RICE UNIVERSITY

Objective Approaches to Single-Molecule Time Series Analysis

by

James Nicholas Taylor

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

Dr. Christy F. Landes, Chair
Norman Hackerman-Welch Young Investigator, Assistant Professor of Chemistry, Rice University

Dr. Cecilia Clementi
Professor of Chemistry and Chemical and Biomolecular Engineering, Rice University

Dr. Peter J. A. Nordlander
Professor, and Director, Rice Quantum Institute, Professor of Physics and Astronomy, Rice University

HOUSTON, TEXAS
May 29, 2012

ABSTRACT

Objective Approaches to Single-Molecule Time Series Analysis

by

James Nicholas Taylor

Single-molecule spectroscopy has provided a means to uncover pathways and heterogeneities that were previously hidden beneath the ensemble average. Such heterogeneity, however, is often obscured by the artifacts of experimental noise and the occurrence of undesired processes within the experimental medium. This has subsequently caused in the need for new analytical methodologies. It is particularly important that objectivity be maintained in the development of new analytical methodology so that bias is not introduced and the results improperly characterized. The research presented herein identifies two such sources of experimental uncertainty, and constructs objective approaches to reduce their effects in the experimental results. The first, photoblinking, arises from the occupation of dark electronic states within the probe molecule, resulting in experimental data that is distorted by its contribution. A method based in Bayesian inference is developed, and is found to nearly eliminate photoblinks from the experimental data while minimally affecting the remaining data and maintaining objectivity. The second source of uncertainty is electronic shot-noise, which arises as a result of Poissonian photon collection. A method based in wavelet decomposition is constructed and applied to simulated and experimental data. It is

found that, while making only one assumption, that photon collection is indeed a Poisson process, up to 75% of the shot-noise contribution may be removed from the experimental signal by the wavelet-based procedure. Lastly, in an effort to connect model-based approaches such as molecular dynamics simulation to model-free approaches that rely solely on the experimental data, a coarse-grained molecular model of a molecular ionic fluorophore diffusing within an electrostatically charged polymer brush is constructed and characterized. It is found that, while the characteristics of the coarse-grained simulation compare well with atomistic simulations, the model is lacking in its representation of the electrostatically-driven behavior of the experimental system.

Acknowledgments

I would firstly like to express my gratitude to each and every person who has helped me along the way. However large or small your contribution, I would not have finished this dissertation without it.

I would like to thank my advisor, Dr. Christy Landes, for helping me learn what it is to be a true scientist. She has allowed me the freedom to explore and learn about Life, the Universe, and Everything in my own, unique way while still providing the direction and guidance that are so necessary in these endeavors. I especially thank her for remaining true to me during a time in my life that I would not have endured without her support. None of the ensuing work would have been accomplished without her help, and I will be forever grateful for the thoughtful leadership she has provided to me. She has been a true mentor.

I would like to thank my thesis committee, Drs. Richard Willson, Cecilia Clementi, and Peter Nordlander. Dr. Willson has been a close collaborator since the beginning, and I am grateful to have come to know him and his research. I would also like to thank Dr. Clementi for her helpful input in these last months. This dissertation would not have become what it has become without it.

I would like to thank all the members of the Landes research group, both past and present. Particularly, I would like to thank Drs. Carmen Reznik and Charlisa Daniels for countless discussions and for providing me with models of consistency and discipline throughout their graduate careers; even if I was not able to apply these models to my own life. I would also like to thank Lydia Kisley and David Cooper for allowing me to be

a pseudo-mentor to them in these last two years. They have taught me more things than they know. I would like to thank my friend Dr. Nitesh Poddar for showing me what hard work and persistence really are. Lastly, I would like to thank Bo Shuang for his invaluable help in these final weeks.

I would like to thank my parents, Paul and Debbie Gordy, for being the best parents anyone could have.

I would like to thank Dr. Stephan Link and the entirety of the Link research group. We have worked closely throughout the years, and I am grateful for your input, as my research has most definitely benefited from it.

I would like to thank Dr. Dima Makarov for teaching me how to write in a particular way, and for his invaluable input to much of my work.

Lastly, I would like to thank Paola Gil Mateo. She has inspired me both in life and in death, and has taught me many things about myself that I would not have realized had I never known her.

Contents

List of Figures

List of Tables

List of Equations

Nomenclature

FRET Fluorescence Resonance Energy Transfer

smFRET Single-Molecule FRET

AMPA α-amino-3-hydroxy-5-methyl-4-isoxazole propionate

FCS Fluorescence Correlation Spectroscopy

VEGF Vascular Endothelial Growth Factor

aV aptamer anti-VEGF DNA aptamer

PSS Polystyrene Sulfonate

MD Molecular Dynamics

KMC Kinetic Monte Carlo

HMM Hidden-Markov Model

CG Coarse-Grained

R6G Rhodamine 6G

Cy3 Cyanine 3

Cy5 Cyanine 5

"I think it is because nature has a simplicity,

and therefore a great beauty."

- Richard Feynman

for Julie

Chapter 1

Introduction and Overview

1.1 Introduction

Observing one molecule at a time, as in the experiments of single-molecule spectroscopy, removes the averaging of ensemble techniques and leads to the discovery of pathways and heterogeneities that were previously hidden beneath the ensemble average. For example, our single-molecule fluorescence resonance energy transfer (smFRET) studies of the agonist binding domain of the α-amino-3-hydroxy-5-methyl-4-isoxazole propionate (AMPA) receptor [1], described in chapter 7, reveal that the protein explores a heterogeneous conformational landscape while bound to the agonist glutamate. This indicates that the neurological membrane protein governing the passage of ions through a cellular membrane is not simply a digital gate, but rather that the exploration of this conformational landscape leads to differing extents of activation within the ion channel. These findings, however, were not immediately apparent in the results due to the effects of experimental noise. These initial results indicate a broadly

funneled landscape with no resolvable features, in disagreement with theoretical prediction [2, 3]. However, with the application of the wavelet denoising procedure [4, 5] described in chapters 5 and 6, the effects of the experimental noise were reduced, and persistent states were identified within the noisy data.

Single-molecule spectroscopy continues to advance at an astonishing rate on all scientific fronts. Such advancement is occurring within the distinct realms of system, technique, and method of analysis. Superresolution microscopy is one such technique that is of much current interest because of its promise of imaging cellular activity *in vivo* [6]. Molecular transport phenomena within many different systems, including purely polymeric constructs [7, 8] as well as electrochemical storage constructs [9-11], have been characterized as well, primarily via fluorescence correlation spectroscopy (FCS). Furthermore, the smFRET technique has been particularly useful [1, 12-14]. smFRET is a technique that gives a measure of distance on the scale of 1-10 nm. These are the distance scales that are typically relevant for inter- and intramolecular, biological interactions. Owing to this distance scale, smFRET has been used extensively for conformational dynamical studies, such as ours involving the aV aptamer [14].

