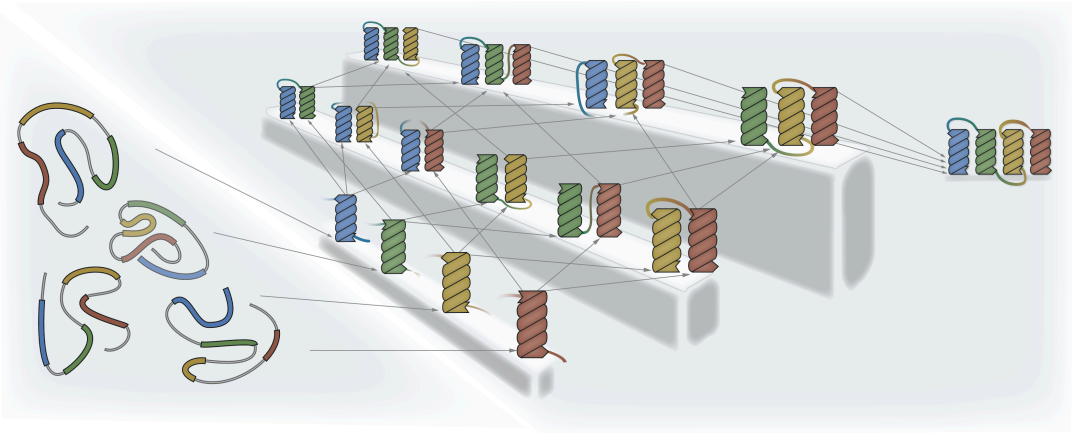


Figure 2.3.2: The folding landscape of a four-helix bundle. Each level c is an ensemble of configurations. The combinatorics are: 1 possible configuration with zero helices ($c = 0$), 4 possible one-helix configurations ($c = 1$), 6 possible two-helix configurations ($c = 2$), 4 possible three-helix configurations ($c = 3$), and 1 possible four-helix configuration ($c = 4$, the folded state). While the model treats all possible tertiary pairings, we show here only adjacent ones, to keep the figure readable.

differ by orders of magnitude. An example is the spectrin superfamily: these proteins have very different folding rates despite nearly identical chain lengths, secondary structure counts, and topologies.[188,189] Another example is the homeodomain superfamily.[53,54] Our dataset includes both the spectrin and homeodomain helix bundles.

The present model is consistent with both the funnel-landscape view that folding is a disorder-to-order transition through many different microscopic routes[31,190–192] and the view of folding based on sequential pathways and "foldons", wherein secondary structural elements fold via particular sequences of events.[68,70] Funnels and foldon paths are not mutually exclusive; they are just different perspectives at different levels of resolution, often directed at different questions and often focused on different parts of the landscape. Some aspects of the foldon path perspective are evident in the present model: the elemental folding unit is the secondary structure, the reaction coordinate is one-dimensional, the free energy increases along the reaction coordinate from U to a transition state and decreases to F, and there is a clear order of folding events through the formation of $c = 1, 2, 3, \ldots N$ secondary structures sequentially. On the other hand,

the funnel perspective is evident too: Fig. 2.3.2 shows the combinatorics of the many different routes of assembling the secondary structures (and there are additional route combinatorics that arise from the many microscopic routes for forming each secondary structure, but those are below the resolution of the present model). While real protein folding surely entails more complexity – different secondary and tertiary structures forming at different rates, some structures are finished when others are only partially completed, etc. — we believe the present model captures the essence of the physics with a minimum of parameters.

There is an important consequence of the fact that protein folding rates have a simple dependence on the number N of secondary structures. In turn, the number of secondary structures in a native protein depends in a simple linear way on the chain length L (number of amino acids); see Fig. 2.3.3. Hence, it follows that folding rates can be predicted, to first approximation, based simply on knowledge of the chain length L alone, without the need for knowledge of protein native structures. An important implication is that we can readily make estimates for the distribution of protein folding and unfolding rates over whole proteomes, since the chain lengths of all the proteins in a proteome are readily determined once a proteome is known.

2.3.2 Folding rates in the *E. coli* proteome

Our model indicates that protein folding rates are mainly a function of the number N of secondary structures in a protein. However, since N is a linear function of the length L of the protein chain, we can predict folding rates vs. chain length, $k_f(L)$, or folding times, $\tau_f(L)$.

However, a protein's folding unit is often a domain, rather than the whole chain.[4] When domains fold as independent units, the total folding time of the protein will be the sum of the folding times of its domains.[193] Domains might also fold cooperatively, but it's not yet clear if inter-domain kinetic cooperativity is widespread in proteomes.[4] Here, we

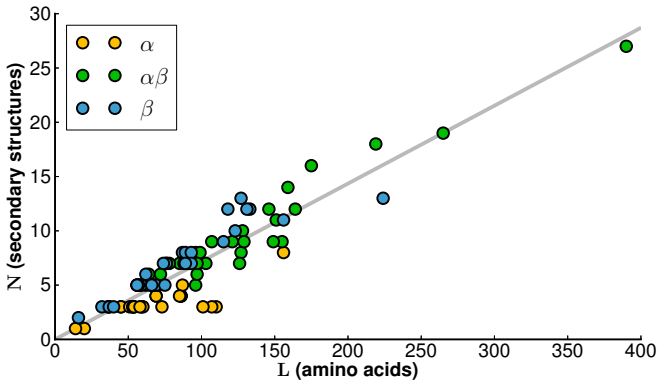


Figure 2.3.3: Number of secondary structures vs. chain length for the 93 proteins in our data set. Our fit line is $N = \gamma L$, where $\gamma = 0.0718$ secondary structures per amino acid. $R^2 = 0.85$. The slope of the line corresponds to an average of ≈ 14 amino acids per secondary structure. However, this fit includes loops, so it represents an overestimate of average secondary structure length.

compute an estimated distribution of folding rates by combining our model with domain annotations from the SUPERFAMILY database,[194] which contains domain annotations for 3003 out of the 4228 proteins in the *E. coli* proteome. In the absence of better information, we assume that each domain folds as an independent unit. We approximate the folding time of each of the 3003 annotated proteins as the folding time of its largest (and slowest) domain.

On this basis, we computed the distribution of intrinsic folding times in *E. coli*. We use the term "intrinsic" to mean in the absence of chaperones, aggregation or other cellular factors. It is not currently known how to account for those factors, but these intrinsic folding times may provide a useful reference point for future efforts that aim to account for additional biological effects on folding in the cell.

The distribution of folding times for the *E. coli* proteome peaks around the one second timescale (Fig. 2.3.4). The distribution predicts that none of the proteins are sub-millisecond folders. There has been much effort to understand ultra-fast folding domains in order to elucidate the ultimate speed limits to folding,[195] but the figure shows that such domains are not rate-limiting for the multidomain structures to which they belong.

In Fig. 2.3.4, we include bars indicating a few other timescales that are relevant to the cell: the left line (dark blue) indicates the roughly 16 seconds that is required to synthesize an average *E. coli* protein (325 amino acids x 0.05 seconds to add each amino acid in translation[196]); the middle line (orange) indicates the roughly 30 seconds it takes for *E. coli*'s GroEL chaperones to refold a protein (a protein spends about 10 seconds in the chaperone cavity, and takes about 3 recycling events to fold[10,197]); and the right line (teal) indicates *E. coli*'s mimimum doubling time of 20 minutes. The figure shows that much of the cell's protein folding activity takes place on a time scale between 10 ms at the fast end and $10 - 100$ sec on the slow end. It also shows that the folding of the slowest proteins could be bottlenecks at the fastest doubling rates of *E. coli.*

However, the figure also illuminates a huge gap in our current knowledge–how do large domains fold? Over 600 of the proteins are predicted to fold on timescales slower than the doubling time, due to large, slow-folding domains (> 400 amino acids). One explanation is that these large domains may actually be made up of subdomains that fold independently, even though current domain annotations treat them as single domains. It also seems likely that many factors may mitigate problems from slow folding times, including chaperones, folding on the ribosome, and kinetic cooperativity between protein domains.

2.4 Conclusions

We have developed a simple model of protein folding kinetics. Drawing on an earlier treatment of Zwanzig et al.,[159,160] our model posits that secondary structures are the units of folding assembly, that they are relatively unstable, that isolated units flicker in and out of structure, and that individual secondary structures are stabilized and escorted along the folding route by neighboring secondary structures. This model leads to the prediction that increasing amounts of structure are uphill in free energy, so the last step is the

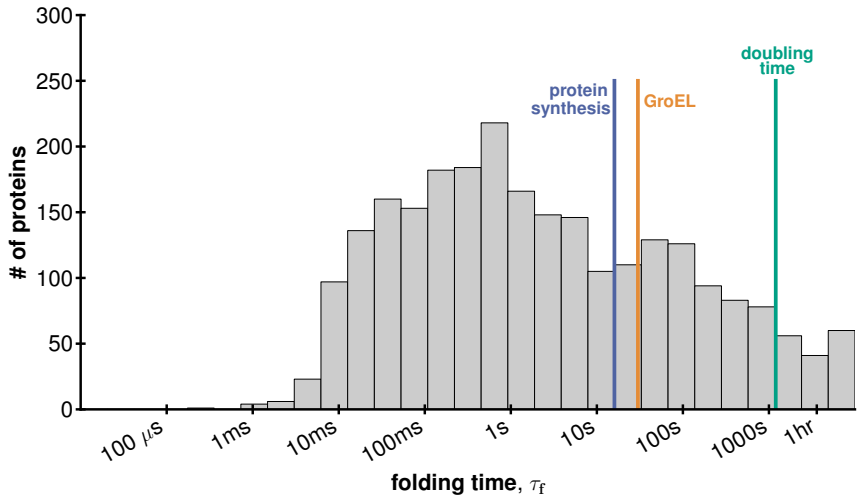


Figure 2.3.4: E. coli folding time distribution. Colored lines indicate timescales for key cellular processes: (dark blue) ribosomal protein synthesis, (orange) GroEL refolding, (teal) doubling time.

slowest, so increasingly structured chains can be regarded as nested transition states for preceding structures. The model is consistent with the two-state nature of much protein folding kinetics, and with the observed variation of folding rates of 93 proteins with numbers N of secondary structure elements (or, correspondingly, with chain length L).

We have combined this model for $\tau_f(L)$ with domain annotations from the SUPERFAMILY database to obtain an estimate for the intrinsic folding rate distribution for the *E. coli* proteome. It shows that most protein folding times range between 10 msec and 10 sec. However, a key unknown is how large domains fold. The present model predicts that the folding times of large domains could take longer than it takes to duplicate the cell.

2.5 Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship and by NSF grant PHY 1205881. We thank the Laufer Center for support and Sarina Bromberg for creating Fig. 2.3.2 and the helix bundle graphics for Fig. 2.2.2. We would also like to thank Daniel Farrell, Kingshuk Ghosh, Adam de Graff, T.J. Lane, and Justin MacCallum for helpful discussions.

2.6 Supporting Information

2.6.1 Simple metric fits

Various scaling laws have been proposed for the dependence of folding rates on protein length. Early on, Thirumalai proposed that folding rates should scale as the square root of chain length based on polymer theory arguments.[198] Later studies have suggested a range of exponents for the scaling law.[96–98,101,107,167,174,177,199,200] A recent manuscript from Lane and Pande argues that current available data is insufficient to infer the correct scaling law.[201] We show in Fig. 2.6.1 such correlations, using our data set of 93 proteins (detailed in Tables S3-S5).

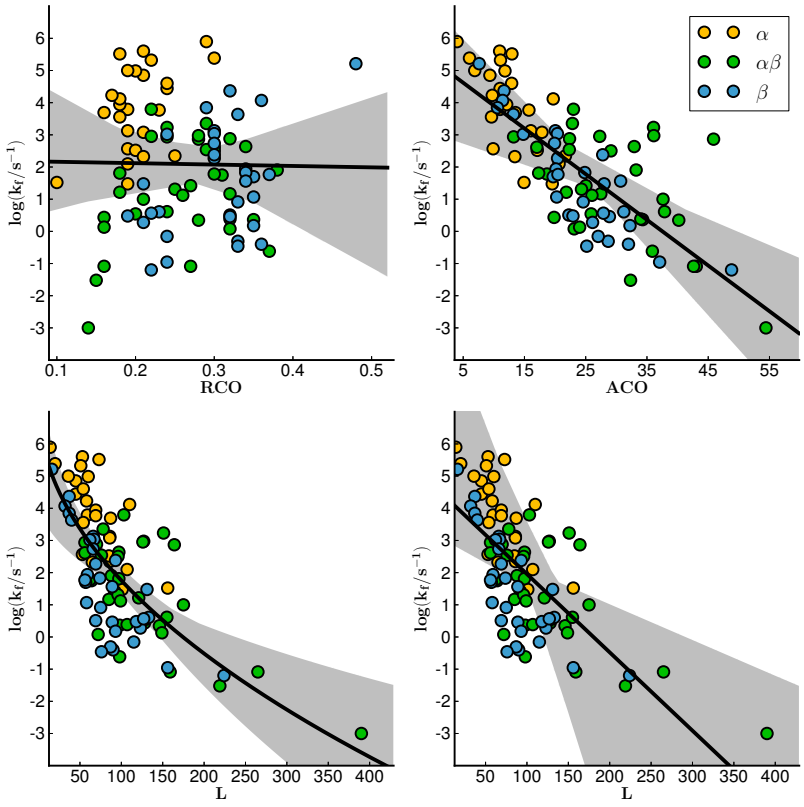


Figure 2.6.1: Folding rates vs. simple metrics (RCO, ACO, $\sqrt{L}$, and L). Proteins are colored based on structural class. The black line is the fit to the data. The gray bands represent the 95% confidence interval. Fit parameters are tabulated in Table S1.

Figure 2.6.1 shows fits to experimental folding rates for 93 proteins, both two-state and multi-state proteins. For the multi-state proteins, we fitted to the slowest phase. We show fits to some of the presently most prominent metrics: relative contact order (RCO),[93] absolute contact order (ACO),[96–98] chain length (L), and the square root of chain length ($\sqrt{L}$).[101,102,177,200] We fit the function $log\,(k_f) = log\,(k_0) - ax$, where x is RCO, ACO, $\sqrt{L}$, or L using k_0 and a as adjustable parameters.

The shading in the figures show 95% confidence intervals, which we obtain by bootstrapping[202,203] the data (resampling with replacement). Because $\sqrt{L}$ gives a reasonable fit to the data, and because the principal difference between RCO and ACO is simply that the latter contains the chain length,[96–98,101] a key conclusion is the importance of the chain length in predicting protein folding rates, more important than the topology, *per se*.

Table 2.6.1: Simple metric fit parameters and fit quality. Values in parentheses represent 95% confidence intervals from bootstrapping.

Model	R^2	rmse	$log(k_0)$	a
RCO	$0.00\ (0.00, 0.08)$	$3.46\ (2.51, 4.22)$	$2.21\ (0.53, 4.09)$	$0.43\ (-5.56, 7.40)$
ACO	$0.59\ (0.41, 0.75)$	$1.40\ (0.96, 1.88)$	$5.33\ (4.60, 6.03)$	$0.14\ (0.11, 0.17)$
L	$0.48\ (0.32, 0.61)$	$1.82\ (1.39, 8.51)$	$4.39\ (3.35, 9.64)$	$0.02\ (0.01, 0.07)$
$\sqrt{L}$	$0.53\ (0.35, 0.66)$	$1.62\ (1.19, 2.03)$	$7.24\ (6.33, 8.16)$	$0.55\ (0.46, 0.65)$

2.6.2 The kinetic model

Our kinetic model is a "one-step" continuous time Markov process,[135] a process that consists of hops between adjacent sites along a 1D lattice (Fig. 2.6.2). Each lattice site is labeled with an integer. In our folding model, we refer to the lattice sites as "states" since they correspond to configurational states of the protein. Our model has $N + 1$ states, where N is the number of secondary structures in the protein. The states span the integer

range $c \in [0, 1, 2, \ldots, N-1, N]$, where c represents the number of folded secondary structures.

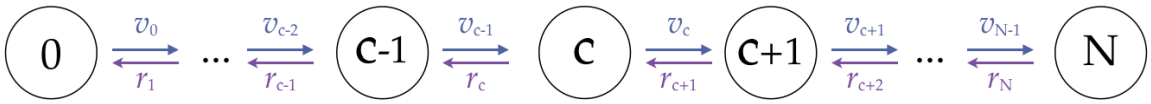


Figure 2.6.2: A one-step Markov process. Our kinetic model is a Markov process with hops between adjacent protein configurational states. Each hop in the forward direction adds another secondary structure to the folding protein.