The development of new experimental methodologies, as illustrated in the opening example, has subsequently caused the need for the development of new analytical methodologies. Much work has already been done in this area, such as the works of Gopich and Szabo on photon statistics in smFRET data [15, 16]. Other statistical and information theoretical methods have been applied as well, such as the implementations of Fisher information matrices to achieve optimal time resolution [17]

and positional accuracy [18], statistical correlation functions to reveal single-molecule kinetic heterogeneities [19], and hidden-Markov models to extract the most likely sequence of events from smFRET time trajectories [20, 21]. More recently, statistical correlation is combined with wavelet decomposition in attempt to describe kinetic heterogeneities in single-molecule systems [22]. Still other methods involve energy landscape theory [23], which involves the symbolization of the single-molecule time series and the subsequent extraction of the free energy landscape describing it. Lastly, we discuss another method utilizing information theory by using past events to determine the probability of future events and extract a conformational-space network that is free of *a priori* assumptions [24, 25]. Despite the relative success of these implementations, much is still left to be desired from the resolution of single-molecule experiments. Physical events in these experiments still remain hidden under guesses, optimization parameters, and the artifacts of experimental noise.

This realm of analytical methodology has been the main focus of my research. In the characterization of single-molecule results, it is particularly important to maintain an objective approach. In this work, I will describe two such approaches that were developed in my efforts to increase the information content, i.e. reduce the effects of noise, within the experimental results. To this end, we identify two types of experimental noise. The first, photoblinking, is of photophysical origin, and the other, shot noise, is an electronic error associated with the detection of photons. To identify photoblinks, an objective approach based in Bayesian inference is developed (see Chapter 4). Next, we modify the wavelet denoising method of Donoho [26] to reduce

the effects of shot-noise in the experimental photon trajectories (see chapters 5 and 6). It is found, as illustrated in the opening example, that the implementation of these objective methodologies greatly increases the amount of information available in the experimental results, thereby greatly increasing the amount of detail than can be extracted about the experimental system.

Often, such theoretical methodology falls into one of two schools of thought. Many theoretical methods, such as molecular dynamics, build the system from the ground up using microscopic first principles. Other methods, such as the information theoretical conformational space networks described in [24], characterize the system based solely on the experimental data, relying on sequences and probabilities to reconstruct the system from the top down. Another focus of my research has been to find convergence between these two, differing schools of thought. In particular, I seek to merge the two approaches by first using molecular dynamics to generate spatial information about a particular system, then using stochastic modeling to generate photophysical information, i.e., synthetic photon trajectories. Once such trajectories have been obtained, it is then possible to apply the top-down approaches mentioned above to reconstruct the model system. This convergence approach facilitates the incorporation of molecular interactions and photophysical phenomena that often plague the experiments [27-29], and results in a theoretical method capable of reproducing these effects and their manifestations at the analytical level. The approach then offers the advantages of providing a microscopically detailed model of

troublesome phenomena as well as providing a means to characterize their manifestations at the macroscopic level.

1.2 Overview of Research Aims

In this section we discuss the specific aims of my research. We begin with an experimental system involving the interaction of a small oligonucleotide with vascular endothelial growth factor. Next, we discuss the Bayesian photoblink detection method and its ramifications within simulated and experimental data. We continue with a discussion of the wavelet denoising method, and then extend the discussion to include further modifications that are specific to smFRET. We then apply these analytical methodologies to another experimental system, the AMPA receptor, and find improved experimental resolution. Lastly, we discuss a theoretical modeling approach whose goal is to find convergence between methods based on microscopic first principles and model-free reconstruction.

Research Aim 1: Identify conformational dynamics within the aV/VEGF system.

Vascular endothelial growth factor (VEGF) is a signaling protein that is known to regulate angiogenetic behavior in blood endothelial cells [30]. Hindrance of the VEGF-receptor interaction has been shown to reduce the rate of tumor growth *in vivo* [31], indicating that VEGF inhibition may be effective for disorders that are aided by neovascularization [32]. As a result, several anti-VEGF agents of diverse molecular composition, such as soluble VEGF receptors [33] VEGF-binding antibodies [31], and aptamers have been developed to inhibit angiogenesis. This research focuses on a small

DNA aptamer with high VEGF binding affinity. Aptamers are most commonly small oligonucleotides whose sequences have been iteratively optimized to recognize specific biological targets [34, 35]. Recognition of these targets can result in the disruption or enhancement of a specific biological process, thus causing aptamers to be potential therapeutics in many applications. To observe the aV/VEGF interaction, we performed smFRET experiments, as described in chapter 3, and successfully observed the interaction of the aptamer with its target protein. However, because of the large amount of experimental noise inherent to these types of experiments, we were unable to draw more specific conclusions. It became clear that more advanced methods would be required to characterize the results of this experimental system.

Research Aim 2: Objectively identify photoblinks within single-molecule trajectories.

Photoblinking arises primarily due to the triplet excited state of the fluorophores that are used in single-molecule experiments [28]. It is characterized in the data by a drop in the emission intensity of the fluorophore to zero for some amount of time, and then the return of the emission intensity to normal values. Photoblinking is particularly troublesome to smFRET measurements, where the occurrence of a photoblink causes many undesired photophysical phenomena [28]. This research aim focuses on developing a method based in Bayesian inference to identify these photoblinks. Bayesian inference is a probabilistic method that frees us from the use of subjective techniques to identify the events. The method is described in chapter 4, and is shown to identify photoblinks effectively in a broad range of simulated experimental circumstances.

Research Aim 3: Objectively reduce the effects of experimental shot-noise.

Our next research aim seeks to reduce the shot noise contribution in the experimental trajectories. Shot noise is a phenomenon of electronic origin, in that, to measure the intensity of the incident light, detectors measure an electrical current. This current fluctuates due to the discreteness of the electron, and the result is a photon trajectory containing a randomly distributed Poisson noise. The properties of the Poisson distribution are used along with a procedure involving the wavelet transformation to devise a method to reduce the shot noise contribution in the photon trajectories obtained from smFRET experiments. The wavelet transformation is a mathematical operation that transforms a function from its native domain to the wavelet domain. While in the wavelet domain, a thresholding operation is performed on the input function to reduce the noise contribution, and then the wavelet transformation is inverted to obtain a denoised version of the input function. I describe and implement the basic method on both simulated and experimental trajectories in chapter 5, and I extend the method with smFRET-specific modifications in chapter 6. It is shown that up to 75% of the shot noise contribution can be removed from the photon trajectories, thereby increasing the information content within the experimental results.

Research Aim 4: Identify conformational dynamics of the agonist binding domain of the AMPA receptor.

The wavelet denoising and photoblink filtering procedures were applied to experimental trajectories obtained from the agonist binding domain of the AMPA receptor, a neurological membrane protein that functions as an ion channel. This

receptor is a mediator of neurotransmission, and is a molecule of intense interest in the fields of learning and memory. When the extracellular agonist binding domain of the AMPA receptor binds the agonist glutamate, the membrane protein undergoes a series of conformational shifts leading to a permeable cation channel, and the channel closes upon subsequent unbinding of the agonist. Furthermore, the degree of binding cleft closure within the agonist binding domain is postulated [36] to control the extent of activation of this ion channel. The conformational dynamics of this agonist binding domain are therefore important to understand. We performed smFRET experiments to observe these conformational dynamics. Prior to denoising the smFRET trajectories, we find a smooth and featureless conformational distribution yielding little information as in the case of aV/VEGF. However, after we denoise the trajectories with the wavelet-based algorithm, we find persistent states within the conformational distributions. Specifically, the agonist bound protein was predicted to have four conformational states via molecular dynamics simulation [2, 3], which corresponds exactly to the number of states found in our single-molecule analyses, thereby providing experimental evidence for the theoretical prediction. These results are described in chapter 7.

Research Aim 5: Develop a combination model involving coarse-grained molecular dynamics and stochastic photon kinetics for simulation of single-molecule experiments.

The discrete nature of the results of single-molecule experiments causes them to often veer from the ensemble average. The source of such disagreement may lie within the realm of dynamic disorder, as in the AMPA receptor system, or may lie within the

realm of photophysics as in the pi-pi stacking interactions observed between cyanine fluorophores and unpaired nucleobases within single-stranded DNA [29]. To differentiate between such occurrences, one must know the statistical manifestation of such photophysical behavior within the experimental data so that it may be objectively and appropriately identified. To provide such insight, I develop a combination model that is comprised of both molecular and photophysical components. Specifically, a coarse-grained model of a system, described in Reznik, *et al* [9], consisting of a charged organic fluorophore, rhodamine 6G (R6G), diffusing within a charged polystyrene sulfonate (PSS) polymer brush is constructed, and stochastic Monte Carlo methods are used to generate synthetic photon trajectories as one would obtain from FCS experiments. A coarse-grained description of the excited state fluorophore is constructed and placed within a simulation volume containing a coarse-grained description of the polymer brush. Molecular dynamics trajectories are obtained, and these trajectories are then used along with a Gaussian irradiance profile to generate the synthetic trajectories. The end result is a simple, effective simulation that contains both molecular and photophysical information for the characterization and differentiation of such phenomena. The details of this research aim are discussed in chapter 8.

Chapter 2

Experimental and Analytical Methods

This chapter contains descriptions of the relevant optical and experimental techniques, as well as descriptions of the methods by which the data are analyzed and the basic concepts behind these methods. We begin by describing the optical system and the concepts of the two experimental methods that will be featured in the later chapters. We continue to describe in detail how the data are processed and categorized, and describe statistical tests used in determining whether or not a trajectory contains contamination. In the last section, we discuss the details of several analytical methods that will be prevalent in the following chapters. These methods include curve fitting methods as well as other kinetic analyses and simulations such as waiting time distributions, hidden-Markov models (HMMs) kinetic Monte Carlo simulations.

2.1 Optical Aspects of Scanning, Confocal, Single-Molecule Microscopy

2.1.1 *The Optical System*

The inverted microscope is the centerpiece of the optical system. It is equipped with an objective lens, which, for the purposes of the experiments described here, is

typically a 100x magnification objective lens with a numerical aperture of 1.3. The microscope is also equipped with a piezoelectric stage capable of 3-dimensional movement with nanometer precision. The translational scale along the two horizontal axes of the stage is 100 *μm*, and is 200 *nm* in the vertical dimension. We position a sample atop this piezoelectric stage to provide the ability to translate our sample with such 3-dimensional precision. Excitation comes from a continuous wave, 532 *nm* laser source that is often diverted through various optics, such as neutral density filters, beam profilers, and polarizing optics prior to reaching the microscope objective. After passing through the objective, the incident beam is focused within the sample volume, exciting a small volume of the sample.

Fluorescence from the system under observation within the excitation volume is collected by the same objective lens (epifluorescence), and sent along the detection path. The fluorescence first transmits through an optical filter designed to reflect wavelengths shorter than the excitation wavelength, then transmits through a notch, or band-stop, filter with wavelength range of +/- 10 *nm* around the central wavelength, 532 *nm*. After this filter, the fluorescence is imaged onto one of two detectors. The nature of this separation of the incident light onto the detectors depends on the experiment.

In the case of smFRET, the light incident on the detectors is separated by wavelength such that light at shorter wavelengths reflects from an optical filter that is designed to pass longer wavelengths. This is known as a long-pass optical filter, or equivalently, a dichroic mirror. As such, the longer wavelengths are transmitted by the

filter. We position one detector to collect the shorter wavelengths, and the other detector to collect longer wavelengths, and this enables the experiment to separately record numbers of photons emitted by the donor and acceptor fluorophores in a smFRET experiment. The first difference in FCS experiments involves the use of an optical aperture, or pinhole, in the detection path. The pinhole serves to maintain the confocal condition required by FCS by rejecting emission that is out of focus. The next difference lies in the separation of the fluorescence onto the two detectors. Instead of separating the fluorescence by wavelength, a polarizing beam splitter is used to separate the incident light into its parallel and perpendicular components. Such detection of the components of polarization allows are orientational as well as translational information to be extracted from the FCS experiments. Lastly, photons pass through bandpass filters that are designed are to reject any wavelengths that are out of range for the fluorophore(s) used in the experiment. These filters are positioned in front of one of two avalanche photodiode detectors in single-photon counting mode.

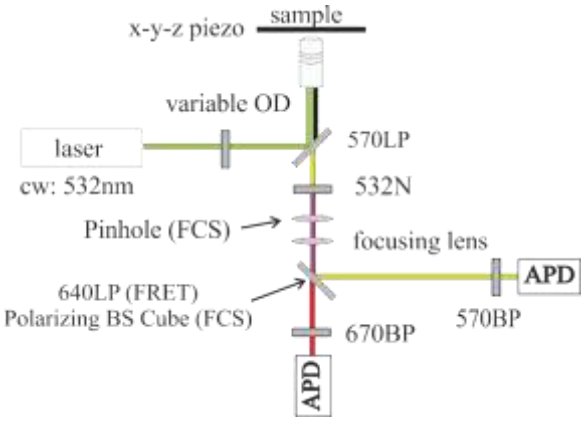


Figure 2.1. The scanning, confocal FRET/FCS setup.

2.1.2 *Concepts of FRET*

In this section we discuss the concepts of fluorescence, or Förster, resonance energy transfer. The theory was originally published in 1948 by Theodor Förster. In this

theory, the excitation energy of an electronically excited fluorophore, the donor, is transferred via a nonradiative mechanism to another fluorophore with similar electronic characteristics, the acceptor. This efficiency of this energy transfer is proportional to the square of its rate constant. This rate constant is dependent on the inverse third power of the radial distance *r* between the donor and acceptor fluorophores, so the efficiency of energy transfer is dependent on the inverse sixth power of this distance. Specifically, we have

$$E = \frac{1}{1 + (r/R_0)^6} \qquad (2.1)$$

Here, R_0 is a quantity known as the Forster radius. This is the distance, which is specific to a particular pair of fluorophores, at which the efficiency of energy transfer is at 50%. It is given by the equation

$$R_0 = \frac{9Q_D(\ln 10)\kappa^2 J(\lambda)}{128\pi^5\eta^4 N_A} \qquad (2.2)$$

The parameters involved in this equation are the quantum yield of the donor, Q_D, the dipole orientation factor, κ^2, the spectral overlap integral between donor emission and acceptor absorption, $J(\lambda)$, the refractive index of the medium, η, and Avogadro's number, N_A.