The forward (v) and reverse (r) hopping rates are

$$v_c = (N - c)\, k_1 \tag{2.12}$$

$$r_{c+1} = \frac{(c+1)\, k_1}{K_2 K_3^{n_{c+1} - n_c}} \tag{2.13}$$

where k_1 is a rate constant for the folding of an isolated secondary structure and $N - c$ represents the number of secondary structures still waiting to fold, given that c have already folded. The reverse hopping rate, r_{c+1}, represents unfolding a secondary structure, and we derive it from detailed balance:

$$w(c) v_c = w(c+1) r_{c+1} \tag{2.14}$$

$$r_{c+1} = v_c \frac{w(c)}{w(c+1)} \tag{2.15}$$

Where $w(c)$ and $w(c+1)$ are the Boltzmann weights of states c and $c+1$, respectively. Plugging in equation 1 from the main text for $w(c)$ and $w(c+1)$, we get

$$r_{c+1} = v_c \frac{\frac{N!}{c!(N-c)!} K_2^c K_3^{n_c}}{\frac{N!}{(c+1)!(N-(c+1))!} K_2^{c+1} K_3^{n_{c+1}}} \tag{2.16}$$

$$= v_c \frac{(c+1)c!(N-c-1)!}{c!(N-c)(N-c-1)!} \frac{1}{K_2 K_3^{n_{c+1} - n_c}} \tag{2.17}$$

where in the second step we divided through and rewrote $(c+1)!$ as $(c+1)c!$ and $(N-(c+1))!$ as $(N-c)\,(N-c-1)!$. Several combinatoric terms cancel and we're left with

$$r_{c+1} = v_c \frac{c+1}{N-c} \frac{1}{K_2 K_3^{n_{c+1}-n_c}} \tag{2.18}$$

Finally, we replace v_c with equation 2.12 to get the result shown above in equation 2.13:

$$r_{c+1} = (N-c)\, k_1 \frac{c+1}{N-c} \frac{1}{K_2 K_3^{n_{c+1}-n_c}} = \frac{(c+1)k_1}{K_2 K_3^{n_{c+1}-n_c}} \tag{2.19}$$

Escape from the folded state has an additional factor, K_f, which stabilizes the folded state:

$$r_N = \frac{N k_1}{K_2 K_3^{n_N - n_{N-1}} K_f} \tag{2.20}$$

which we can rewrite in terms of Q_F, the folded partition function

$$r_N = k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_F} \tag{2.21}$$

Given the forward and reverse rates, we write the kinetics of our model as a master equation:

$$\frac{dp_c}{dt} = r_{c+1} p_{c+1} + v_{c-1} p_{c-1} - (r_c + v_c) p_c \tag{2.22}$$

2.6.3 Analytical expressions for the folding and unfolding rates

To compute the folding and unfolding rates, we follow Zwanzig[160] and posit that the rate-limiting step is the transition from the "first-excited state" ($c = N-1$) to the folded state ($c = N$). Based on Eqn. 2.22, we can write the rate of change of the population of

the folded state as

$$\frac{\mathrm{d}p_N}{\mathrm{d}t} = v_{N-1}p_{N-1} - r_N p_N \tag{2.23}$$

When we substitute Eqn. 2.12 for v_{N-1} and Eqn. 2.21 for r_N, we get

$$\frac{\mathrm{d}p_N}{\mathrm{d}t} = k_1 p_{N-1} - k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_F} p_N \tag{2.24}$$

We follow Zwanzig's local thermodynamic equilibrium (LTE) approximation that says, if the highest barrier is $c = N - 1$, all the states $c < N$ rapidly equilibrate, conditional on $p_N(t)$,

$$p_c(t) = \frac{w(c)}{Q_U}\left(1 - p_N\left(t\right)\right) \quad \text{for } c < N \tag{2.25}$$

We substitute Eqn. 2.25 for p_{N-1} in Eqn. 2.24 to get

$$\frac{\mathrm{d}p_N}{\mathrm{d}t} = k_1 \frac{w(N-1)}{Q_U}\left(1 - p_N\left(t\right)\right) - k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_F} p_N \tag{2.26}$$

$$= k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_U}\left(1 - p_N\left(t\right)\right) - k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_F} p_N \tag{2.27}$$

This shows that the rate of change of the folded state population is a competition between a rate of gain, $k_1 N K_2^{N-1} K_3^{n_{N-1}} Q_U^{-1}$, and a rate of loss, $k_1 N K_2^{N-1} K_3^{n_{N-1}} Q_F^{-1}$. This is analogous to a two-state folding reaction $U \rightleftarrows F$ in which

$$\frac{\mathrm{d}[F]}{\mathrm{d}t} = k_f[U] - k_u[F] \tag{2.28}$$

Comparing Eqn. 2.28 with Eqn. 2.27, we see that k_f is our rate of gain and k_u is our rate of loss:

$$k_f = k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_U} \tag{2.29}$$

$$k_u = k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_F} \tag{2.30}$$

This result is shown as Eqn. 10 and Eqn. 11 in the main text.

2.6.4 Numerical integration of master equation

The master equation (Eqn. 2.22) can be written in matrix form as

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{A} \tag{2.31}$$

where $\mathbf{P}(t)$ is a vector of the state probabilities at time t and $\mathbf{A}$ is the transition rate matrix. The matrix elements of $\mathbf{A}$ are:

$$\mathbf{A} = \begin{pmatrix} -v_0 & v_0 & 0 & \cdots & 0 & 0 & 0 \\ r_1 & -(r_1+v_1) & v_1 & \cdots & 0 & 0 & 0 \\ 0 & r_2 & -(r_2+v_2) & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(r_{N-2}+v_{N-2}) & v_{N-2} & 0 \\ 0 & 0 & 0 & \cdots & r_{N-1} & -(r_{N-1}+v_{N-1}) & v_{N-1} \\ 0 & 0 & 0 & \cdots & 0 & r_N & -r_N \end{pmatrix} \tag{2.32}$$

Each off-diagonal element $\mathbf{A}_{ij}$ represents the transition rate from state i to state j, but only adjacent states (that differ by one secondary structure) have non-zero transition rates. Each diagonal element $\mathbf{A}_{ii}$ is defined so that the rows sum to zero (satisfying the

condition that total probability density should be conserved). We can numerically solve equation 2.31 to get the state populations at time t based on the populations at $t = 0$:

$$\mathbf{P}(t) = \mathbf{P}(0)\, e^{\mathbf{A}t} \tag{2.33}$$

We compute state probabilities as a function of time using equation 2.33 with initially all of the probability density localized to the fully unfolded state ($p_0(0) = 1$, $p_i(0) = 0$ for $i > 0$). We plot the results of one such calculation in Fig. 2B in the main text. Then, to get the folding rate, we compute the rate spectrum using the *ratespec* python package of Voelz and Pande.[204] Given a time trace, *ratespec* calculates the rate spectrum using regularized linear regression.

2.6.5 Eigendecomposition of rate matrix

We can also compute a folding rate from the eigen values of our rate matrix, $\mathbf{A}$. We diagonalize $\mathbf{A}$ as follows:

$$\mathbf{A} = \mathbf{B}\mathbf{\Lambda}\mathbf{B}^{-1} \tag{2.34}$$

where $\mathbf{B}$ is a matrix of the eigenvectors of $\mathbf{A}$ and $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues of $\mathbf{A}$ along the diagonal. The smallest eigenvalue is $\lambda_1 = 0$, and its corresponding eigenvector represents the equilibrium populations of the states of the model. The folding rate is obtained from the smallest non-zero eigenvalue, $k_f = -\lambda_2$. The larger eigenvalues represent dynamics occuring on faster timescales. We observed two-state folding in our model: there was a clear separation of timescales (Fig. 2.6.3).

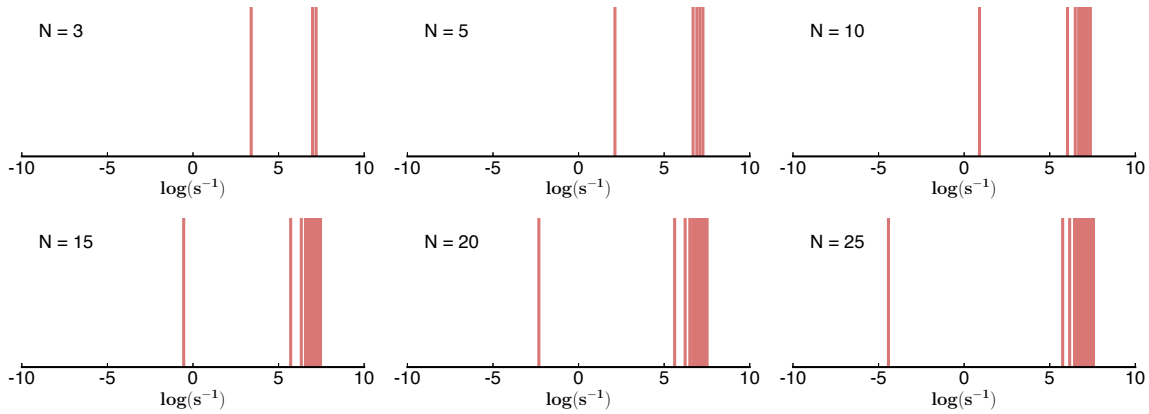


Figure 2.6.3: Eigen spectra for different values of N. At each value of N, we see two-state folding. There is a clear separation of time scales between the folding mode and faster modes.

2.6.6 Numerical validation of analytical expression for folding rate

Here we show that our analytical expression for the folding rate (Eqn. 2.29) is consistent with the results of numerical simulations of our master equation, as well as eigen decomposition of the rate matrix. In Fig. 2.6.4, we plot the results from computing k_f via the three different approaches: (1) from Eqn. 2.29 (gray), (2) from numerical integration of master equation (Eqn. 2.33), and (3) from the smallest non-zero eigenvalue. The overlap between the three methods is very good. We find that the analytical expression for k_f is a good approximation for the more exact numerical evaluations of k_f.

2.6.7 Fitting parameter K_f to protein stability model.

We used the linear relationship between chain length (L) and number of secondary structures (N) (Fig. 5) to compute an average chain length (L_{fit}) at each N. Then, we used the protein stability model of Dill and Ghosh[18,186] to fit K_f for each N (for each L_{fit}). The Dill & Ghosh model predicts stability as a function of L and temperature. We set $T = 300K$. The fit values are tabulated below (Table 2.6.2).

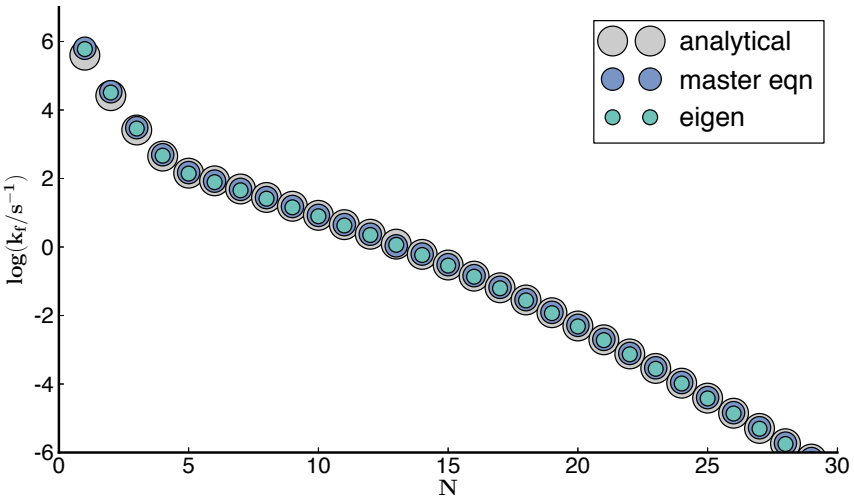


Figure 2.6.4: Comparison of three methods for computing $log(k_f)$ from our model. The agreement between the three methods is very good. We use different dot sizes so that the results from all three methods can be seen.

2.6.8 Comparison with Zwanzig model

Here, for ease of comparison, we explain how the parameters in our model map on to the original Zwanzig model.[160] We've split Zwanzig's $K = \nu e^{-\beta U}$ into our two equilibrium constants, K_2 and K_3. We don't break the equilibrium constants down into enthalpic and entropic contributions, as Zwanzig does with his parameters U and ν, respectively. Zwanzig's ν represents the number of incorrect configurations per residue. This chain entropy is one of the components of our parameters, K_2 and K_3. Zwanzig's stability gap $e^{-\beta\epsilon}$ corresponds to our K_f. Our kinetic rate k_1 is the same as Zwanzig's. Our order parameter c represents the number of correct secondary structures; Zwanzig's order parameter was S, the number of incorrect residues.

Table 2.6.2: Fit values of model parameter $\mathbf{K_f}$ as a function of $\mathbf{N}$ at $\mathbf{T = 300K}$.

N	L_{fit}	$log\,(K_f)$
1	14	1.75
2	28	3.21
3	41	4.37
4	55	5.23
5	69	5.79
6	83	6.35
7	97	6.90
8	111	7.46
9	124	8.02
10	138	8.58
11	152	9.15
12	166	9.71
13	180	10.27
14	194	10.84
15	207	11.41
16	221	11.99
17	235	12.56
18	249	13.15
19	263	13.74
20	277	14.33
21	290	14.94
22	304	15.55
23	318	16.18
24	332	16.81
25	346	17.46
26	360	18.11
27	373	18.77
28	387	19.44
29	401	20.11

2.6.9 Proteins in data set

This set of proteins is largely identical to the set used by Ouyang and Liang.[101] To that set, we've added some additional two-state proteins from the data set of Zou and Ozkan,[105] as well as the spectrins R15, R16, and R17,[188] the homeodomain, Pit1,[54] and the L9 helix characterized by Mukherjee et al.[205]

Table 2.6.3: List of two-state proteins in data set

Name	PDB	Structure	Length	N	$log(k_f)$
a3D	2A3D	α	73	3	5.52
Abp1 SH3	1JO8	β	58	5	1.07
AcP	1APS	$\alpha\beta$	98	7	-0.62
AcP common	2ACY	$\alpha\beta$	98	7	0.36
ADA2h	1O6X	$\alpha\beta$	70	5	2.88
Albumin bd	1PRB	α	53	3	5.60
bACBP	2ABD	α	86	4	2.35
BBL	2WXC	α	45	3	4.85
Bc Csp	1C9O	β	66	5	3.13
c myb	1FEX	α	59	3	3.79
CheW	1K0S	$\alpha\beta$	151	11	3.23
CI2	2CI2	$\alpha\beta$	64	6	1.75
CspA	1MJC	β	69	5	2.27
CspB	1CSP	β	67	5	3.04
Cyclophilin A	1LOP	$\alpha\beta$	164	12	2.87
CTL9	1DIVC	$\alpha\beta$	92	7	1.42
E3BD WW	1W4E	α	45	3	4.44
EC298	1RYK	α	69	4	3.94
FBP WW	1E0L	β	37	3	4.37
FKBP12	1FKB	$\alpha\beta$	107	9	0.39
FNfn9	1FNF9	β	90	7	-0.40
fyn SH3	1SHF	β	59	5	1.94
hbLBD BCKD	1K8M	β	87	8	-0.31
HPr	1HDN	$\alpha\beta$	85	7	1.17
Im7	1AYI	α	86	4	3.13
Im9	1IMQ	α	86	4	3.08
L23	1N88	$\alpha\beta$	96	5	1.31
L9 helix	n/a	α	14	1	5.90
lambda	1LMB	α	87	5	3.69
MerP	2HQI	$\alpha\beta$	72	6	0.08
NTL9	1DIVN	$\alpha\beta$	56	5	2.94
P13	1QTU	β	115	9	-0.16
PI3 SH3	1PKS	β	76	7	-0.46
POB	1W4J	α	51	3	5.32

Table 2.6.4: List of two-state proteins in data set (continued)

Name	PDB	Structure	Length	N	$log(k_f)$
protA Y15W	1SS1	α	60	3	4.99
protG	1PGB	$\alpha\beta$	56	5	2.62
protG hairpin	1PGBb	β	16	2	5.21
protL	2PTL	$\alpha\beta$	62	5	1.78
PsaE	1PSE	β	69	5	0.51
R15	R15	α	110	3	4.12
R16	R16	α	107	3	2.10
R17	R17	α	101	3	1.48
RafRBD	1RFA	$\alpha\beta$	78	7	3.35
Rap1	1IDY	α	54	3	3.56
S6	1RIS	$\alpha\beta$	97	6	2.64
Sho1 SH3	2VKN	β	75	5	0.92
spectrin SH3	1SHG	β	57	5	1.70
src SH2	1SPR	$\alpha\beta$	103	7	3.80
src SH3	1FMK	β	56	5	1.77
sso7d	1SSO	β	62	6	3.02
Tendamistat	2AIT	β	74	7	1.83
Tm1083	1J5U	$\alpha\beta$	127	8	2.98
Tm Csp	1G6P	β	66	5	2.74
TNfn3	1TEN	β	89	8	0.46
Trf1	1BA5	α	53	3	2.57
TrpCage	1L2Y	α	20	1	5.38
Twitchin	1WIT	β	93	8	0.18
U1A	1URN	$\alpha\beta$	96	8	2.50
Ubq	1UBQ	$\alpha\beta$	76	7	2.54
Urm1	2QJL	$\alpha\beta$	99	8	1.12
Villin	1VII	α	36	3	5.00
Villin 14T	2VIK	$\alpha\beta$	126	7	2.95
WW pin	1PIN	β	32	3	4.07
WW prototype	1E0M	β	37	3	3.84
WW YAP	1K9Q	β	40	3	3.63

Table 2.6.5: List of multi-state proteins in data set

Name	PDB	Structure	Length	N	$log(k_f)$
AlphaLactAlb	1HMK	$\alpha\beta$	121	9	1.22
ApoPseuAz	1ADW	β	123	10	0.28
BetaLactoGlob	1BEB	β	156	11	-0.95
CheY	3CHY	$\alpha\beta$	128	10	0.43
Colicin E7	1CEI	α	85	4	2.52
CPGK	1PHPc	$\alpha\beta$	219	18	-1.52
CRBP II	1OPA	β	133	12	0.61
Cro	2CRO	α	65	5	2.32
DHFR	1RA9	$\alpha\beta$	159	14	-1.09
EnHD	1ENH	α	54	3	4.60
FF HYPA	1UZC	α	69	4	3.77
FNfn10	1FNF10	β	93	7	2.38
GFP	1B9C	β	224	13	-1.20
GroEL apical	1DK7	$\alpha\beta$	146	12	0.35
HEWL	1HEL	$\alpha\beta$	129	9	0.54
HisActPhil	1HCD	β	118	12	0.48
IFABP rat	1IFC	β	131	12	1.48
ILBP	1EAL	β	127	13	0.56
NHypF	1GXT	$\alpha\beta$	88	7	1.91
NPGK	1PHPn	$\alpha\beta$	175	16	1.00
P16	2A5E	α	156	8	1.52
Pit1	1AU7	α	58	3	4.23
RNase HI	2RN2	$\alpha\beta$	155	9	0.61
StaphNuc	1JOO	$\alpha\beta$	149	9	0.13
Suc1	1SCE	$\alpha\beta$	97	7	1.81
TrypSynthAlpha	1QOPa	$\alpha\beta$	265	19	-1.09
TrypSynthBeta	1QOPb	$\alpha\beta$	390	27	-3.00
Twitchin Ig	1TIT	β	89	7	1.56

3

Continuous time aggregated Markov models for maximum likelihood estimation

We now shift from protein folding to a new topic: super-resolution microscopy. In the chapter that follows this one, we will describe a maximum likelihood method for analyzing super-resolution microscopy data using aggregated Markov models. As a prelude to that, in this chapter we will introduce aggregated Markov models and explain how they've been used in the ion channel literature.