Furthermore, we may relate the efficiency of energy transfer to the intensities observed to be emitted from the acceptor, I_A, and the donor, I_D in the following manner,

$$E = \frac{I_A}{I_A + I_D} \qquad (2.3)$$

This equation allows for the determination of energy transfer efficiency in an experimental setting. Single-molecule FRET experiments are typically carried out by tagging the molecule(s) of interest with donor and acceptor fluorophores, and using an optical setup similar to the one described above. There are many complications, however. These complications include issues such as photoblinking and detector shot-noise, which will be discussed at length in the upcoming chapters.

2.1.3 *Concepts of Fluorescence Correlation Spectroscopy*

In this section we discuss the concepts of fluorescence correlation spectroscopy (FCS). FCS uses the fluorescence characteristics of some probe molecule to determine its diffusion behavior. This involves placing a sample containing the system to be studied, that contains the probe molecule, on a setup similar to the one described above and measuring fluorescent intensities as the molecule diffuses through the focal volume created by passing excitation light through the microscope objective. Construction of autocorrelation function of this intensity allows for the determination of diffusive properties within the system.

We construct this autocorrelation function $G(\tau)$ with the deviation of the local intensity from the mean intensity, $\delta I(t)$, with the equation

$$G(\tau) = \frac{\langle \delta I(t) \rangle \langle \delta I(t+\tau) \rangle}{\langle I(t) \rangle^2} \qquad (2.4)$$

Using the Gaussian irradiance profile of the $1/e^2$ irradiance volume, we may relate the intensity autocorrelation to the radius r_0 and long axis z_0 of the Gaussian ellipsoid

describing the focal volume, and to the characteristic diffusion time of the probe molecule, τ_D

$$G(\tau) = G(0)\left(1 + \frac{\tau}{\tau_D}\right)^{-1}\left(1 + \left(\frac{r_0}{z_0}\right)^2\left(\frac{\tau}{\tau_D}\right)\right)^{-1/2} \qquad (2.5)$$

This characteristic diffusion time may be related to the Stokes-Einstein diffusion constant *D*, by the relation

$$D = \frac{{r_0}^2}{4\tau_D} \qquad (2.6).$$

FCS is often used to study transport properties within various experimental systems that range from biological systems to inorganic nanoparticles to charged environments. There are useful variations to the basic method, such as those that translate the focal volume in one dimension as time evolves to measure directed transport, and others that use polarization-dependent detection to measure the 3-dimensional orientation of the probe molecule as it evolves with time.

2.2 Methods of Data Processing

The methods by which the data are processed, that is, the methods by which the data are categorized and stored, play a central role in in the accessibility of the data for analysis at a later time. Careful attention to detail in the initial stages saves considerable time in future analyses. This section describes these initial processing steps and finishes with basic calculations and statistical tests that are used to determine the validity of each trajectory in the data set.

2.2.1 *Processing smFRET Trajectories*

Typically, raw smFRET data, i.e., the experimental measurements obtained directly from the detectors, consists of two discretely binned vectors, one containing acceptor photon counts and the other containing donor photon counts. More specifically, each discrete bin contains the number of photon counts recorded on some uniform time interval. We describe these bins as acquisition bins. The time interval spanned by an acquisition bin is typically on the order of 1 ms, and will be taken to be such in this discussion. This type of discrete interval in time is sometimes called the sampling interval, or inversely equivalently, sampling frequency.

Due to the Poisson noise inherent to the detection method, these measurements must typically be sampled at a longer interval for analysis. In this process, which is sometimes referred to as 'binning up', acquisition bins that fall within a certain, longer time interval, say 10 ms, are added together to produce a single value for that particular sampling interval. These 10 ms sampling intervals are referred to as analysis bins. This action is performed to over the entire signal to produce a new signal with 10 ms sampling intervals.

Increasing the sampling interval increases the number of photons within each analysis bin. Because the noise within the acquired signal is a Poisson noise whose variance scales as the square root of the number of occurrences, an increase in the

number of photons results in a relative decrease in the noise contribution to the overall signal. A smaller noise contribution leads to more accurate calculations.

After increasing the signal's sampling interval, we can now distinguish three distinct regions within the trajectories. Fig. 2.2 shows an idealized acceptor trajectory in

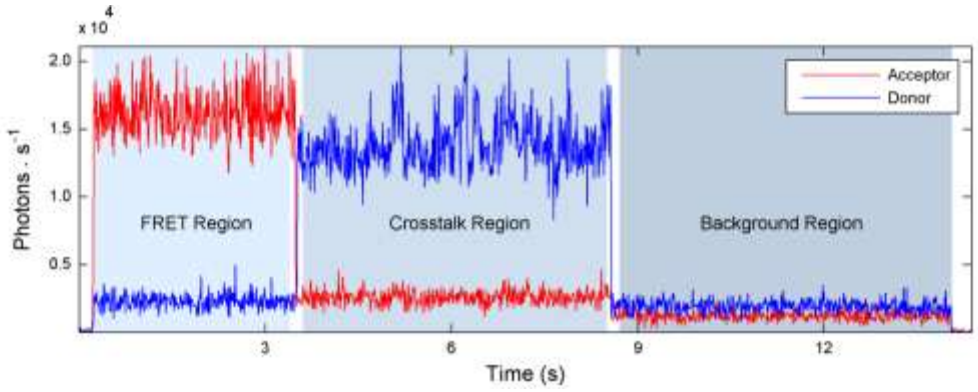


Figure 2.2. Regions in an smFRET time trajectory. The region to the far left shows the FRET region, the center region shows the crosstalk region, and to the right is the background region.

red along with its counterpart donor trajectory in blue. The shaded areas indicate the areas of the trajectory that we refer to as the background region, the crosstalk region, and the FRET region. These regions consist of three differing photophysical situations. Namely, the background region occurs after both fluorophores have photobleached, the crosstalk region after the acceptor has photobleached but while the donor is still active, and the FRET region while both fluorophores are still active.

These regions are selected manually within a custom graphical user interface (known as *fretchop*), and the trajectories are processed, or 'chopped' into their component regions. Average background and crosstalk, defined as leakage of donor photons onto the acceptor detector, values are obtained, statistical exclusion tests are performed, and FRET efficiency values are calculated at each analysis bin. Ensemble distributions of photons, efficiencies and background, are compiled as well. Output plots containing time trajectories (photons and efficiencies), histograms, and all relevant

statistics are then rendered to portable document format. Each raw trajectory, its counterpart at increased sampling interval, each FRET, background, and crosstalk region, as well as all the relevant statistics, filenames, and exclusion criterion are redundantly saved in a MATLAB file format. One copy contains information about that particular trajectory, and the other copy exists within a larger file containing all information about all trajectories within a particular set.

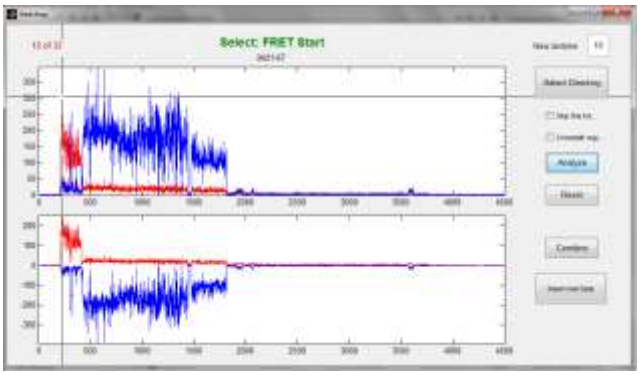


Figure 2.3. Screenshot of the FRET analysis package.

Background, crosstalk, and efficiency calculations are described in detail in Section 4.2. The statistical exclusion tests are described in the next Subsection 2.2.2. The methods discussed in the entirety of Section 2.2.1 form the framework for a FRET analysis package, that was developed and coded in MATLAB, and implemented in the graphical user interface known as *fretchop.* A screenshot of the interface is shown in Fig. 2.3.

2.2.2 *Statistical Tests for Trajectory Exclusion*

Often, a FRET trajectory will be contaminated by additional photons from an external source. This external source may be an additional fluorophore attached to a biomolecule, possibly a malfunction in the microfluidic flow system, or it could be another contaminant altogether. It is therefore prudent that we develop means to identify trajectories that contain contamination from external sources. We must,

however, avoid the introduction of bias by simply excluding those trajectories that we deem to be visibly contaminated. We must instead develop methods to identify this behavior statistically, and apply this statistical test to all trajectories within a particular set.

We identify such sets by their time and condition of acquisition. All trajectories within a set are acquired sequentially in time under the same experimental conditions within the same experimental sample. Statistical exclusion tests are not performed across differing sets of trajectories.

The first and simplest statistical test is a test of the local background measurement. If the conditions between each acquisition are taken to be equivalent, then the distribution of background measurements across a set of trajectories would be normally distributed within some standard error. Therefore, if a trajectory's background measurement is an outlier to this distribution, there is a local contamination within this trajectory, and we are provided a statistical means to exclude it from our analysis. In the case of smFRET trajectories, we determined outliers to deviate from the mean measurement by more than 2 standard deviations.

The second statistical test is based on the sum (acceptor plus donor) photon counts within the FRET region of a particular data set. The method is analogous to the statistical test for background, but the mean of the sum of acceptor and donor counts within each trajectory is used in place of the background measurement. This test is necessary for the exclusion of trajectories whose photon counts are not large enough to yield accurate calculations relative to the rest of the trajectories in the set. The

disproportionate error resulting from these trajectories skews the results of wavelet and state-finding algorithms, so it is important that these trajectories be excluded from analyses.

The third statistical test concerns the temporal length of the FRET region of the trajectories. Trajectories that are very long relative to the rest of the set have negative effects of the analyses. To avoid the over- or underrepresentation of any one molecule's behavior in the overall picture, we develop a method to exclude the very long and very short, relative to the mean of the data set, trajectories. This method is comprised of developing a cumulative exponential probability distribution based on the mean length of the FRET region of each trajectory, and then excluding very short trajectories with probability less than 5% and those with probability greater than the top 5% of this distribution. This test therefore provides a very efficient means to avoid overrepresentation of behavior within a very long trajectory that may, in fact, be rare behavior local to only the one trajectory.

2.3 Methods of Data Analysis

2.3.1 *Curve Fitting*

All fits of all data sets are nonlinear least squares regressions whose parameters have been further optimized by a Nelder-Mead unconstrained simplex optimization. In general, the method of least squares seeks to optimize a parametric model function with respect to some set of observable data. The model function represents some hypothesis about some observable data set, and the least squares regression is a test of

said hypothesis. The parameters of the model function are independent variables that are adjusted to conform, if possible, to the characteristics of the particular observable.

Most often we use a nonlinear least squares regression to avoid errors associated with linearization. An example of a nonlinear model function is an exponential function

$$m(x, \beta) = \beta_0 e^{\beta_1 x} \qquad (2.7).$$

Here, β_0 and β_1 are the parameters to be optimized, and x is the independent variable. The regression's objective is to find the values of the parameters of the model function that minimize the sum of squared residuals. The sum of squared residuals, where d_i are the observable data, is expressed as

$$S = \sum_i (d_i - m(x_i, \beta))^2 \qquad (2.8).$$

This quantity represents what is often called the unexplained variance. This quantity is "unexplained" because it represents variation in the data that isn't described by the model. In conjunction with the sum of total squared error S_{tot}

$$S_{tot} = \sum_i (d_i - \langle d \rangle)^2 \qquad (2.9),$$

we obtain the coefficient of determination reported in all fits

$$R^2 = 1 - \frac{S}{S_{tot}} \qquad (2.10).$$

2.3.2 *Kinetic Analysis via Waiting Time Distributions*

Suppose we have a time trajectory describing the occupation of two states, A and B, such as that obtained from a 2-state hidden-Markov model of a smFRET trajectory. We wish to extract the kinetic rate constants for transitions between these two states from this trajectory. One method to accomplish this is via the distributions of waiting times before transition from one state to the other. For the two state system under consideration with states A and B, we compile a histogram of the waiting times prior to transition out of state A, and similarly compile a histogram for state B. Considering that the transition lifetimes of each distribution contain only the transitions out of the particular state, any transitions leading back into the same state need not be considered in this analysis. This means that the distribution of state lifetimes will then be exponentially distributed.

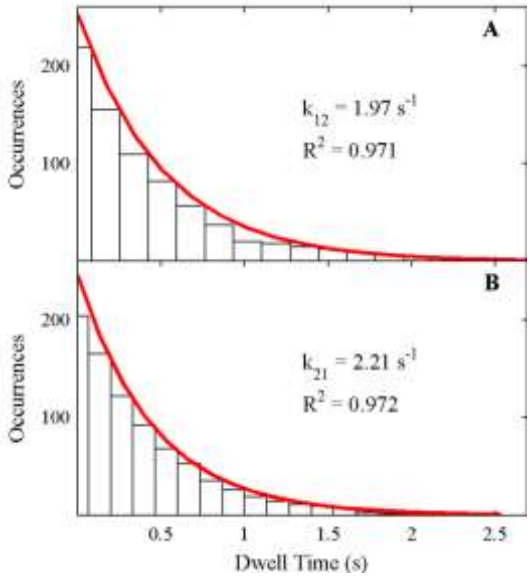


Figure 2.4. Dwell time analysis. A) shows the dwell time distribution of state A with ≈2s lifetime, and B) shows that of state B with slightly longer lifetime.

In the case of our two state system, this distribution is described by a monoexponential distribution with escape rate $k = k_A$ *or* k_B, depending on whether we are analyzing the transition rate from A to B or from B to A, respectively. We may describe this distribution with the equation

$$P(t_i) = k_i e^{-k_i t_i} dt_i \qquad (2.11).$$

Here, the subscript *i* refers to either state A or state B. This equation allows us to fit the waiting time distributions to exponential decays, and this leads to the rate constants for transitions between states. This process is illustrated in Fig. 2.4.