3.1 Aggregated Markov models for ion channel gating kinetics

The development of patch clamp experiments in the 70s and 80s made it possible to record the currents of single ion channels in membranes.[206–210] Patch clamp recordings revealed that channels fluctuate between a closed (low conductance) and an open (high conductance) state. We can think of the gating dynamics in terms of a two-state kinetic model (Fig. 3.1.1). The rate of hopping from the closed state to the open state is k_{co}, and the reverse rate is k_{oc}. One can determine the kinetic rates by fitting the observed

distribution of closed and open times. We show an example from simulated data in Fig. 3.1.2. The statistics from the five time traces shown in the left panel of Fig. 3.1.2, along with many other time traces, are combined to get the histograms shown in the two panels on the right. The rates are then determined by fitting equations 3.1 and 3.2 to the closed and open histograms, respectively.

$$f_c(t_c) = k_{co} e^{-k_{co} t_c} \tag{3.1}$$

$$f_o(t_o) = k_{oc} e^{-k_{oc} t_o} \tag{3.2}$$

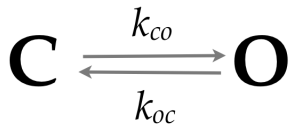


Figure 3.1.1: An ion-channel gating model with two states: closed (C) and open (O).

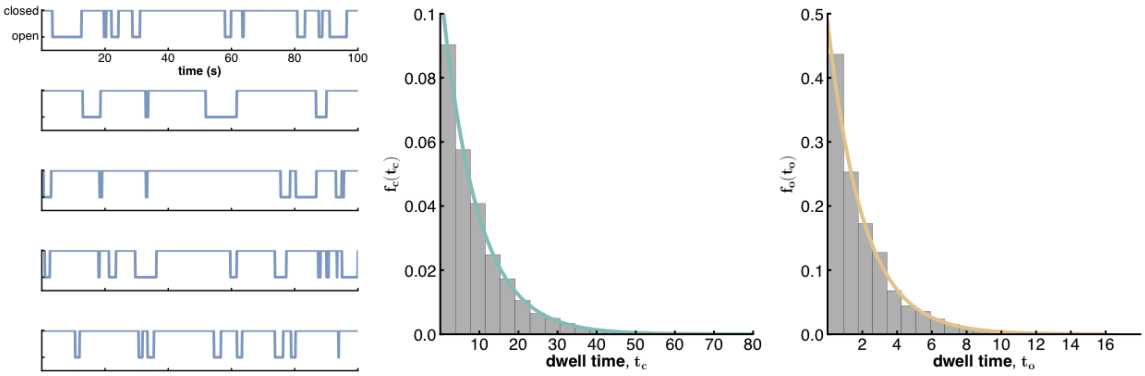


Figure 3.1.2: (left) Simulated trajectories, $k_{co} = 0.1\ s^{-1}$ and $k_{oc} = 0.5\ s^{-1}$ (center) lifetime distribution for closed state, (right) lifetime distribution for open state. The histograms are well-fit by single-exponential distributions.

The histogram method works well for the 2-state scheme presented above, but real ion channel gating is not that simple. Real ion channels have multiple closed and open states (Fig. 3.1.3), and the observed output from the experiment does not uniquely specify the

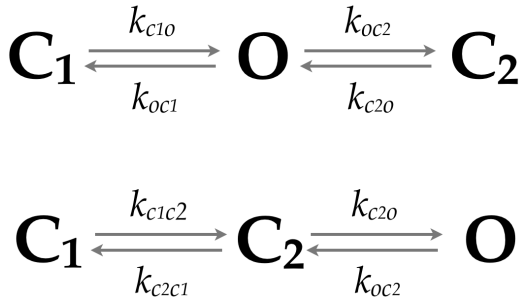


Figure 3.1.3: Ion-channel gating models with two closed states (C_1, C_2) and one open state (O). (top) COC model; (bottom) CCO model. Based on equilibrium data alone, these two models would be indistinguishable.[211]

state of the channel, e.g. C_1 and C_2 in Fig. 3.1.3 would have the same conductance. To extract kinetic rates for these more complicated topologies, we need to use aggregated Markov models (AMMs). During the 80s and 90s, Colquhoun, Hawkes, Sachs and others pioneered the use of AMMs to extract kinetic rates from single channel recordings[211–221] and processive molecular motors.[222,223] The term "aggregated" in AMM refers to a partitioning of the state space into classes, in this case a closed class and an open class. It is only possible to observe which class the system is in, not the specific state. For example, given a three-state model like those shown in Fig. 3.1.3, the best we can do is observe that the channel is closed, we cannot know with certainty whether it is in C_1 or C_2. Additionally, both models would have biexponential dwell time distributions, so they can't be distinguished, even in theory. As discussed in previous chapters, the term "Markov" in AMM refers to the fact that the probablility of transitioning from one state to another depends only on the current state, not on previously occupied states.

AMMs are a special case of hidden Markov models (HMMs).[224] In AMMs, the output probabilities for each state are fixed to zero or one. For example, closed states have probability zero of producing open conductance and probability one of producing closed conductance. The converse is true for open states. In HMMs, the outputs depend probabilistically on the state.

3.2 Maximum likelihood estimation of AMM channel kinetics

Unlike the simple single-exponential distributions in Fig. 3.1.2, the lifetime distribution of an aggregated class in an AMM is often multi-exponential. It is still possible to infer kinetic rates by constructing lifetime histograms, as in the two-state example above, but this approach is not efficient.[218] Lifetime histograms do not take into account the correlations between dwell times. Consequently, useful information is discarded. Another problem is that large amounts of data are needed for accurate histogram fitting.

A smarter way to fit AMMs to data is to use a maximum likelihood approach that uses the joint probability density of observed time traces. This approach was initially proposed by Horn & Lange in terms of a discretization of a continuous time Markov process.[214] A more efficient likelihood method was developed by Ball & Sansom who dealt with the continuous time Markov process directly, rather than discretizing it.[216] Sachs further developed the likelihood method to incorporate missed events (a channel gating event that happens faster than the resolution of the experiment) and to incorporate a more efficient forward-backward recursive procedure for computing the likelihood.[218,219]

The key idea in the AMM maximum likelihood approach is to compute the probability of escape from an aggregated class of states. Above, in equations 3.1 and 3.2, we showed examples of how to compute the probability of escape from a single state. We can express the probability density of escape from an aggregated class of states as:

$$f(c,t) = \mathbf{P}_0 e^{\mathbf{Q}_{cc}t}\mathbf{Q}_{co}\mathbf{1} \tag{3.3}$$

Here, $\mathbf{Q}$ is a rate matrix, and the matrix exponential function is shorthand for an infinite series ($e^{Qt} = I + Qt + Q^2t^2/2! + \ldots$). The vector $\mathbf{P}_0$ specifies the initial probability of each state. The vector $\mathbf{1}$ is a column vector of ones, which serves to collapse the matrix product down to a scalar value. The matrix exponential term accounts for all possible

exchanges between states within the closed class during the dwell. In the case of the COC model, no such exchanges are possible because the open state acts as a gateway state between the two closed states. In the CCO model, the channel can transition back and forth between the C_1 and C_2 states many times during the dwell before eventually transitioning to the open class. The matrix exponential is numerically expensive to compute for large matrices, but a vast literature exists on efficient ways compute it.[225,226] The elements of the rate matrix $\mathbf{Q}$ for the CCO model would look like this:

$$\mathbf{Q} = \begin{pmatrix} -k_{c_1c_2} & k_{c_1c_2} & 0 \\ k_{c_2c_1} & -(k_{c_2c_1} + k_{c_2o}) & k_{c_2o} \\ 0 & k_{oc_2} & -k_{oc_2} \end{pmatrix} \tag{3.4}$$

We can divide $\mathbf{Q}$ into submatrices based on the aggregated classes:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{cc} & \mathbf{Q}_{co} \\ \mathbf{Q}_{oc} & \mathbf{Q}_{oo} \end{pmatrix} \tag{3.5}$$

$$\mathbf{Q_{cc}} = \begin{pmatrix} -k_{c_1c_2} & k_{c_1c_2} \\ k_{c_2c_1} & -(k_{c_2c_1} + k_{c_2o}) \end{pmatrix} \tag{3.6}$$

$$\mathbf{Q_{co}} = \begin{pmatrix} 0 \\ k_{c_2o} \end{pmatrix} \tag{3.7}$$

$$\mathbf{Q_{oc}} = \begin{pmatrix} 0 & k_{oc_2} \end{pmatrix} \tag{3.8}$$

$$\mathbf{Q_{oo}} = \begin{pmatrix} -k_{oc_2} \end{pmatrix} \tag{3.9}$$

$$\tag{3.10}$$

Equation 3.3 represents the probability density of starting in one of the closed states (with initial probability specified by $\mathbf{P}_0$), staying in the closed aggregated class for time t, and then transitioning out of the closed aggregated class to a state in the open aggregated

class. We can now extend equation 3.3 to the case of two dwells and ask, "What's the probability density of dwelling in the closed aggregated class for time t_1, transitioning to the open aggregated class, dwelling there for time t_2, and finally transitioning back to the closed class?" The corresponding equation for this scenario would be:

$$f\left(\{c,o\},\{t_1,t_2\}\right) = \mathbf{P}_0 e^{\mathbf{Q}_{cc}t_1}\mathbf{Q}_{co}e^{\mathbf{Q}_{oo}t_2}\mathbf{Q}_{oc}\mathbf{1} \tag{3.11}$$

Equation 3.11 represents the joint probability density for the series of two dwells described above. Following the same logic, we can easily extend this equation to compute the joint density of a long series of dwells:

$$f\left(\mathbf{h},\mathbf{t}\right) = \mathbf{P}_0 e^{\mathbf{Q}_{cc}t_1}\mathbf{Q}_{co}e^{\mathbf{Q}_{oo}t_2}\mathbf{Q}_{oc}\ldots e^{\mathbf{Q}_{oo}t_L}\mathbf{Q}_{oc}\mathbf{1} \tag{3.12}$$

where L is the total number of dwells in the time series. We use $\mathbf{h}$ to represent the set of aggregated classes in the dwell series and $\mathbf{t}$ to represent the set of dwell times, so $\mathbf{h} = \{h_1, h_2, \ldots, h_L\}$ and $\mathbf{t} = \{t_1, t_2, \ldots, t_L\}$. Each element of $\mathbf{h}$ is either the closed or open class, i.e. $h_i \in \{c, o\}$. We've assumed in this example that the dwell series ends with a dwell in the open class followed by a transition to the closed class.

In maximum likelihood analysis, the kinetic rates and N are treated as variables and the observed data is treated as a fixed set of parameters. To reflect this shift in perspective, we rewrite the joint density (Eqn. 3.12) in terms of a likelihood function:

$$f\left(\theta \mid \mathbf{h},\mathbf{t}\right) = \mathbf{P}_0 e^{\mathbf{Q}_{cc}t_1}\mathbf{Q}_{co}e^{\mathbf{Q}_{oo}t_2}\mathbf{Q}_{oc}\ldots e^{\mathbf{Q}_{oo}t_L}\mathbf{Q}_{oc}\mathbf{1} \tag{3.13}$$

where θ is the set of rates, $\theta = \{k_{c_1c_2}, k_{c_2c_1}, k_{c_2o}, k_{oc_2}\}$. Equation 3.13 is the key equation in the maximum likelihood approach of Sachs and others.[214,216,218,219] The goal of the maximum likelihood approach is to find the parameter set $\hat{\theta}$ that maximizes the likelihood function. The set $\hat{\theta}$ represents the best estimate of the kinetic rates, given the

dwell series data. Note that although we've used the CCO three-state gating model as an illustrative example here, equation 3.13 is a general concept that we can apply to any Markov model with a finite number of discrete states. It can be easily extended to kinetic models that have larger numbers of states and/or that have more than two aggregated classes.

If the correct kinetic model is unknown, one can compute the likelihood of a series of models and determine the best model based on which one maximizes the likelihood. However, Kienker showed that some models (related by a similarity transformation) are indistinguishable if only equilibrium statistics are available.[211] The pair of three-state models shown above (Fig. 3.1.3) is one such example. But Kienker also showed that perturbation experiments can help resolve indistinguishability issues by transiently increasing the number of experimental observables.

One final note, it is possible to modify the likelihood function to account for the activity of a collection of independent channels[218,227,228] (see chapter 7 in van Kampen[135] for a general discussion of collective systems). The solution is to represent the state of the collection of channels with a set of occupation numbers $\mathbf{N} = \{N_{c_1}, N_{c_2}, N_o\}$, so that N_{c_1}, for example, represents the number of channels in the collection that are occupying the first closed state, C_1.

In the next chapter, we will adapt the AMM approach discussed here to analyze kinetic data from super-resolution microscopy, with the ultimate goal of counting single molecules *in vivo*. We will show how to write a likelihood function for the photophysics of fluorescent proteins, and we will make use of the Yeo et al.[227] formalism for collections of channels to write the likelihood function in terms of a collection of independent proteins that cluster together in a diffraction-limited volume.

4

An aggregated Markov model approach to the molecular counting problem in super-resolution microscopy

This chapter contains a manuscript in preparation for publication.

Abstract

We develop a maximum likelihood method for quantifying fluorophores in a diffraction-limited volume measured by super-resolution microscopy. The method is an extension of aggregated Markov methods developed in the ion channel literature for studying gating dynamics. We show that the method accurately and precisely (1) quantifies fluorophores in simulated data and (2) determines the kinetic rates that govern the photophysics of the fluorophores. We apply the method to *in vitro* Dendra2 data and *in vivo* data of Dendra2 fused to the bacterial flagellar motor protein FliM. Our estimate of the number of FliM subunits in a single motor is 15, roughly half the expected value based on previous Cryo-EM and FRAP studies. We discuss possible reasons for the discrepancy between our estimate and the literature values.

4.1 Introduction

Conventional optical microscopy is diffraction-limited and typically cannot resolve images below the 250nm range, but super-resolution (SR) methods can probe the nanometer level.[229–234] Photoactivated localization microscopy (PALM) is one such SR method. PALM can image molecules closer than the diffraction limit by separating their fluorescent signals in time.[232] PALM works by illuminating a sample under low light intensity, which stochastically triggers fluorophore activation. Once active, the fluorophore is excited by light of a different wavelength and releases a burst of photons (an emission burst). A short time later, it will irreversibly photobleach. The light intensity can be modulated to increase the average time separation between fluorophore activation events.[235] In PALM, the fluorophores are genetically encoded photoactivatable fluorescent proteins (PA-FPs),[236] which are fused to proteins of interest. Currently, mEos2[237] and Dendra2[238,239] are two of the most common PA-FPs used in PALM.

PALM has the potential to provide molecular counting with single molecule sensitivity. However, several obstacles remain:

1. **PA-FP "blinking" leads to severe overcounting biases.** Blinking refers to a process by which a PA-FP produces a series of intermittent emission bursts, instead of one continuous burst.[233,240] This is a problem because, ideally, one would count molecules by summing up the number of observed emission bursts. However, due to blinking, a simple sum of bursts will overcount the true number of molecules (Fig. 4.1.1).

2. **Unknown blinking statistics.** The blinking properties of common PA-FPs have been characterized *in vitro*, but not *in vivo* where the actual experiments are done. Current analysis methods are incapable of extracting such information from *in vivo* data.

3. **The missed event problem.** Some photoactivation or blinking events are missed due to the finite temporal resolution of the imaging methods (≈ 50 ms).

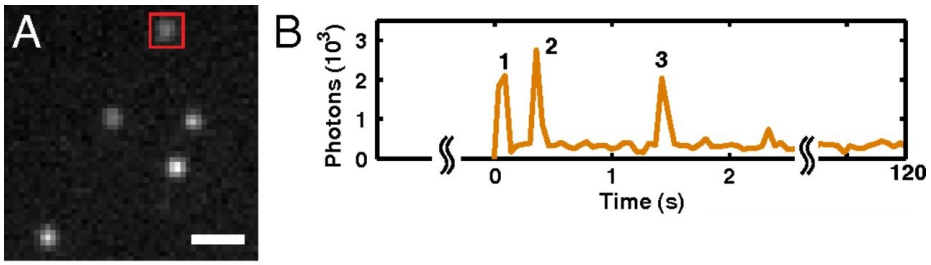


Figure 4.1.1: Dendra2 blinks *in vitro*. Several emission bursts from the same fluorophore are observed. Simply counting the number of emission bursts would overcount the true number of molecules. The three bursts shown in B actually come from a single Dendra2 molecule highlighted in A. Reprinted with permission from Lee et al.[235] Copyright (2012) National Academy of Sciences, USA. See page 99 for license information.

Here, we describe a maximum-likelihood approach for dealing with these obstacles. Our approach is an adaptation of the continuous time aggregated Markov model (AMM) techniques developed in the ion channel literature to estimate kinetic rates for channel opening and closing events in patch clamp experiments (reviewed in the previous chapter).[211–220] We extend these ideas from the ion channel world to handle a new challenge: molecular quantification in PALM.

Previous studies have addressed the PALM counting problem by setting a temporal threshold (τ_{crit}).[235,241] In those studies, a pair of emission bursts separated by a time shorter than τ_{crit} are grouped together and assigned to a single PA-FP. Bursts separated by a time longer than τ_{crit} are considered to be from separate PA-FPs.

Our method overcomes several important limitations of thresholding methods. First, thresholding methods require advance knowledge of kinetic rates to determine the optimal value of τ_{crit}. Our method doesn't require knowledge of kinetic rates beforehand. Kinetic rates are an output of our method, rather than an input. Second, thresholding methods can't account for missed events, but our method can. The missed events problem was solved by Roux[215] and later Sachs[218] for ion channel problems, and

their solution can be applied to PALM data. And third, our method is not tied to a specific kinetic model. We can easily explore alternative kinetic models, such as models with multiple blinked states or with time-varying rates.

4.2 An aggregated Markov model for the photophysics of a collection of PA-FPs in a diffraction-limited volume

4.2.1 States of the model

Consider the model shown in Fig. 4.2.1 for the photophysics of a single PA-FP. There are four possible states: inactive (I), active (A), dark (D), or photobleached (B). Once active, the PA-FP has two options: (1) it can blink to the dark state, or (2) it can irreversibly photobleach. Fluorescence is only detected in the active state, not in the other three states.

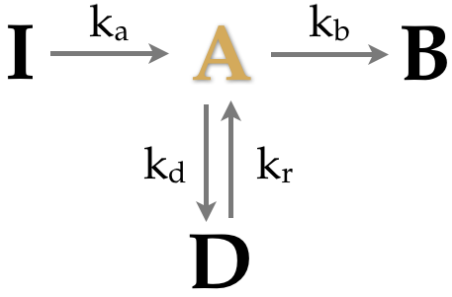


Figure 4.2.1: This kinetic model has four states inactive (I), active (A), dark (D), and photobleached (B). The only fluorescent state is A. We name the transitions between states this way: activation ($I \to A$), blinking ($A \to D$), recovery ($D \to A$), and photobleaching ($A \to B$).

Now consider a collection of N identical PA-FPs, each of which is governed by the model of Fig. 4.2.1. We assume that each fluorophore is independent of the others: the state of one fluorophore doesn't affect the state of any other fluorophore. We describe the state of a system of N fluorophores as a vector of the populations of the inactive, active, dark, and photobleached states: $\{N_I, N_A, N_D, N_B\}$. To avoid confusion, we will use the term *microstate* to refer to the state of a single fluorophore (i.e. I, A, D, or B), and we will

use *macrostate* to refer to a population vector that describes the collection of fluorophores. For example, macrostate i for a collection of two PA-FPs in which both PA-FPs are inactive would be $\mathbf{s_i} = \{2, 0, 0, 0\}$. The set of all macrostates is $\mathbf{S} = \{\mathbf{s_1}, \ldots, \mathbf{s_M}\}$, where each $\mathbf{s_i}$ in $\mathbf{S}$ is a population vector and M is the total number of macrostates.

Computing M is a common combinatorial problem: the number of unique ways to partition N indistinguishable objects into x bins. In this case, the objects are PA-FPs and the bins are the microstates. There are four microstates, so $x = 4$.

$$M = \binom{N + x - 1}{x - 1} = \binom{N + 3}{3} \tag{4.1}$$

For example if the collection contains two fluorophores the following 10 macrostates (obtained from $\binom{5}{3}$) are available:

Table 4.2.1: Macrostates for a collection with $N = 2$ PA-FPs

macrostate	$\{I, A, D, B\}$	aggregated class
$\mathbf{s_1}$	$\{2, 0, 0, 0\}$	dark
$\mathbf{s_2}$	$\{0, 0, 2, 0\}$	dark
$\mathbf{s_3}$	$\{0, 0, 0, 2\}$	dark
$\mathbf{s_4}$	$\{1, 0, 1, 0\}$	dark
$\mathbf{s_5}$	$\{1, 0, 0, 1\}$	dark
$\mathbf{s_6}$	$\{0, 0, 1, 1\}$	dark
$\mathbf{s_7}$	$\{0, 2, 0, 0\}$	bright
$\mathbf{s_8}$	$\{1, 1, 0, 0\}$	bright
$\mathbf{s_9}$	$\{0, 1, 1, 0\}$	bright
$\mathbf{s_{10}}$	$\{0, 1, 0, 1\}$	bright

We model the collection of PA-FPs as an aggregated Markov model (AMM). As shown in table 4.2.1, each macrostate belongs to an aggregated class, either dark or bright. Macrostates with at least one active PA-FP ($A > 0$) are assigned to the bright class. Macrostates with zero active PA-FPs ($A = 0$) are assigned to the dark class.

Here, the bright class corresponds to the detection of fluorescence and the dark class

corresponds to the absence of fluorescence. The aggregated classes are necessary because PALM experiments can't distinguish between the various dark states or between the various bright states. The dark and bright classes here are analogous to the closed and open classes in the ion channel AMMs discussed in the previous chapter.

Another possibility would be that multiple levels of fluorescence are observable: bright2, bright3, etc. For the purposes of this chapter, we focus on the scenario in which all macrostates with $A > 0$ are grouped into one bright class. Extending our approach to the scenario with more aggregated classes would be straightforward. Additional aggregated classes would actually simplify the analysis problem because it would enable us to better identify the state of the collection of PA-FPs in each dwell.

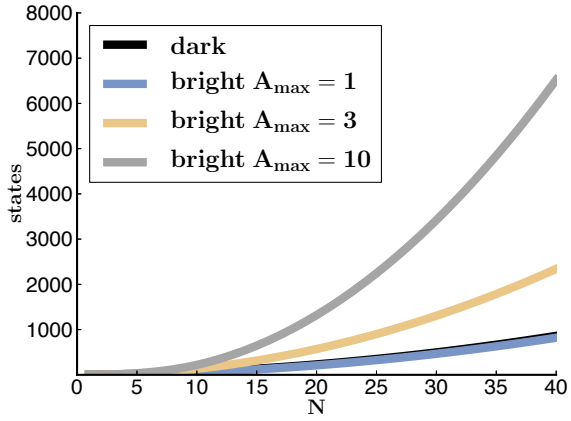


Figure 4.2.2: As the number of PA-FPs increases, the number of macrostates grows exponentially. We tune the growth rate by setting A_{max}, the number of PA-FPs that we allow to be simultaneously active.

The number of macrostates grows exponentially (Fig. 4.2.2) with N. This is a concern for the numerical calculations discussed in the next section; the computational time of the likelihood depends on the number of macrostates. As such, we define a quantity A_{max}, which represents the maximum number of PA-FPs we allow to be simultaneously photoactive. We use this quantity as a way to tune the size of the state space to save on computational time in situations where we expect photoactivation events to be well

separated in time.

4.2.2 Model kinetics

The transition rate from macrostate $\mathbf{s_i}$ to macrostate $\mathbf{s_j}$ is simply the transition rate for one PA-FP muliplied by a combinatoric factor for the population of the appropriate microstate. The macrostate transition rates are summarized in table 4.2.2.

Table 4.2.2: Kinetic rates of a collection of PA-FPs

Transition type	Change in population	Rate $\mathbf{s_i} \rightarrow \mathbf{s_j}$
activate	$\{-1, +1, 0, 0\}$	$N_{I,i}k_a(t)$
blink	$\{0, -1, +1, 0\}$	$N_{A,i}k_d$
recover	$\{0, +1, -1, 0\}$	$N_{D,i}k_r$
photobleach	$\{0, -1, 0, +1\}$	$N_{A,i}k_b$

The dynamics of the PA-FP AMM are governed by a rate matrix, $\mathbf{Q}$. Each off-diagonal matrix element q_{ij} equals the transition rate of $\mathbf{s_i} \rightarrow \mathbf{s_j}$. The diagonal elements are set so that each row sums to zero: $q_{ii} = -\sum_{i \neq j} q_{ij}$. The macrostate transition probabilities at any time t are given by the Kolmogorov equation[135,212]:

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{Q} \tag{4.2}$$

whose solution is given by

$$\mathbf{P}(t) = \mathbf{P}(0)\, e^{\mathbf{Q}t} \tag{4.3}$$

In the case of dark and bright observation classes, we can partition the rate matrix $\mathbf{Q}$ into four submatrices, based on the dark (subscript d) and bright (subscript b) aggregated

classes:

$$Q = \begin{pmatrix} Q_{dd} & Q_{db} \\ Q_{bd} & Q_{bb} \end{pmatrix} \tag{4.4}$$

The submatrix $\mathbf{Q}_{dd}$ contains the rates for transitions from states in class d to other states in class d; $\mathbf{Q}_{db}$ contains the rates for transitions from the states in class d to the states in class b. The other two submatrices are similarly defined.

4.3 Likelihood function

The likelihood function that we describe in this section provides an answer to the following question: "Given a kinetic model and a set of kinetic rates, what's the likelihood of observing the data?" Our goal is to determine the kinetic rates and N, and we can do so by finding the values of the rates and N that maximize the likelihood function, i.e. maximize the likelihood of observing the data that was collected by PALM.

Suppose we have a trajectory of L dwells representing the dynamics of a collection of N PA-FPs. Associated with each dwell is an observed aggregated class and a dwell time. The set of observation classes is $\mathbf{h} = \{h_1, ..., h_L\}$. The set of dwell times is $\mathbf{t} = \{t_1, ..., t_L\}$. So, during dwell i we observe class $h_i \in \{d, b\}$ for duration t_i. See Fig. (4.3.1) for an illustration. The probability densities for dwelling in the dark class for time t and then transitioning to the bright class are given by the elements of the following matrix:

$$\mathbf{G}_{db}(t) = e^{\mathbf{Q}_{dd}t}\mathbf{Q}_{db}. \tag{4.5}$$

The $(i, j)^{th}$ element of $\mathbf{G}_{db}$ is the probability density of entering class d from its i^{th} state, dwelling in class d for time t and then transitioning to the j^{th} state of class b.

We wish to calculate the likelihood of the dwell trajectory $\mathbf{h}$, given the model

parameters θ, where θ is the set of parameters (N and the transition rates) that go into the rate matrix $\mathbf{Q}$. The likelihood function then reads as follows:

$$f(\theta \mid \mathbf{t}, \mathbf{h}) = \mathbf{P}_{init} \cdot \mathbf{G}_{db}(t_1)\mathbf{G}_{bd}(t_2) \ldots \mathbf{G}_{bd}(t_L)e^{\mathbf{Q}_{dd}t_{final}} \cdot \mathbf{P}_{final} \tag{4.6}$$

where $\mathbf{P}_{init}$ is a probability vector with all probability density in the fully inactive macrostate. The parameters, θ, determine the elements of the rate matrix $\mathbf{Q}$. The final factor of $e^{\mathbf{Q}_{dd}t_{final}}$ comes from the fact that all of the PA-FPs irreversibly photobleach by the end of the trajectory. After all photobleaching events occur, the system will dwell in the dark class indefinitely. We represent this with a long final dwell in the dark class for time $t_{final} = 10^4$ seconds. $\mathbf{P}_{final}$ is a probability vector with all probability density in the fully photobleached macrostate.

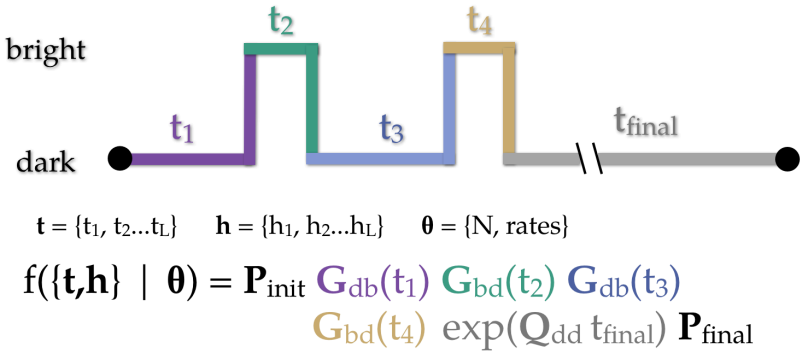


Figure 4.3.1: An idealized time trace. Each dwell is color-coded with its corresponding term in the likelihood function.

The likelihood function, as it's presented in equation 4.6 and Fig. 4.3.1, assumes that no fluorescence events are missed by the instrument. In reality, current experiments have a resolution of about 50 milliseconds, so some events will be missed. In the SI, we discuss a modification of the likelihood function that accounts for missed transitions.

4.3.1 Numerical evaluation of the likelihood function

Our goal is to find the set of parameters $\hat{\theta}$ that maximizes the likelihood function given the data (a dwell trajectory represented by $\mathbf{t}$ and $\mathbf{h}$). In practice, we accomplish this goal by maximizing the likelihood function with respect to the rates for a fixed value of N, and then we repeat the maximization process for other values of N.

We maximize the likelihood function via the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) minimizer, implemented in Scipy.[242–244] BFGS is a quasi-Newton method that performs well on non-smooth optimization problems. Our goal is to maximize the likelihood but, in practice, we minimize $-log\left[f(\{\mathbf{t},\mathbf{h}\}\mid\theta)\right]$. The computational time for maximizing the likelihood function is plotted in Fig. 4.3.2. The scaling depends on A_{max}, and it is advantageous to set A_{max} to a small value when possible (when activation events are well separated).

As an aside, we would like to point out that $A_{max}=1$ is still distinct from thresholding methods. Consider the following scenario: a PA-FP activates, and then blinks. While the first PA-FP is in the blinked state, a second PA-FP activates and photobleaches before the first molecule recovers from blinking. This scenario would still obey $A_{max}=1$, but it would be forbidden in thresholding.

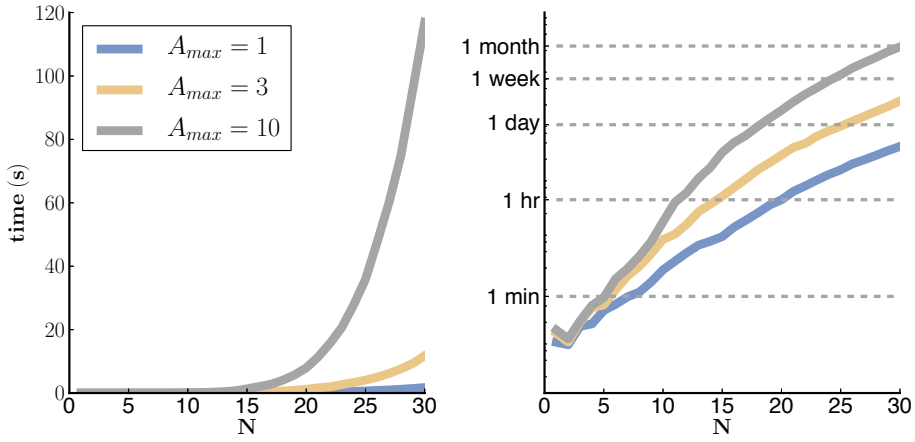


Figure 4.3.2: (left) computation time for a single matrix exponential calculation; (right) estimated time for full likelihood maximization.