2.3.3 *Hidden Markov Models*

Hidden Markov modeling (HMM) is a mathematical method for determining the most likely sequence of states in a trajectory. The hidden aspect of the model is derived from the assumption that the trajectory is contaminated by noise, and that the states are hidden from the observer by this noise. The model also assumes that the kinetic transitions in the trajectory are Markovian, i.e., they are stochastic processes in which the current state of the system depends only on the state immediately prior, and not on any other past or future states. The HMM relies on a pair of probabilities that are sometimes called the transition and emission probabilities.

The transition probabilities are the probabilities that the state of the system will change from its current state to one of possibly many other states after some particular amount of time has elapsed since the last change occurred. These probabilities are related to the kinetic rates of transition among the states in the system. The other probability measure used in the HMM is the emission probability. This probability examines the likelihood that a state located at some central value in the observable measurement “emits” a particular observable data point. For example, the HMMs most commonly used in smFRET target a the probability that the occupation of a certain state, characteristically located near some central efficiency with normal probability

about that efficiency, results in the observation of a particular efficiency. These transition and emission probabilities are multiplied to give the total probability that the system transitions between two states at a particular time. HMMs were developed in the context of single-molecule time series by Andrec, *et al* [21].

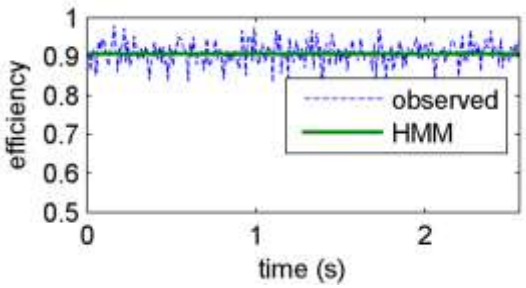


Figure 2.5. The HMM (green) acting on a noisy, stationary efficiency trajectory (blue).

In the implementation of the HMM, the most likely sequence of events is typically extracted with an algorithm such as the Viterbi algorithm [37]. Then the transition probabilities and the locations and widths of the states are used as parameters in an optimization algorithm that iterates until it converges on parameters that yield a maximum in total probability. This procedure requires that guesses as to the number of states, and to the states' locations and widths be input to the optimization algorithm. The HMM acting on a smFRET trajectory is shown as an example in Fig. 2.5.

2.3.4 *Stochastic Modeling and Kinetic Monte Carlo Simulations*

Kinetic Monte Carlo simulations are a simple and effective method to simulate the temporal behavior of a system with discrete states whose transitions obey exponential kinetics. The stochastic behavior of the waiting times between transitions is

simulated in a Metropolis Monte Carlo [38] style by selecting a uniformly distributed random number to define a particular waiting time.

Specifically, the waiting time distributions produced by the kinetic rate of transition between a particular pair of states are randomly sampled to produce a time trajectory that is comprised of a series of waiting times. In the context of smFRET trajectories for systems with discrete states, we may simulate trajectories for a system with known efficiency states and transition rate constants. We first define these efficiencies and rates then generate state-to-state trajectories via kinetic Monte Carlo methods until we meet a predetermined time cutoff. We use the efficiencies of these states to generate acceptor and donor photon trajectories. Shot-noise and background contamination are sampled into the photon trajectories according to Poisson statistics.

2.3.5 *Molecular Dynamics Simulations*

All molecular dynamics simulations were performed in GROMACS version 4.5.5 [39-42]. The Groningen machine for molecular simulation is open-source molecular modeling software that is capable of performing simulations of many different systems under many different conditions. Simulations in this work were constructed specifically for GROMACS by the production of 3 input files. The first file is a structure file containing the dimensions of the simulation volume and the coordinates for each atom included in the simulation. The second file is a topology file that defines the parameters for all bonded and nonbonded, i.e., Coulombic and Lennard-Jones forces, interactions to be used in the simulation. Lastly, the molecular dynamics parameter file contains the simulation parameters. These include the type of integration to be performed, the

number and length of time steps, physical constants such as friction coefficients, and cutoff radii.

Chapter 3

Dynamics of an anti-VEGF DNA Aptamer: A Single-Molecule Study

The contents of this chapter were adapted from an article originally published in *Biochemical and Biophysical Research Communications* on August 22, 2008.

Taylor, J.N.; Darugar, Q.D.; Kourentzi, K.; Willson, R.C.; Landes, C.F. "Dynamics of an anti-VEGF DNA Aptamer: A Single-Molecule Study." *Biochem. Biophys. Res. Comm. 2008. 373(2): 213-218.*

ABSTRACT

Single-molecule fluorescence resonance energy transfer (SMFRET) was used to study the interaction of a 25-nucleotide (nt) DNA aptamer with its binding target, vascular endothelial growth factor (VEGF). Conformational dynamics of the aptamer were studied in the absence of VEGF in order to characterize fluctuations in the unbound nucleic acid. SMFRET efficiency distributions showed that, while the aptamer favors a base-paired conformation, there are frequent conversions to higher energy conformations. Conversions to higher energy structures were also demonstrated to be

dependent on the concentration of Mg^{2+}-counterion by an overall broadening of the SMFRET efficiency distribution at lower Mg^{2+} concentration. Introduction of VEGF caused a distinct increase in the frequency of lower SMFRET efficiencies, indicating that favorable interaction of the DNA aptamer with its VEGF target directs aptamer structure towards a more open conformation.

3.1 Introduction

The emergence of aptamers as a new class of therapeutic molecules [43] has necessitated an understanding of the relationships among structure, dynamics, and function of these molecules. Aptamers are small molecules, typically oligonucleotides that have been optimized to bind specific biological targets [34, 35]. This binding can affect the biological process involving the target, thus causing aptamers to be potential therapeutics in many applications. The work of Ruckman [44] lead to the development of pegaptanib for the treatment of age-related macular degeneration [45]. Pegaptanib became known as Macugen (Eyetech, Pfizer) upon approval by the FDA in 2004, becoming the first aptamer to be approved by the FDA for therapeutic use.

VEGF is a signaling protein that is known to regulate angiogenic behavior of blood endothelial cells [30]. Hindrance of the VEGF-receptor interaction has been shown to reduce the rate of tumor growth *in vivo* [31], indicating that VEGF inhibition may be effective for disorders that are aided by neovascularization [32]. As a result, several anti-VEGF agents of diverse molecular composition, such as soluble VEGF receptors [33] and VEGF-binding antibodies [31], have been developed to inhibit angiogenesis. However, in the case of pegaptanib, it was necessary to make 2'-fluoro or

2′-methoxy substitutions to increase *in vivo* stability of the RNA aptamer [44]. DNA molecules lack this highly reactive hydroxyl group and are more resistant to nuclease degradation inside the cell. In combination with their structural similarity to RNA aptamers as well as their high specificity for binding targets [46], DNA aptamers are promising therapeutic agents.