4.4 Results and Discussion

4.4.1 Analysis of simulated data

We analyzed simulated trajectories of a collection of $N = 5$ PA-FPs (Fig. 4.4.1) to test the ability of our method to extract kinetic rates and estimate N from data. Our simulated trajectories were generated by the Gillespie stochastic simulation algorithm.[245] We simulated three different blinking scenarios: moderate blinking (set 1), fast blinking (set 2), and slow blinking (set 3). The parameters for the three sets are summarized in Table 4.4.1. We tuned the ratio between the blinking rate and the photobleaching rate (k_d/k_b) to control the blinking behavior. If k_d is large, relative to k_b, then many blinking events will occur before photobleaching. If k_d is much smaller than k_b, the PA-FP is more likely to photobleach without blinking.

An initial question is "does the likelihood function have a maximum in the correct location?" This question is important because it will determine whether or not we expect likelihood maximization runs to converge to the correct parameter values. If the

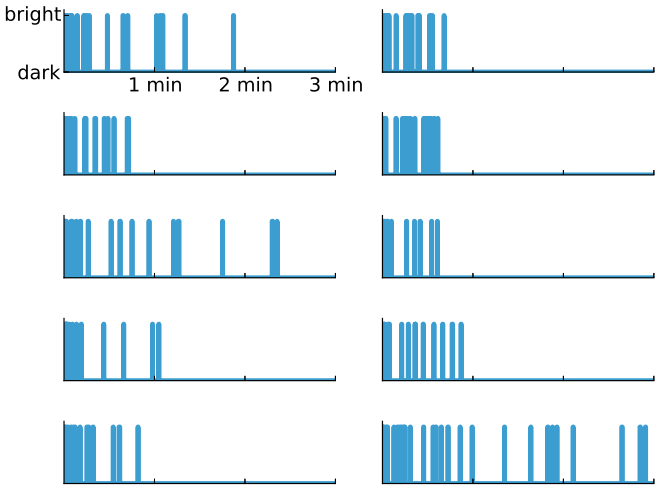


Figure 4.4.1: Sample traces from Gillespie simulations with parameters: $N = 5$, $k_a = 0.5$, $k_d = 3.0$, $k_r = 0.1$, $k_b = 1.0$. Rates in units of s^{-1}.

Table 4.4.1: Kinetic rates used to generate simulated data sets. Rates in units of s^{-1}.

Rate constant	Set 1	Set 2	Set 3
k_a	0.5	0.5	0.5
k_d	3.0	10.0	0.1
k_r	0.1	0.1	0.1
k_b	1.0	1.0	1.0
k_d/k_b	3	10	0.1

maximum lies elsewhere in parameter space, we would expect to see a bias in our results when we analyze the data. To answer this question, we computed 1D slices in parameter space around the true value of each parameter. Each panel of Fig. 4.4.2 was obtained by holding four of the five parameters (k_a, k_d, k_r, k_b, and N) constant at their true values and then varying the remaining parameter near its true value. We see that the four kinetic rates are peaked in the correct location at their true values. In the 1D slice for N, we see a maximum that spans $N = 4$ and $N = 5$.

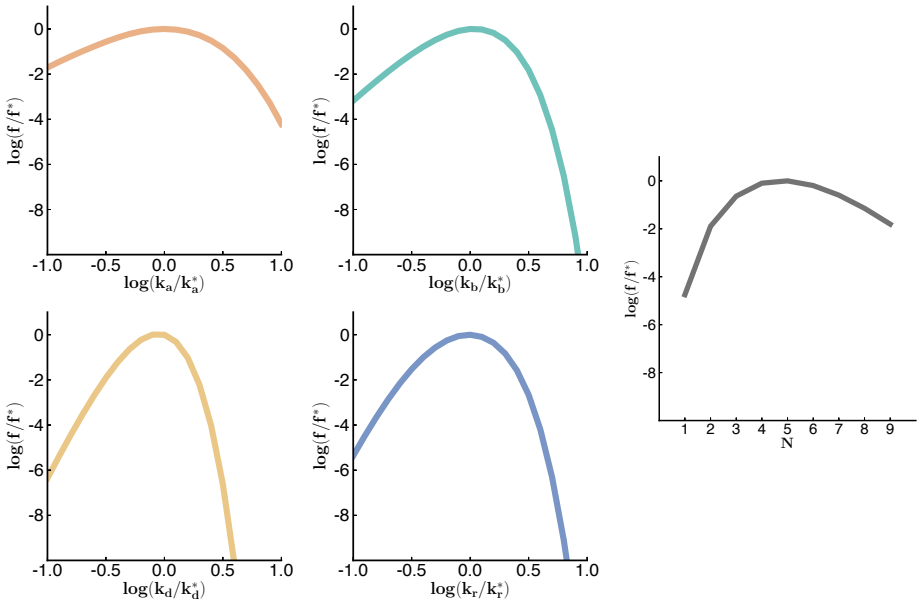


Figure 4.4.2: In each panel, one parameter is varied while the other parameters are held at their true values. We find the the likelihood is maximized at the true parameter values.

Next, we assessed the ability of the numerical maximization procedure to converge to the correct parameter values. The 1D slices discussed above suggest that the likelihood function maximum is in the correct region of parameter space, but a separate question is "can we simultaneously determine all five parameters?" We found that the likelihood maximization procedure converges to the correct kinetic rates within 100 cycles. Figure 4.4.3 shows the convergence of the rate estimates from one of the maximization runs.

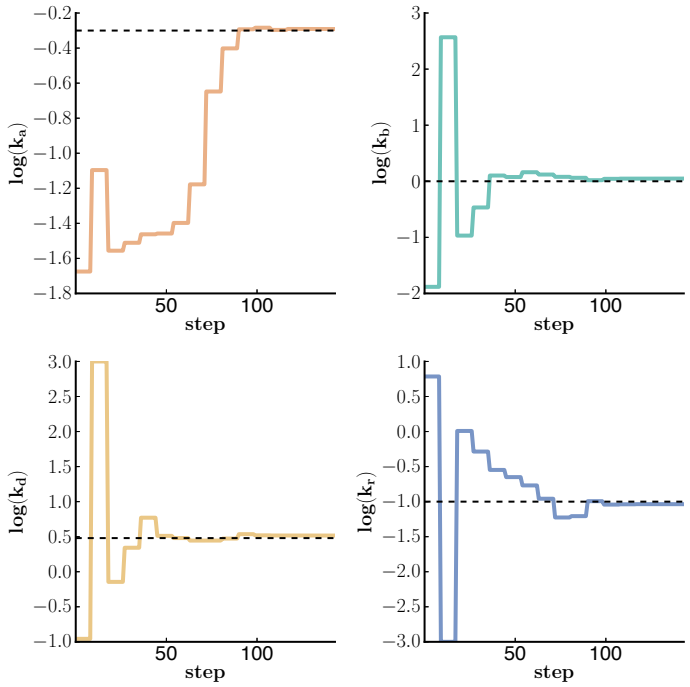


Figure 4.4.3: Convergence of likelihood maximization of data set A. The maximization converges close to true the parameter values. The parameters are estimated from the results of many independent maximization runs, like the one depicted here. In these runs, $N = 5$.

We used a bootstrapping approach (resampling data with replacement) to determine the precision of our parameter estimates.[202,203] We randomly selected a subset of trajectories and determined the rates that maximized the sum of the log-likelihoods of the selected trajectories. We constructed a distribution of rates by repeating this process for other randomly selected sets of trajectories. Our estimate for each parameter is the mean of the corresponding distribution, and we compute the 95% confidence interval based on percentiles of the distribution.

Now, we discuss the results for simulated fast (set 2) and slow blinking (set 3). We found that, like with set 1, convergence of the likelihood maximization occured within 100 cycles (Fig. 4.4.5 and Fig. 4.4.6). The boostrap results show that the parameters were determined precisely (Fig. 4.4.7 and Fig. 4.4.8), but we observed a small bias toward slower k_d in set 2 (Fig. 4.4.7). This set has a faster blinking rate than the other two sets.

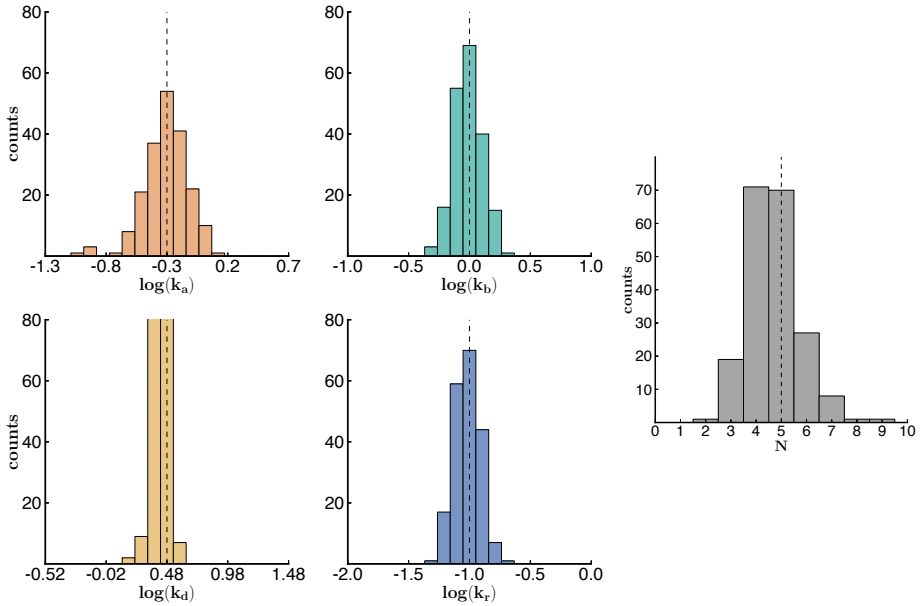


Figure 4.4.4: Histogram of bootstrapping results from simulated data set A. We fitted 200 bootstraps of the data with 10 time traces per bootstrap.

Given that we've imposed a 50 ms temporal resolution on our simulated data (to mimic the resolution of PALM experiments), it's possible that the bias is due to missed blinking events. It's also possible that this causes bias toward smaller N that we observe for this data set. Even so, our estimated k_d is correct to within a factor of two.

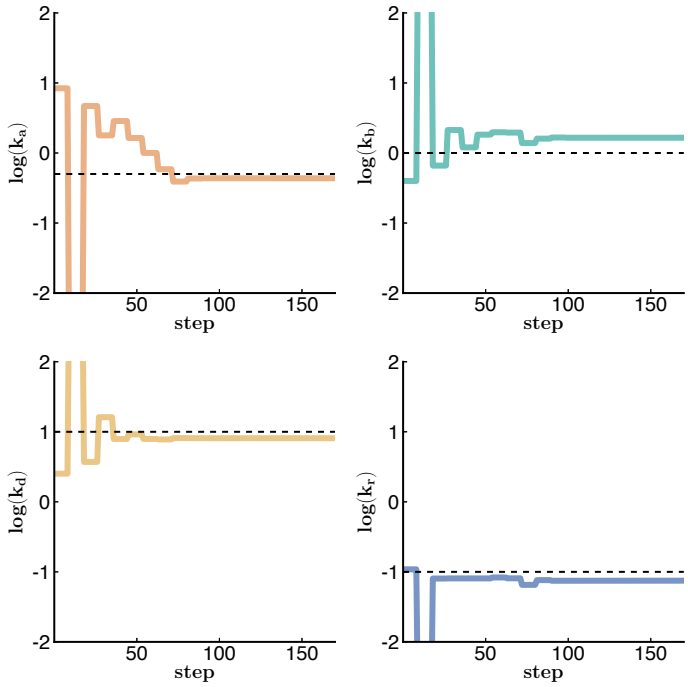


Figure 4.4.5: Convergence of likelihood maximization of simulated data set B. In these runs, $N = 5$.

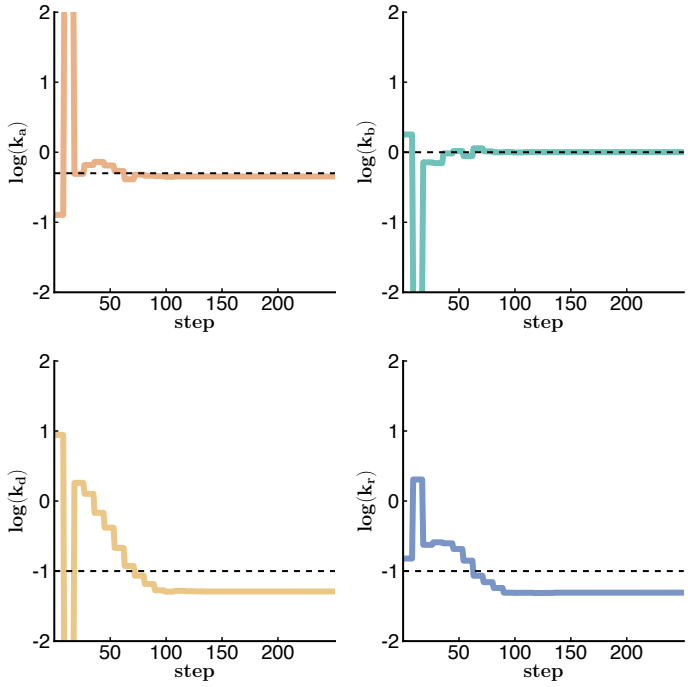


Figure 4.4.6: Convergence of likelihood maximization of simulated data set C. We observe a bias toward slower rates for k_d and k_r in this particular maximization run. In these runs, $N = 5$.

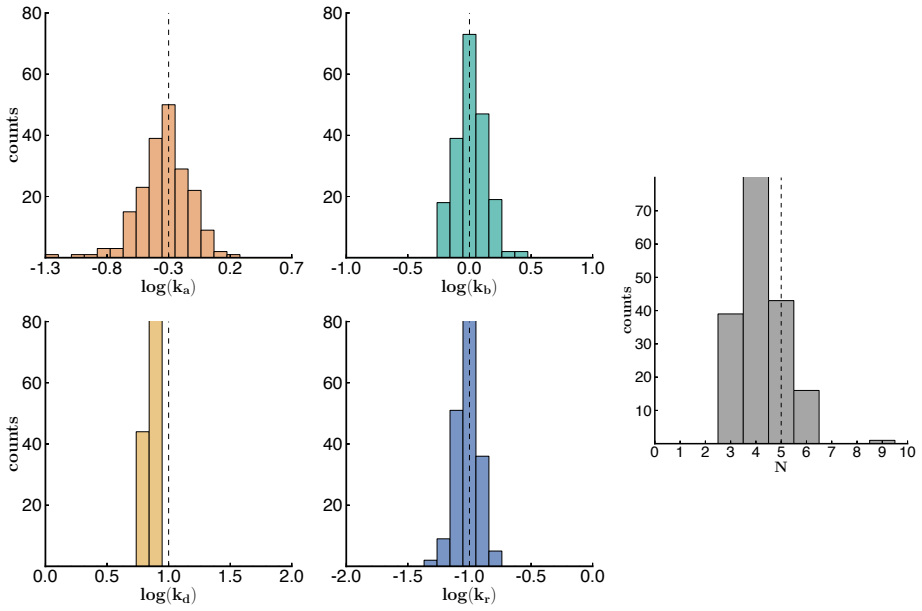


Figure 4.4.7: Histogram of bootstrapping results from simulated data set B. We fitted 200 bootstraps of the data with 5 time traces per bootstrap. An overall bias toward slow k_d is observed in the k_d distribution.

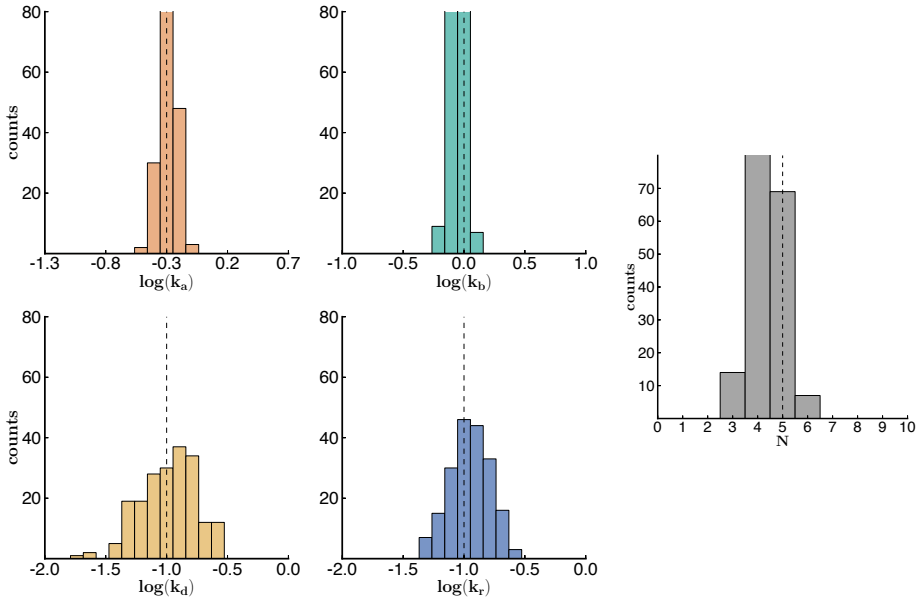


Figure 4.4.8: Histogram of bootstrapping results from simulated data set C. We fitted 200 bootstraps of the data with 40 time traces per bootstrap.