A 25-nt DNA aptamer, with sequence 5′-CCGTCTTCCAGACAAGAGTGCAGGG-3′ and designation aV (anti-VEGF), has shown to have nanomolar binding affinity for VEGF [47]. Through a series of single-base substitutions, its binding affinity has also shown to be dependent on higher order structure in the aptamer [47]. Specifically, increased stability of the stem region results in considerable loss of VEGF-affinity [47], suggesting a relationship between structural flexibility and anti-VEGF activity of the aptamer. Because higher order structure in nucleic acids has been noted to be highly dependent on counterion concentrations [48-50], and multivalent cations in particular [51-54], it is therefore of interest to observe an experimental Mg^{2+}-dependence in the structural dynamics of the aV aptamer.

smFRET has often been used in combination with biomolecular systems [55-58]. Specifically, it has been used successfully to measure nucleic acid structural dynamics as a function of counterion-dependence [59, 60], protein interactions [61], and complementary oligonucleotides [62, 63]. Advances in sample preparation and processing algorithms can make SMFRET even more useful by reducing or eliminating complications such as shot noise [64], fluorophore photoblinking [65], and state

assignments [20, 21], thereby enhancing the ability detect the presence of specific conformations.

The immobilized aV aptamer's structural fluctuations are characterized using SMFRET. Global analyses reveal that a closed conformation is favored by the aptamer at 2 mM Mg^{2+}, but transformations to less-favored conformations are observed. As predicted, the prevalence of higher energy states increases at decreased Mg^{2+} concentration. We also observe that the aV-VEGF interaction is accompanied by a shift in the equilibrium structure towards a less-favored conformation, suggesting that association with VEGF results in a loss of base-pairing in the aV hairpin. Lifetime analyses show that the aV-VEGF complex is not associated with a specific conformation under these conditions. However, despite this observation, dissociation of the complex may be observed by a specific transition of the aptamer from an open conformation to one that is fully closed.

3.2 Materials and Methods

3.2.1 *Sample Preparation*

Standard 22x22 mm glass microscope slides were cleaned with low pressure oxygen plasma (Plasmod, March Plasma Systems), and functionalized with Vectabond (Vector Laboratories, 20µL/mL in >99.9% acetone) by immersion in solution for 5 minutes. Functionalization was quenched by immersion in molecular biology (MB) grade water (HyClone, Thermo) for 30 seconds. A 100:1 mixture of 5 kDa, methoxy-terminated, N-succinimidyl polyethylene glycol (Fluka, 33% w/w PEG in MB water) and 5

kDa biotin-terminated PEG (NOF Corporation, 2.5% w/w in MB water), buffered with $NaHCO_3$ (1% v/v, pH 8.0), was applied to each slide, atmospherically dried, and excess was rinsed with MB water.

Custom hybriwell chambers, secure seal spacers (Grace Bio-Labs), coned ports, tips, and teflon tubing (Western Analytical Products) were used to construct a flow chamber that was secured to each slide. Each slide was treated with 40 µL of 20% w/w streptavidin (Invitrogen) in 25 mM HEPES buffer (Sigma) for several minutes. A 256 pM aptamer solution (10 µL, 5 nM aV in 188 µL, 25 mM HEPES) was denatured at 80˚C, annealed at 60˚C, and cooled at 0˚C after addition of 2 µl $MgCl_2$ (1M, Sigma). The aptamer solution was then added in ten 20 µL increments, and incubated in a closed atmosphere. Typical coverslip coverage was 2-3 aptamers per 5 μm^2. Genie Plus syringe pumps (Kent Scientific) maintained flow of 2.5 mM HEPES buffer, 145 mM Na^+, 2 mM or 0.2 mM Mg^{2+}, and an O_2 scavenger – 4.3% w/w glucose, 1% w/w glucose oxidase, and 0.1% v/v catalase in MB water saturated with Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid, Sigma) [66] – through the sample chamber during experimentation. Effective VEGF concentration was 2.2 µM in all experiments.

Modifications to the aV aptamer (Trilink Biotechnologies) are illustrated in Figure 3.1A. An 8-nt thymidine spacer equipped with a terminal biotin was attached to the 3'-G25 to reduce surfacial interactions during experimentation. The donor fluorophore, Cy3, was placed at 5'-C1, and the acceptor fluorophore, Cy5, was placed at A15**.** VEGF was expressed as inclusion bodies in *E. coli*. The inclusion bodies were purified using

cation exchange chromatography, then refolded, and the protein was further purified using heparin affinity chromatography [67].

3.2.2 *Data Acquisition and Analysis*

Single-molecule trajectories were obtained using a Verdi V-6 diode-pumped laser (Coherent) operating at 532 nm and ~1.1 W/cm^2 excitation irradiance. Emission from the donor/acceptor pair was detected on two channels using avalanche photodiode detectors (APDs, Perkin Elmer, Inc.), and a confocal microscope (Carl Zeiss MicroImaging, Inc.) equipped with a piezo stage (Physik Instrumente). Images were rendered by XPMPro (RHK Technology), and photons were counted by PMS (Becker & Hickl GmbH). The experimental apparatus is illustrated in Figure 3.1B. Raw trajectories were obtained at acquisition times of 100 µs or 1 ms by focusing the laser spot on a single molecule until both fluorophores photobleached. To ensure that each trajectory

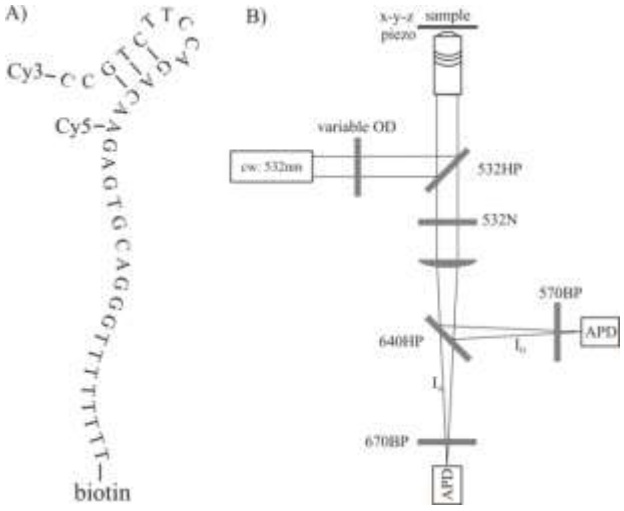


Figure 3.1. Modifications to the aV aptamer are shown in **A)**. Cy3 is attached at 5'-C1, and Cy5 at A15. An biotinylated 8-T spacer is attached at the 3'-G25. The scanning confocal microscope is shown in **B)**. Optical filters are high-pass (HP), notch (N), or bandpass (BP) filters.

represented a single FRET pair, those that showed multi-step photobleaches were excluded from analyses.

All analysis software was coded in-house using MATLAB (version 7.2.0.232, The Mathworks) with the exception of HMMs, which were compiled using the HaMMy GUI [20] (available at http://bio.physics.uiuc.edu/). A processing algorithm binned each trajectory to 10ms time steps, subtracted background and crosstalk, and identified photoblinks. Efficiency was calculated at each 10ms bin via $E_A = I_A/(I_A + I_D)$, where I_A and I_D represent photon counts detected on each respective APD. HaMMy searched for the presence of five states in each time trajectory. Global analyses filtered trajectories that have low total signal, then compiled efficiency histograms of observed and modeled values. Modeled efficiencies were used in lifetime analyses under the assumption of a three-state model. All displayed fits are nonlinear least squares regressions that were optimized using a simplex-based algorithm.