4.4.2 Analysis of *in vitro* data

In addition to the simulated data presented above, we analyzed the *in vitro* data of Lee et al.[235] In this data set, biotinylated Dendra2 molecules were immobilized on a streptavidin-coated glass coverslip. The sample was illuminated with a 405 nm laser to photoactivate the Dendra2 and then excited with a 561 nm laser until the molecules were photobleached (see Lee et al.[235] for further details). Individual emission bursts from the EMCCD output (Fig. 4.1.1A) were processed into single molecule time traces (Fig. 4.1.1B) for analysis.

We simultaneously extracted the four kinetic rates (k_a, k_d, k_r, k_b) and found that Dendra2 blinking is slow ($k_d/k_b \approx 0.3$); its behavior most closely resembles that of set 3 of our simulated data. Dendra2 molecules are more likely to photobleach upon activation than blink. Our N distribution is sharply peaked at 1 (Fig. 4.4.9), as expected for this data set, since the experiments were designed to separate and isolate Dendra2 molecules on the coverslip. Our rate estimates compare well with those of Lee et al., who found a similar blinking rate (Table 4.4.2). Our results differ from their results most significantly for k_r, the rate of recovery from D to A. They fit the distribution of fluoresence-off times to determine k_r and found that the distribution fit poorly to a single exponential, but that it was well fit by a double exponential ($k_{r1}e^{-k_{r1}t} + \alpha k_{r2}e^{-k_{r2}t}$). The poor fit to a single exponential agrees with the fact that our histogram of k_r is heavy tailed towards slower rates (Fig. 4.4.9). Their findings and ours suggest that perhaps the kinetic model of Dendra2 (Fig. 4.2.1) should be amended to two blinked states, rather than one. Fitting the data with alternative kinetic models will be included in future work.

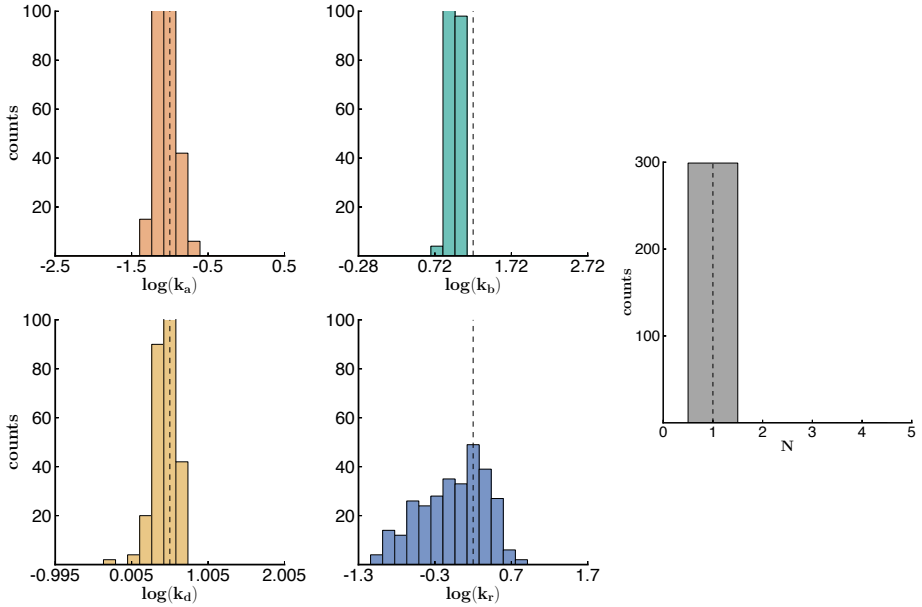


Figure 4.4.9: Histogram of bootstrapping results from *in vitro* data. We fitted 300 bootstraps of the data with 50 time traces per bootstrap. The dashed lines show the parameter values from Lee et al.[235]

Table 4.4.2: *in vitro* Dendra2 kinetic rates from the current AMM analysis and from Lee. Units: s^{-1}

Rate constant	Our analysis	Lee et al.[235]
k_a	0.009	0.01
k_d	2.8	3.2
k_r	0.87	1.6
k_{r2}	n/a	18, $\alpha = 3.2$
k_b	9.2	16.6
k_d/k_b	0.3	0.2

4.4.3 Analysis of *in vivo* data

The ultimate goal of our work is to enable accurate *in vivo* molecular counting. To that end, we also analyzed the data of Dendra2 fused to the flagellar motor protein FliM in *E. coli* from Lee et al.[235] (Fig. 4.4.10). The PALM data for FliM-Dendra2 was acquired by continuous illumination of a 561 nm excitation laser, but the 405 nm activation laser power was modulated according to a "Fermi" activation scheme. Fermi activation refers to the gradual ramping of laser power to maximize the temporal separation of PA-FP activation events. Under the Fermi protocol, the activation rate is given by:

$$k_a(t) = \frac{1}{T}\frac{e^{(t-t_F/T)}}{[1+e^{(t-t_F/T)}]log[1+e^{(t-t_F/T)}]} \tag{4.7}$$

We simulated the Fermi activation protocol by recomputing k_a according to equation 4.7 for each dwell, i.e. for each G matrix in the likelihood function (equation 4.6). We used the parameters reported by Lee et al.: $t_F = 2.2$ min and $T = 12$ sec. A sample of the FliM time traces are shown in Fig. 4.4.11. Total photobleaching occurs within four minutes.

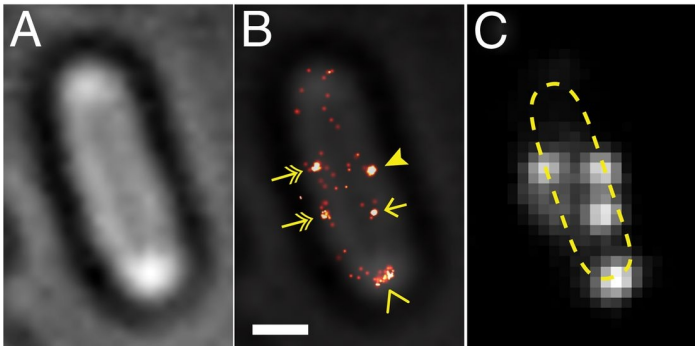


Figure 4.4.10: (A) Bright-field image of bacterial cell. (B) FliM-Dendra2 PALM overlay image. The motor proteins mainly localize as clusters at the cell membrane. The cluster indicated with a solid arrow head was selected for molecular counting. Other clusters that were elongated, located at the cell pole, not on the membrane, or surrounded by dispersed molecules were not selected for counting. (Scale bar, 500 nm) (C) Diffraction-limited FliM-Dendra2 image. Reprinted with permission from Lee et al.[235] Copyright (2012) National Academy of Sciences, USA. See page 99 for license information.

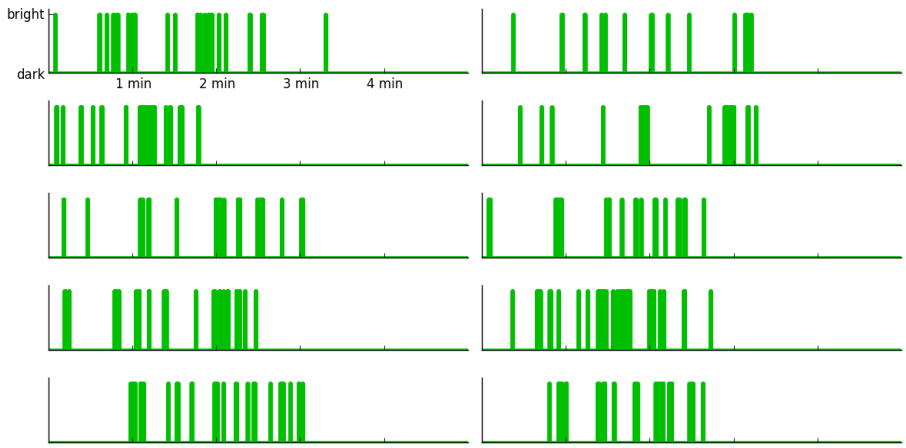


Figure 4.4.11: Ten FliM time traces out of the eighty-eight that were analyzed. Data was collected in *E. coli* cells with a Fermi activation protocol to facilitate well-spaced activation events.

Using the kinetic rates we extracted from the *in vitro* data in the previous section (Table 4.4.2), we computed the likelihood as a function of N for the FliM data. The center of our N distribution is at $N = 15$ (Fig. 4.4.12), roughly half of the expected value. Cryo-EM studies found that a mature copy of the flagellar motor has 34 copies of FliM.[246,247] Another study found 30 ± 6 using a photobleaching approach.[248] Lee et al. analyzed the same data with their thresholding method and found a value for N that agrees well with the Cryo-EM and FRAP results.

There are several factors that could cause the undercounting that we observe. First, the current analysis doesn't account for missed events. Any events faster than 50 ms will be missed. In the SI, we discuss a method for correcting the likelihood function for missed events. Second, the current FliM analysis assumes the *in vitro* kinetic rates are the same *in vivo*. We are currently working on determining the rates from the *in vivo* data. Finally, the raw data is very noisy and prone to drift. The current strategy for converting the raw data into dwell-based time traces uses simple intensity thresholding. An improvement would be to detrend the data to account for drift and apply one of the many piece wise constant (PWC) denoising methods that exist in the literature (recently reviewed by Little

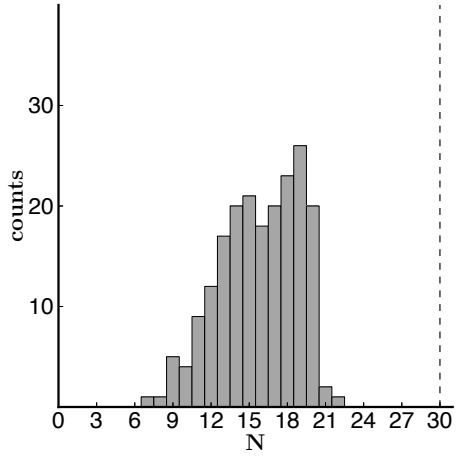


Figure 4.4.12: N was estimated from 200 independent likelihood maximizations, using kinetic rates extracted from *in vitro* data. The dashed line indicates the expected number of FliM subunits, based on independent experimental data.

and Jones[249,250]). These improvements will be the focus of future work.

4.5 Conclusions

We have adapted maximum likelihood aggregated Markov methods, originally developed for studying ion channel gating dynamics, to the PALM counting problem. Our approach accounts for fluorescent protein blinking in a robust way that doesn't rely on thresholds, doesn't require advance knowledge of kinetic rates, and isn't limited to models with only one blinked state. We've successfully used the method to extract kinetic rates and count the number of molecules accurately and precisely in simulated and *in vitro* data. However, our current *in vivo* results suggest that we undercount the number of FliM molecules in bacterial flagellar motors. Future work will focus on denoising the *in vivo* data and accounting for missed events in order to address the undercounting discrepancy.

4.6 Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship. We thank S.H. Lee, J.Y. Shin, A. Lee, and C. Bustamante for providing the Dendra2 *in vitro* and FliM *in vivo* data, and for numerous helpful discussions about PALM and the blinking problem.

4.7 Supporting Information

4.7.1 How to account for missed transitions

We consider a dwell of length t in the bright state. As shown in equation 4.8, the probability density for a dwell of length t in the bright class followed by a transition to the dark class is

$$\mathbf{G_{bd}}(t) = e^{\mathbf{Q_{bb}}t}\mathbf{Q_{bd}}. \tag{4.8}$$

Following work by Qin et al.[218] and Roux and Sauvé,[215] we assume, for sake of concreteness, that within this dark dwell we miss rapid transitions to the bright state. We could just as well have assumed that within the bright dwell we miss rapid transitions to the dark state. In order to be missed, the transitions must be shorter than some time dead time or acquisition time t_d.

We define a new probability density which is the sum over all possible missed transitions:

$$\tilde{\mathbf{G_{bd}}}(t) = \sum_{n=0}^{\infty} \Gamma(n) \tag{4.9}$$

where $\Gamma(n)$ is the probability density for n missed transitions where, in general, we have

assumed that n runs from 0 to infinity. $\Gamma(n)$, is defined as follows

$$\Gamma(n) = \int' d\tau_1 \cdots \tau_{2n+1} \left(\prod_{i=1}^{n} \mathbf{G_b d}(\tau_{2i-1}) \mathbf{G_d b}(\tau_{2i}) \right) \mathbf{G_b d}(\tau_{2n+1}). \tag{4.10}$$

where the integral over time accounts for the fact that we must sum over all possible times during which the transitions can occur within the dwell t. The prime on the integral indicates that there are restrictions in taking this integral. These conditions are the following:

$$\sum_i \tau_i = t \tag{4.11}$$

$$\tau_1 > t_d \tag{4.12}$$

$$\tau_{2i} \leq t_d. \tag{4.13}$$

The first condition says that all dwell times within t must eventually last a total time t. The second condition says that τ_1 must exceed t_d (otherwise the first missed transition would be lumped in with the previous dark dwell). The third condition simply states that, in order to be missed, all transitions to the dark class must be shorter than t_d.

In the limit that the amount of time spent in missed state (in this case the dark state) is much smaller than the total t (i.e. the time spent in the bright state including missed events), then the integral given by equation 4.10 simplifies considerably. The transition probability $\tilde{\mathbf{G}}_{bd}(t)$ then becomes

$$\tilde{\mathbf{G_{bd}}}(t) = e^{\mathbf{Q_{bb}} t_d} e^{(\mathbf{Q_{bb}} - \mathbf{Q_{bd}}(1 - e^{\mathbf{Q_{bb}} t_d})\mathbf{Q_{db}}^{-1}\mathbf{Q_{db}})(t - t_d)} \mathbf{Q_{bd}} \tag{4.14}$$

References

[1] Anfinsen C (1973) Principles that govern folding of protein chains. *Science* 181:223–230.

[2] Kendrew J, et al. (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* 181:662–666.

[3] Perutz M, et al. (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution by X-ray analysis. *Nature* 185:416–422.

[4] Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J (2007) The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.* 8:319–330.

[5] Batey S, Nickson AA, Clarke J (2008) Studying the folding of multidomain proteins. *HFSP J.* 2:365–377.

[6] Uversky VN, Dunker AK (2010) Understanding protein non-folding. *BBA-Proteins Proteomics* 1804:1231–1264.

[7] White S, Wimley W (1999) Membrane protein folding and stability: Physical principles. *Annu. Rev. Biophys. Biomolec. Struct.* 28:319–365.

[8] Arinaminpathy Y, Khurana E, Engelman DM, Gerstein MB (2009) Computational analysis of membrane proteins: the largest class of drug targets. *Drug Discov. Today* 14:1130–1135.

[9] Landry S, Gierasch L (1994) Polypeptide interactions with molecular chaperones and their relationship to *in vivo* protein folding. *Annu. Rev. Biophys. Biomolec. Struct.* 23:645–669.

[10] Horwich AL, Fenton WA (2009) Chaperonin-mediated protein folding: using a central cavity to kinetically assist polypeptide chain folding. *Q. Rev. Biophys.* 42:83–116.

[11] Dobson C (2003) Protein folding and misfolding. *Nature* 426:884–890.

[12] Balch WE, Morimoto RI, Dillin A, Kelly JW (2008) Adapting proteostasis for disease intervention. *Science* 319:916–919.

[13] Hartl FU, Bracher A, Hayer-Hartl M (2011) Molecular chaperones in protein folding and proteostasis. *Nature* 475:324–332.

[14] Lumry R, Biltonen R, Brandts J (1966) Validity of 2-state hypothesis for conformational transitions of proteins. *Biopolymers* 4:917–944.

[15] Dill K (1990) Dominant forces in protein folding. *Biochemistry* 29:7133–7155.

[16] Robertson A, Murphy K (1997) Protein structure and the energetics of protein stability. *Chem. Rev.* 97:1251–1267.

[17] Baldwin RL (2007) Energetics of protein folding. *J. Mol. Biol.* 371:283–301.

[18] Dill KA, Ghosh K, Schmit JD (2011) Physical limits of cells and proteomes. *Proc. Natl. Acad. Sci. U. S. A.* 108:17876–17882.

[19] Pace C (2009) Energetics of protein hydrogen bonds. *Nat. Struct. Mol. Biol.* 16:681 – 682.

[20] Dobson C, Sali A, Karplus M (1998) Protein folding: A perspective from theory and experiment. *Angew. Chem.-Int. Edit.* 37:868–893.

[21] Levinthal C (1968) Are there pathways for protein folding? *J. Chim. Phys.-Chim. Biol.* 65:44–45.

[22] Tsong T, Baldwin R, Elson E (1971) Sequential unfolding of ribonuclease-A - detetection of a fast initial phase in kinetics of unfolding. *Proc. Natl. Acad. Sci. U. S. A.* 68:2712–2715.

[23] Ikai A, Tanford C (1971) Kinetic evidence for incorrectly folded intermediate states in refolding of denatured proteins. *Nature* 230:100–102.

[24] Wetlaufe D (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U. S. A.* 70:697–701.

[25] Kim P, Baldwin R (1982) Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.* 51:459–489.

[26] Udgaonkar J, Baldwin R (1988) NMR evidence for an early framework intermediate on the folding pathway of ribonuclease-A. *Nature* 335:694–699.

[27] Kim P, Baldwin R (1990) Intermediates in the folding reactions of small proteins. *Annu. Rev. Biochem.* 59:631–660.

[28] Weissman J, Kim P (1991) Reexamination of the folding of BPTI - predominance of native intermediates. *Science* 253:1386–1393.

[29] Weissman J, Kim P (1992) Kinetic role of nonnative species in the folding of bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* 89:9900–9904.

[30] Baldwin R (1995) The nature of protein folding pathways - the classical versus the new view. *J. Biomol. NMR* 5:103–109.

[31] Dill K, Chan H (1997) From Levinthal to pathways to funnels. *Nat. Struct. Biol.* 4:10–19.

[32] Onuchic J, LutheySchulten Z, Wolynes P (1997) Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545–600.

[33] Wales D (2003) *Energy landscapes*, Cambridge molecular science (Cambridge University Press).

[34] Leopold P, Montal M, Onuchic J (1992) Protein folding funnels - a kinetic approach to the sequence structure relationship. *Proc. Natl. Acad. Sci. U. S. A.* 89:8721–8725.

[35] Bryngelson J, Onuchic J, Socci N, Wolynes P (1995) Funnels, pathways, and the energy landscape of protein-folding - a synthesis. *Proteins* 21:167–195.

[36] Krishna MMG, Englander SW (2007) A unified mechanism for protein folding: predetermined pathways with optional errors. *Protein Sci.* 16:449–464.

[37] Myers J, Oas T (2002) Mechanisms of fast protein folding. *Annu. Rev. Biochem.* 71:783–815.

[38] Borgia A, Williams PM, Clarke J (2008) Single-molecule studies of protein folding. *Annu. Rev. Biochem.* 77:101–125.

[39] Maxwell K, et al. (2005) Protein folding: Defining a standard set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci.* 14:602–616.

[40] Itzhaki L, Otzen D, Fersht A (1995) The structure of the transition state for folding of chymotrypsin inhibitor-2 analyzed by protein engineering methods - evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* 254:260–288.

[41] Matouschek A, Kellis J, Serrano L, Fersht A (1989) Mapping the transition state and pathway of protein folding by protein engineering. *Nature* 340:122–126.

[42] Fersht A, et al. (1985) Hydrogen-bonding and biological specificity analyzed by protein engineering. *Nature* 314:235–238.

[43] Jackson S, Fersht A (1991) Folding of chymotrypsin inhibitor-2 .2. Influence of proline isomerization on the folding kinetics and thermodynamic characterization of the transition-state of folding. *Biochemistry* 30:10436–10443.

[44] Riddle D, et al. (1999) Experiment and theory highlight role of native state topology in SH3 folding. *Nat. Struct. Biol.* 6:1016–1024.

[45] Chiti F, et al. (1999) Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.* 6:1005–1009.

[46] Kragelund B, et al. (1999) The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. *Nat. Struct. Biol.* 6:594–601.

[47] McCallister E, Alm E, Baker D (2000) Critical role of beta-hairpin formation in protein G folding. *Nat. Struct. Biol.* 7:669–673.

[48] Kim D, Fisher C, Baker D (2000) A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* 298:971–984.

[49] Guerois R, Serrano L (2000) The SH3-fold family: Experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* 304:967–982.

[50] Mayor U, Johnson C, Daggett V, Fersht A (2000) Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc. Natl. Acad. Sci. U. S. A.* 97:13518–13522.

[51] Fowler S, Clarke J (2001) Mapping the folding pathway of an immunoglobulin domain: Structural detail from phi value analysis and movement of the transition state. *Structure* 9:355–366.

[52] Kuhlman B, O'Neill J, Kim D, Zhang K, Baker D (2002) Accurate computer-based design of a new backbone conformation in the second turn of protein L. *J. Mol. Biol.* 315:471–477.

[53] Gianni S, et al. (2003) Unifying features in protein-folding mechanisms. *Proc. Natl. Acad. Sci. U. S. A.* 100:13286–13291.

[54] Banachewicz W, Religa TL, Schaeffer RD, Daggett V, Fersht AR (2011) Malleability of folding intermediates in the homeodomain superfamily. *Proc. Natl. Acad. Sci. U. S. A.* 108:5596–5601.

[55] Nauli S, Kuhlman B, Baker D (2001) Computer-based redesign of a protein folding pathway. *Nat. Struct. Biol.* 8:602–605.

[56] Alm E, Baker D (1999) Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. U. S. A.* 96:11305–11310.

[57] Munoz V, Eaton W (1999) A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. U. S. A.* 96:11311–11316.

[58] Galzitskaya O, Finkelstein A (1999) A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. U. S. A.* 96:11299–11304.

[59] Alm E, Morozov A, Kortemme T, Baker D (2002) Simple physical models connect theory and experiment in protein folding kinetics. *J. Mol. Biol.* 322:463–476.

[60] Merlo C, Dill K, Weikl T (2005) phi values in protein-folding kinetics have energetic and structural components. *Proc. Natl. Acad. Sci. U. S. A.* 102:10171–10175.

[61] Naganathan AN, Munoz V (2010) Insights into protein folding mechanisms from large scale analysis of mutational effects. *Proc. Natl. Acad. Sci. U. S. A.* 107:8611–8616.

[62] Roder H, Elove G, Englander S (1988) Structural characterization of folding intermediates in cytochrome-c by H-exchange labeling and proton NMR. *Nature* 335:700–704.

[63] Bai Y, Milne J, Mayne L, Englander S (1993) Primary structure effects on peptide group hydrogen-exchange. *Proteins* 17:75–86.

[64] Sosnick T, Mayne L, Hiller R, Englander S (1994) The barriers in protein folding. *Nat. Struct. Biol.* 1:149–156.

[65] Bai Y, Sosnick T, Mayne L, Englander S (1995) Protein folding intermediates - native state hydrogen exchange. *Science* 269:192–197.

[66] Chamberlain A, Handel T, Marqusee S (1996) Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH. *Nat. Struct. Biol.* 3:782–787.

[67] Englander S (2000) Protein folding intermediates and pathways studied by hydrogen exchange. *Annu. Rev. Biophys. Biomolec. Struct.* 29:213–238.

[68] Rumbley J, Hoang L, Mayne L, Englander S (2001) An amino acid code for protein folding. *Proc. Natl. Acad. Sci. U. S. A.* 98:105–112.

[69] Wildes D, Marqusee S (2004) Hydrogen-exchange strategies applied to energetics of intermediate processes in protein folding. *Methods Enzymol.* 380:328–349.

[70] Englander SW, Mayne L, Krishna MMG (2007) Protein folding and misfolding: mechanism and principles. *Q. Rev. Biophys.* 40:287–326.

[71] Skinner JJ, Lim WK, Bedard S, Black BE, Englander SW (2012) Protein dynamics viewed by hydrogen exchange. *Protein Sci.* 21:996–1005.

[72] Wales T, Engen J (2006) Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrom. Rev.* 25:158–170.

[73] Hu W, et al. (2013) Stepwise protein folding at near amino acid resolution by hydrogen exchange and mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.*

[74] Schuler B, Eaton WA (2008) Protein folding studied by single-molecule FRET. *Curr. Opin. Struct. Biol.* 18:16–26.

[75] Schuler B, Hofmann H (2013) Single-molecule spectroscopy of protein folding dynamics-expanding scope and timescales. *Curr. Opin. Struct. Biol.* 23:36–47.

[76] Zoldak G, Rief M (2013) Force as a single molecule probe of multidimensional protein energy landscapes. *Curr. Opin. Struct. Biol.* 23:48–57.

[77] Borgia MB, et al. (2011) Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature* 474:662–U142.

[78] Chung HS, McHale K, Louis JM, Eaton WA (2012) Single-Molecule Fluorescence Experiments Determine Protein Folding Transition Path Times. *Science* 335:981–984.

[79] Carrion-Vazquez M, et al. (1999) Mechanical and chemical unfolding of a single protein: A comparison. *Proc. Natl. Acad. Sci. U. S. A.* 96:3694–3699.

[80] Elms PJ, Chodera JD, Bustamante C, Marqusee S (2012) The molten globule state is unusually deformable under mechanical force. *Proc. Natl. Acad. Sci. U. S. A.* 109:3796–3801.

[81] Jagannathan B, Elms PJ, Bustamante C, Marqusee S (2012) 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.

[82] Stigler J, Ziegler F, Gieseke A, Gebhardt JCM, Rief M (2011) The Complex Folding Network of Single Calmodulin Molecules. *Science* 334:512–516.

[83] Shank EA, Cecconi C, Dill JW, Marqusee S, Bustamante C (2010) The folding cooperativity of a protein is controlled by its chain topology. *Nature* 465:637–640.

[84] Fersht A (1997) Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* 7:3–9.

[85] Daggett V, Fersht A (2003) Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* 28:18–25.

[86] Baldwin R (1989) How does protein folding get started? *Trends Biochem.Sci.* 14:291–294.

[87] Karplus M, Weaver D (1979) Diffusion-collision model for protein folding. *Biopolymers* 18:1421–1437.

[88] Karplus M, Weaver D (1994) Protein folding dynamics: The diffusion-collision model and experimental data. *Protein Sci.* 3:650–668.

[89] Dill K, Bromberg S (2011) *Molecular driving forces: statistical thermodynamics in chemistry, physics, biology, and nanoscience* (Garland Science), 2nd edition.

[90] Kiefhaber T (1995) Kinetic traps in lysozyme folding. *Proc. Natl. Acad. Sci. U. S. A.* 92:9029–9033.

[91] Wildegger G, Kiefhaber T (1997) Three-state model for lysozyme folding: Triangular folding mechanism with an energetically trapped intermediate. *J. Mol. Biol.* 270:294–304.

[92] Sanchez I, Kiefhaber T (2003) Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding. *J. Mol. Biol.* 325:367–376.

[93] Plaxco K, Simons K, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* 277:985–994.

[94] Zhou H, Zhou Y (2002) Folding rate prediction using total contact distance. *Biophys. J.* 82:458–463.

[95] Gong H, Isom D, Srinivasan R, Rose G (2003) Local secondary structure content predicts folding rates for simple, two-state proteins. *J. Mol. Biol.* 327:1149–1154.

[96] Galzitskaya O, Garbuzynskiy S, Ivankov D, Finkelstein A (2003) Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins* 51:162–166.

[97] Ivankov D, et al. (2003) Contact order revisited: Influence of protein size on the folding rate. *Protein Sci.* 12:2057–2062.

[98] Ivankov D, Finkelstein A (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl. Acad. Sci. U. S. A.* 101:8942–8944.

[99] Gromiha M (2005) A statistical model for predicting protein folding rates from amino acid sequence with structural class information. *J. Chem Inf. Model.* 45:494–501.

[100] Gromiha MM, Thangakani AM, Selvaraj S (2006) FOLD-RATE: prediction of protein folding rates from amino acid sequence. *Nucleic Acids Res.* 34:W70–W74.

[101] Ouyang Z, Liang J (2008) Predicting protein folding rates from geometric contact and amino acid sequence. *Protein Sci.* 17:1256–1263.

[102] Chang L, Wang J, Wang W (2010) Composition-based effective chain length for prediction of protein folding rates. *Phys. Rev. E* 82.

[103] Gao J, et al. (2010) Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility. *Proteins* 78:2114–2130.

[104] Guo J, Rao N, Liu G, Yang Y, Wang G (2011) Predicting Protein Folding Rates Using the Concept of Chou's Pseudo Amino Acid Composition. *J. Comput. Chem.* 32:1612–1617.

[105] Zou T, Ozkan SB (2011) Local and non-local native topologies reveal the underlying folding landscape of proteins. *Phys. Biol.* 8.

[106] Rustad M, Ghosh K (2012) Why and how does native topology dictate the folding speed of a protein? *J. Chem. Phys.* 137.

[107] Garbuzynskiy SO, Ivankov DN, Bogatyreva NS, Finkelstein AV (2013) Golden triangle for folding rates of globular proteins. *Proc. Natl. Acad. Sci. U. S. A.* 110:147–150.

[108] Ponder J, Case D (2003) Force fields for protein simulations. *Adv. Protein Chem.* 66:27+.

[109] Mackerell A, Feig M, Brooks C (2004) Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* 25:1400–1415.

[110] Case D, et al. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.* 26:1668–1688.

[111] Hornak V, et al. (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* 65:712–725.

[112] Brooks BR, et al. (2009) CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* 30:1545–1614.

[113] Kaminski G, Friesner R, Tirado-Rives J, Jorgensen W (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* 105:6474–6487 Symposium on Molecular Dynamics - The Next Millennium, COLUMBIA UNIV, NEW YORK, NEW YORK, JUN 02-03, 2000.

[114] Lindorff-Larsen K, et al. (2012) Systematic Validation of Protein Force Fields against Experimental Data. *PLoS One* 7.

[115] Beauchamp KA, Lin YS, Das R, Pande VS (2012) Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J. Chem. Theory Comput.* 8:1409–1414.

[116] Shirts M, Pande V (2005) Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J. Chem. Phys.* 122.

[117] Onufriev A, Bashford D, Case D (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* 55:383–394.

[118] Karplus M, McCammon J (2002) Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 9:646–652.

[119] Dror RO, Dirks RM, Grossman JP, Xu H, Shaw DE (2012) Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* 41:429–452.

[120] Duan Y, Kollman P (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282:740–744.

[121] Shirts M, Pande V (2000) Computing - Screen savers of the world unite! *Science* 290:1903–1904.

[122] Allen F, et al. (2001) Blue Gene: A vision for protein science using a petaflop supercomputer. *IBM Syst. J.* 40:310–327.

[123] Shaw DE, et al. (2008) Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* 51:91–97.

[124] Shaw DE, et al. (2010) Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* 330:341–346.

[125] Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How Fast-Folding Proteins Fold. *Science* 334:517–520.

[126] Voelz VA, Bowman GR, Beauchamp K, Pande VS (2010) Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1-39). *J. Am. Chem. Soc.* 132:1526+.

[127] Lane TJ, Bowman GR, Beauchamp K, Voelz VA, Pande VS (2011) Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. *J. Am. Chem. Soc.* 133:18413–18419.

[128] Noe F, Fischer S (2008) Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* 18:154–162.

[129] Bowman GR, Beauchamp KA, Boxer G, Pande VS (2009) Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* 131.

[130] Noe F, Schuette C, Vanden-Eijnden E, Reich L, Weikl TR (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U. S. A.* 106:19011–19016.

[131] Beauchamp KA, et al. (2011) MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* 7:3412–3419.

[132] Prinz JH, Keller B, Noe F (2011) Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. *Phys. Chem. Chem. Phys.* 13:16912–16927.

[133] Lane TJ, Shukla D, Beauchamp KA, Pande VS (2013) To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* 23:58–65.

[134] Dickson A, Brooks, III CL (2013) Native States of Fast-Folding Proteins Are Kinetic Traps. *J. Am. Chem. Soc.* 135:4729–4734.

[135] Van Kampen N (2007) *Stochastic processes in physics and chemistry* (Elsevier), 3rd edition.

[136] Taketomi H, Ueda Y, Gō N (1975) Studies on protein folding, unfolding, and fluctuations by computer simulation. *Int. J. Pept. Protein Res.* 7:445–459.

[137] Chan H, Dill K (1989) Intrachain loops in polymers - effects of excluded volume. *J. Chem. Phys.* 90:492–509.

[138] Chan H, Dill K (1989) Compact polymers. *Macromolecules* 22:4559–4573.

[139] Chan H, Dill K (1990) Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. U. S. A.* 87:6388–6392.

[140] Dill K, et al. (1995) Principles of protein folding - a perspective from simple exact models. *Protein Sci.* 4:561–602.

[141] Thirumalai D, Klimov D (1999) Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models. *Curr. Opin. Struct. Biol.* 9:197–207.

[142] Hoang T, Cieplak M (2000) Sequencing of folding events in Go-type proteins. *J. Chem. Phys.* 113:8319–8328.

[143] Clementi C, Nymeyer H, Onuchic J (2000) Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298:937–953.

[144] Shea J, Brooks C (2001) From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.* 52:499–535.

[145] Li L, Shakhnovich E (2001) Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations. *Proc. Natl. Acad. Sci. U. S. A.* 98:13014–13018.

[146] Karanicolas J, Brooks C (2002) The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* 11:2351–2361.

[147] Kaya H, Chan H (2003) Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: How adequate is native-centric topological modeling? *J. Mol. Biol.* 326:911–931.

[148] Zhang Z, Chan HS (2012) Transition paths, diffusive processes, and preequilibria of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* 109:20919–20924.

[149] Wang J, et al. (2012) Topography of funneled landscapes determines the thermodynamics and kinetics of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* 109:15763–15768.

[150] Ghosh K, Dill KA (2009) Theory for Protein Folding Cooperativity: Helix Bundles. *J. Am. Chem. Soc.* 131:2306–2312.

[151] Zimm B, Bragg J (1959) Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* 31:526–531.

[152] Lifson S, Roig A (1961) On the theory of helix-coil transition in polypeptides. *J. Chem. Phys.* 34:1963–1974.

[153] Munoz V, Serrano L (1994) Elucidating the folding problem of helical peptides using empirical parameters. *Nat. Struct. Biol.* 1:399–409.

[154] Sorin E, Pande V (2005) Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.* 88:2472–2493.

[155] Burton R, Myers J, Oas T (1998) Protein folding dynamics: Quantitative comparison between theory and experiment. *Biochemistry* 37:5337–5343.

[156] Myers J, Oas T (2001) Preorganized secondary structure as an important determinant of fast protein folding. *Nat. Struct. Biol.* 8:552–558.

[157] Islam S, Karplus M, Weaver D (2002) Application of the diffusion-collision model to the folding of three-helix bundle proteins. *J. Mol. Biol.* 318:199–215.

[158] Islam S, Karplus M, Weaver D (2004) The role of sequence and structure in protein folding kinetics: The diffusion-collision model applied to proteins L and G. *Structure* 12:1833–1845.

[159] Zwanzig R, Szabo A, Bagchi B (1992) Levinthal's paradox. *Proc. Natl. Acad. Sci. U. S. A.* 89:20–22.

[160] Zwanzig R (1995) Simple model of protein folding kinetics. *Proc. Natl. Acad. Sci. U. S. A.* 92:9801–9804.

[161] Munoz V, Thompson P, Hofrichter J, Eaton W (1997) Folding dynamics and mechanism of beta-hairpin formation. *Nature* 390:196–199.

[162] Henry E, Eaton W (2004) Combinatorial modeling of protein folding kinetics: free energy profiles and rates. *Chem. Phys.* 307:163–185.

[163] Bruscolini P, Pelizzola A (2002) Exact solution of the Munoz-Eaton model for protein folding. *Phys. Rev. Lett.* 88.

[164] Bruscolini P, Naganathan AN (2011) Quantitative Prediction of Protein Folding Behaviors from a Simple Statistical Model. *J. Am. Chem. Soc.* 133:5372–5379.

[165] Galzitskaya OV, Glyakina AV (2012) Nucleation-based prediction of the protein folding rate and its correlation with the folding nucleus size. *Proteins* 80:2711–2727.

[166] Naganathan AN, Doshi U, Munoz V (2007) Protein folding kinetics: Barrier effects in chemical and thermal denaturation experiments. *J. Am. Chem. Soc.* 129:5673–5682.

[167] De Sancho D, Munoz V (2011) Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys. Chem. Chem. Phys.* 13:17030–17043.

[168] Bicout D, Szabo A (2000) Entropic barriers, transition states, funnels, and exponential protein folding kinetics: A simple model. *Protein Sci.* 9:452–465 Spring School on Self-Organization in Biological Systems, ECOLE PHYS, LES HOUCHES, FRANCE, MAY 24-29, 1998.

[169] Makarov D, Keller C, Plaxco K, Metiu H (2002) How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. *Proc. Natl. Acad. Sci. U. S. A.* 99:3535–3539.

[170] Wallin S, Chan H (2005) A critical assessment of the topomer search model of protein folding using a continuum explicit-chain model with extensive conformational sampling. *Protein Sci.* 14:1643–1660.

[171] Weikl T, Dill K (2003) Folding rates and low-entropy-loss routes of two-state proteins. *J. Mol. Biol.* 329:585–598.

[172] Weikl T, Palassini M, Dill K (2004) Cooperativity in two-state protein folding kinetics. *Protein Sci.* 13:822–829.

[173] Weikl T, Dill K (2003) Folding kinetics of two-state proteins: Effect of circularization, permutation, and crosslinks. *J. Mol. Biol.* 332:953–963.

[174] Lane TJ, Pande VS (2012) A simple model predicts experimental folding rates and a hub-like topology. *J. Phys. Chem. B* 116:6764–6774.

[175] Baldwin R, Rose G (1999) Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem.Sci.* 24:26–33.

[176] Baldwin R, Rose G (1999) Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem.Sci.* 24:77–83.

[177] Naganathan A, Munoz V (2005) Scaling of folding times with protein size. *J. Am. Chem. Soc.* 127:480–481.

[178] Kubelka J, Henry ER, Cellmer T, Hofrichter J, Eaton WA (2008) Chemical, physical, and theoretical kinetics of an ultrafast folding protein. *Proc. Natl. Acad. Sci. U. S. A.* 105:18655–18662.

[179] Aksel T, Majumdar A, Barrick D (2011) The Contribution of Entropy, Enthalpy, and Hydrophobic Desolvation to Cooperativity in Repeat-Protein Folding. *Structure* 19:349–360.

[180] Naganathan AN (2012) Predictions from an Ising-like Statistical Mechanical Model on the Dynamic and Thermodynamic Effects of Protein Surface Electrostatics. *J. Chem. Theory Comput.* 8:4646–4656.

[181] Lane TJ, Pande VS (2012) Eigenvalues of the homogeneous finite linear one step master equation: Applications to downhill folding. *J. Chem. Phys.* 137.

[182] Baldwin RL, Frieden C, Rose GD (2010) Dry molten globule intermediates and the mechanism of protein unfolding. *Proteins* 78:2725–2737.

[183] Baldwin RL, Rose GD (2013) Molten globules, entropy-driven conformational change and protein folding. *Curr. Opin. Struct. Biol.* 23:4–10.

[184] Shaknovich E, Finkelstein A (1989) Theory of cooperative transitions in protein molecules. *Biopolymers* 28:1667–1680.

[185] Ferguson A, Liu Z, Chan HS (2009) Desolvation Barrier Effects Are a Likely Contributor to the Remarkable Diversity in the Folding Rates of Small Proteins. *J. Mol. Biol.* 389:619–636.

[186] Ghosh K, Dill KA (2009) Computing protein stabilities from their chain lengths. *Proc. Natl. Acad. Sci. U. S. A.* 106:10649–10654.

[187] Wright P, Dyson H, Lerner R (1988) Conformation of peptide fragments of proteins in aqueous solution - implications for initiation of protein folding. *Biochemistry* 27:7167–7175.

[188] Wensley BG, Gaertner M, Choo WX, Batey S, Clarke J (2009) Different Members of a Simple Three-Helix Bundle Protein Family Have Very Different Folding Rate Constants and Fold by Different Mechanisms. *J. Mol. Biol.* 390:1074–1085.

[189] Wensley BG, et al. (2010) Experimental evidence for a frustrated energy landscape in a three-helix-bundle protein family. *Nature* 463:685–U122.

[190] Wolynes P, Onuchic J, Thirumalai D (1995) Navigating the folding routes. *Science* 267:1619–1620.

[191] Ghosh K, Ozkan SB, Dill KA (2007) The ultimate speed limit to protein folding is conformational searching. *J. Am. Chem. Soc.* 129:11920–11927.

[192] Dill KA, MacCallum JL (2012) The protein folding problem, 50 years on. *Science* 338:1042–1046.

[193] Arora P, Hammes GG, Oas TG (2006) Folding mechanism of a multiple independently-folding domain protein: Double B domain of protein A. *Biochemistry* 45:12312–12324.

[194] Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* 313:903–919.

[195] Kubelka J, Hofrichter J, Eaton W (2004) The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.* 14:76–88.

[196] Young R, H B (1976) Polypeptide-chain-elongation rate in Escherichia coli B/r as a function of growth rate. *Biochem J* 160:185–194.

[197] Ewalt K, Hendrick J, Houry W, Hartl F (1997) In vivo observation of polypeptide flux through the bacterial chaperonin system. *Cell* 90:491–500.

[198] Thirumalai D (1995) From minimal models to real proteins - time scales for protein-folding kinetics. *J. Phys. I* 5:1457–1467.

[199] Gutin A, Abkevich V, Shakhnovich E (1996) Chain length scaling of protein folding time. *Phys. Rev. Lett.* 77:5433–5436.

[200] De Sancho D, Doshi U, Munoz V (2009) Protein Folding Rates and Stability: How Much Is There Beyond Size? *J. Am. Chem. Soc.* 131:2074+.

[201] Lane TJ, Pande VS (2013) Inferring the rate-length law of protein folding. arXiv:1301.4302.

[202] Efron B (1979) Bootstrap methods - another look at the jackknife. *Ann. Stat.* 7:1–26.

[203] Efron B, Tibshirani R (1991) Statistical data analysis in the computer age. *Science* 253:390–395.

[204] Voelz VA, Pande VS (2012) Calculation of rate spectra from noisy time series data. *Proteins* 80:342–351.

[205] Mukherjee S, Chowdhury P, Bunagan MR, Gai F (2008) Folding kinetics of a naturally occurring helical peptide: Implication of the folding speed limit of helical proteins. *J. Phys. Chem. B* 112:9146–9150.

[206] Neher E, Sakmann B (1976) Single-channel currents recorded from membrane of denervated frog muscle fibers. *Nature* 260:799–802.

[207] Neher E, Sakmann B, Steinbach J (1978) Extracellular patch clamp method for resolving currents through individual open channels in biological membranes. *Pflugers Arch.* 375:219–228.

[208] Hamill O, Marty A, Neher E, Sakmann B, Sigworth F (1981) Improved patch clamp techniques for high-resolution current recording from cells and cell-free membrane patches. *Pflugers Arch.* 391:85–100.

[209] Nowak L, Bregestovski P, Ascher P, Herbet A, Prochiantz A (1984) Magnesium gates glutamate-activated channels in mouse central neurons. *Nature* 307:462–465.

[210] Neher E, Sakmann B (1992) The patch clamp technique. *Sci. Am.* 266:44–51.

[211] Kienker P (1989) Equivalence of aggregated markov models of ion-channel gating. *Proc. R. Soc. B-Biol. Sci.* 236:269–309.

[212] Colquhoun D, Hawkes A (1981) On the stochastic properties of single ion channels. *Proc. R. Soc. B-Biol. Sci.* 211:205–235.

[213] Colquhoun D, Hawkes A (1982) On the stochastic properties of bursts of single ion channel openings and of clusters of bursts. *Philos. Trans. R. Soc. Lond. Ser. B-Biol. Sci.* 300:1–59.

[214] Horn R, Lange K (1983) Estimating kinetic constants from single channel data. *Biophys. J.* 43:207–223.

[215] Roux B, Sauve R (1985) A general solution to the time interval omission problem applied to single channel analysis. *Biophys. J.* 48:149–158.

[216] Ball F, Sansom M (1989) Ion-channel gating mechanisms - model identification and parameter-estimation from single channel recordings. *Proc. R. Soc. B-Biol. Sci.* 236:385–416.

[217] Colquhoun D, Hawkes A, Srodzinski K (1996) Joint distributions of apparent open and shut times of single-ion channels and maximum likelihood fitting of mechanisms. *Philos. Trans. R. Soc. A-Math. Phys. Eng. Sci.* 354:2555–2590.

[218] Qin F, Auerbach A, Sachs F (1996) Estimating single-channel kinetic parameters from idealized patch-clamp data containing missed events. *Biophys. J.* 70:264–280.

[219] Qin F, Auerbach A, Sachs F (1997) Maximum likelihood estimation of aggregated Markov processes. *Proc. R. Soc. B-Biol. Sci.* 264:375–383.

[220] Colquhoun D, Hatton C, Hawkes A (2003) The quality of maximum likelihood estimates of ion channel rate constants. *J. Physiol.-London* 547:699–728.

[221] Milescu L, Akk G, Sachs F (2005) Maximum likelihood estimation of ion channel kinetics from macroscopic currents. *Biophys. J.* 88:2494–2515.

[222] Milescu LS, Yildiz A, Selvin PR, Sachs F (2006) Extracting dwell time sequences from processive molecular motor data. *Biophys. J.* 91:3135–3150.

[223] Milescu LS, Yildiz A, Selvin PR, Sachs F (2006) Maximum likelihood estimation of molecular motor kinetics from staircase dwell-time sequences. *Biophys. J.* 91:1156–1168.

[224] Rabiner L (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* 77:257–286.

[225] Sidje R (1998) Expokit: A software package for computing matrix exponentials. *ACM Trans. Math. Softw.* 24:130–156.

[226] Moler C, Van Loan C (2003) Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.* 45:3–49.

[227] Yeo G, Edeson R, Milne R, Madsen B (1989) Superposition properties of independent ion channels. *Proc. R. Soc. B-Biol. Sci.* 238:155–170.

[228] Ball F, Milne R, Yeo G (2000) Stochastic models for systems of interacting ion channels. *IMA J. Math. Appl. Med. Biol.* 17:263–293.

[229] Lacoste T, et al. (2000) Ultrahigh-resolution multicolor colocalization of single fluorescent probes. *Proc. Natl. Acad. Sci. U. S. A.* 97:9461–9466.

[230] Qu X, Wu D, Mets L, Scherer N (2004) Nanometer-localized multiple single-molecule fluorescence microscopy. *Proc. Natl. Acad. Sci. U. S. A.* 101:11298–11303.

[231] Gordon M, Ha T, Selvin P (2004) Single-molecule high-resolution imaging with photobleaching. *Proc. Natl. Acad. Sci. U. S. A.* 101:6462–6465.

[232] Huang B, Bates M, Zhuang X (2009) Super-Resolution Fluorescence Microscopy. *Annu. Rev. Biochem.* 78:993–1016.

[233] Heilemann M, Dedecker P, Hofkens J, Sauer M (2009) Photoswitches: Key molecules for subdiffraction-resolution fluorescence imaging and molecular quantification. *Laser Photon. Rev.* 3:180–202.

[234] Sengupta P, Van Engelenburg S, Lippincott-Schwartz J (2012) Visualizing Cell Structure and Function with Point-Localization Superresolution Imaging. *Dev. Cell* 23:1092–1102.

[235] Lee SH, Shin JY, Lee A, Bustamante C (2012) Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM). *Proc. Natl. Acad. Sci. U. S. A.* 109:17436–17441.

[236] Dedecker P, De Schryver FC, Hofkens J (2013) Fluorescent Proteins: Shine on, You Crazy Diamond. *J. Am. Chem. Soc.* 135:2387–2402.

[237] Wiedenmann J, et al. (2004) EosFP, a fluorescent marker protein with UV-inducible green-to-red fluorescence conversion. *Proc. Natl. Acad. Sci. U. S. A.* 101:15905–15910.

[238] Gurskaya N, et al. (2006) Engineering of a monomeric green-to-red photoactivatable fluorescent protein induced by blue light. *Nat. Biotechnol.* 24:461–465.

[239] Fron E, et al. (2013) Revealing the Excited-State Dynamics of the Fluorescent Protein Dendra2. *J. Phys. Chem. B* 117:2300–2313.

[240] Roy A, Field MJ, Adam V, Bourgeois D (2011) The Nature of Transient Dark States in a Photoactivatable Fluorescent Protein. *J. Am. Chem. Soc.* 133:18586–18589.

[241] Coltharp C, Kessler RP, Xiao J (2012) Accurate Construction of Photoactivated Localization Microscopy (PALM) Images for Quantitative Measurements. *PLoS One* 7.

[242] Byrd R, Lu P, Nocedal J, Zhu C (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* 16:1190–1208.

[243] Zhu C, Byrd R, Lu P, Nocedal J (1997) Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.* 23:550–560.

[244] Jones E, Oliphant T, Peterson P, et al. (2001–) SciPy: Open source scientific tools for Python.

[245] Gillespie D (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81:2340–2361.

[246] Thomas D, Morgan D, DeRosier D (1999) Rotational symmetry of the C ring and a mechanism for the flagellar rotary motor. *Proc. Natl. Acad. Sci. U. S. A.* 96:10134–10139.

[247] Young H, Dang H, Lai Y, DeRosier D, Khan S (2003) Variable symmetry in Salmonella typhimurium flagellar motors. *Biophys. J.* 84:571–577.

[248] Delalez NJ, et al. (2010) Signal-dependent turnover of the bacterial flagellar switch protein FliM. *Proc. Natl. Acad. Sci. U. S. A.* 107:11347–11351.

[249] Little MA, Jones NS (2011) Generalized methods and solvers for noise removal from piecewise constant signals. I. Background theory. *Proc. R. Soc. A-Math. Phys. Eng. Sci.* 467:3088–3114.

[250] Little MA, Jones NS (2011) Generalized methods and solvers for noise removal from piecewise constant signals. II. New methods. *Proc. R. Soc. A-Math. Phys. Eng. Sci.* 467:3115–3140.

Appendix

Figure 1.1.1 was adapted with permission from Pace, 2009.[19] The license follows below:

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

May 16, 2013

This is a License Agreement between Geoffrey C Rollins ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3150910451783
License date	May 16, 2013
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Structural and Molecular Biology
Licensed content title	Energetics of protein hydrogen bonds
Licensed content author	C Nick Pace
Licensed content date	Dec 31, 1969
Volume number	16
Issue number	7
Type of Use	reuse in a thesis/dissertation
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Figures	Figure 1a
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Continuous time Markov models of the kinetics of protein folding and fluorescent protein blinking
Expected completion date	Jun 2013

Figures 4.1.1 and 4.4.10 were reproduced from Lee et al., 2012.[235] PNAS allows for reuse of figures for noncommercial purposes. The following text was downloaded on June 3rd, 2013 from `http://www.pnas.org/site/aboutpnas/rightperm.xhtml`:

> "Anyone may, without requesting permission, use original figures or tables published in PNAS for noncommercial and educational use (i.e., in a review article, in a book that is not for sale) provided that the original source and the applicable copyright notice are cited."

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

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

Geoffrey Rollin — Author Signature

6/3/13 — Date