3.3 Results and Discussion

3.3.1 The aV Aptamer without VEGF

SMFRET analysis of the aV hairpin demonstrates that the aV aptamer favors the closed form as predicted by Mfold, but is also quite dynamic. Figures 3.2A and 3.2B contain global histograms of aV SMFRET efficiencies. Na^+ concentration remains constant at 145 mM in both histograms, but the concentration of Mg^{2+} differs by an order of magnitude. Figure 3.2A is compiled from 59 molecules at 2 mM Mg^{2+}, and Figure 3.2B from 46 molecules at 0.2 mM Mg^{2+}.

The higher concentration of Mg^{2+} serves to stabilize the folded conformation of aV relative to the lower concentration. This is shown in the histogram of modeled efficiencies Figure 3.2A (inset), where a striking 70% of the idealized efficiencies fall above the mean value. We therefore conclude that the prevailing conformation of aV under these conditions corresponds to either three or four base pairs (See Supplementary Material). However, it is important to note that the efficiency distribution broadened more than can be attributed to experimental shot noise. Thus, it may also be concluded that there are considerable conformational fluctuations away from the base-paired aV structure.

Significant broadening of the distribution shown in Figure 3.2B illustrates that these fluctuations are intensified at lower Mg^{2+} concentration. The mean efficiency has shifted from 0.92 to 0.86, and the standard deviation has increased from 0.11 to 0.15. The histogram of modeled efficiencies shown in the inset of Figure 3.2B mirrors the observed efficiency distribution, and further emphasizes that fluctuations are more prevalent at lowered Mg^{2+} concentration. The distribution stresses the importance of Mg^{2+} counterions to the stability of the folded state of aV, as the Na^{+} concentration was kept constant at 145mM. We can conclude that the free energy of the closed state is Mg^{2+}-dependent, and that the conformational equilibrium shifts away from the closed conformation at lower Mg^{2+} concentration.

To confirm that efficiency broadening in the aV system is due to structural fluctuations in its relatively weak 2° structure, aV FRET trajectories were compared to those of the HIV-1 transactivation response (TAR) DNA hairpin. This hairpin has been

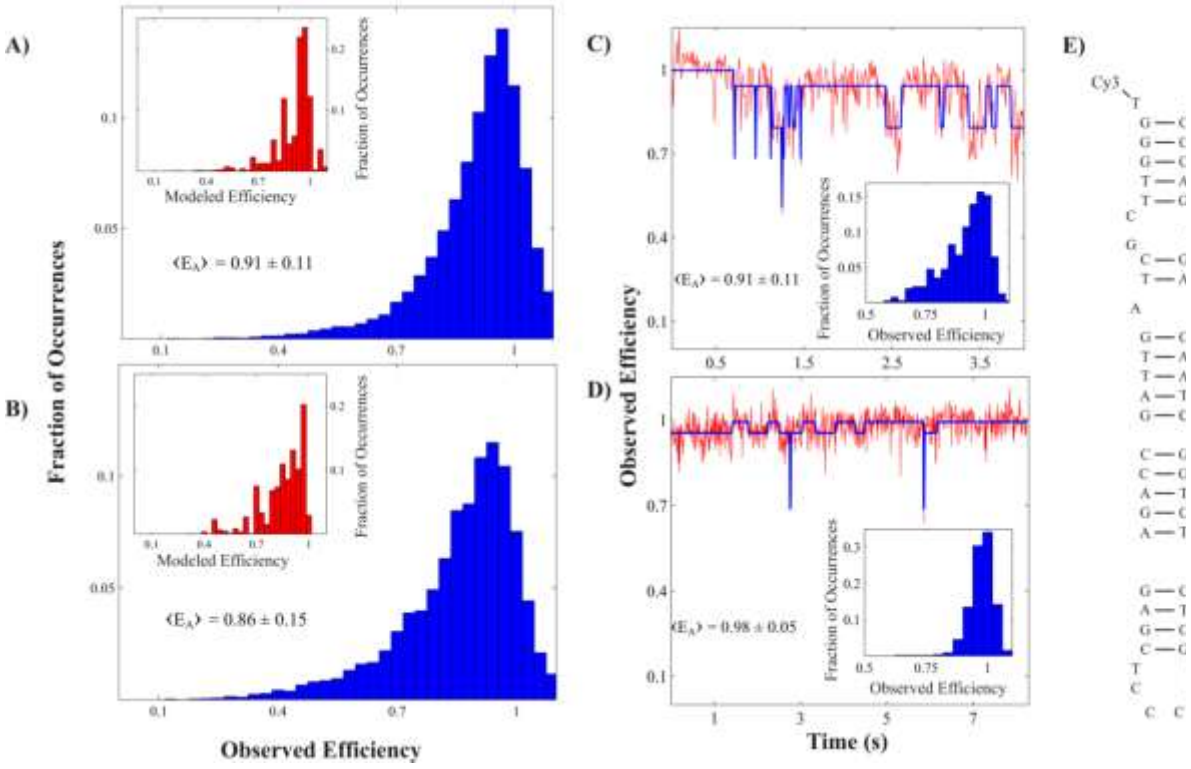


Figure 3.2. Global ensemble histograms and SM time trajectories. The histogram in **A)** was compiled 145 mM Na^{+} and 2 mM Mg^{2+}, while **B)** displays Mg^{2+} dependence as the histogram was compiled at 145 mM Na^{+} and 0.2 mM Mg^{2+}. Corresponding HMM values are displayed in each *Inset*. A representative aV time trajectory is shown in **C)**, and that of TAR is shown in **D)**. The trajectories illustrate the instability of aV 2° structure relative toTAR. A schematic of TAR is shown in **E)**.

extensively studied [62, 63] by SMFRET techniques, and is calculated [68] to have 2° structure that is more stable than that of aV by an order of magnitude. Figure 2C shows a representative trajectory of aV, and Figure 3.2D shows a trajectory of TAR. A schematic of TAR is shown in Figure 3.2E. Indeed, the aV trajectory shows discrete transitions between closed and open states, and the efficiency distribution is not Gaussian. In contrast, the TAR trajectory and its corresponding distribution, reflect that

its hairpin structure retains maximum base pairing without significant fluctuation. This comparison confirms the instability of the base-paired conformation of the aV aptamer in relative to that of a more stable hairpin.

3.3.2 *The aV Aptamer with VEGF*

The addition of VEGF results in the distribution presented in Figure 3.3A. The ionic conditions in these experiments were identical to those used in Figure 3.2A, and the histogram is compiled from 60 molecules. It is clear from the histogram that introduction of VEGF induces a shift in the efficiency distribution, indicating that aV interacts with the VEGF protein. The mean efficiency of this histogram is both lower and broader than its VEGF-free counterpart (0.85 ± 0.16), and a discrete structure appears at an efficiency of ~ 0.5. This feature may appear as a result of the aptamer breaking its base pairs to form a complex with its VEGF target. It is important to note that, although the global efficiency distribution indicates clearly that aV is interacting with VEGF, an irreversible product for the protein-aptamer complex cannot be distinguished. It can be concluded from this information that the aV-VEGF interaction is dynamic and requires further analysis.

Therefore, the trajectories that compose the distribution in Figure 3.3A were examined. It was found that trajectories generally fall into one of three subpopulations. The largest subpopulation is represented by the upper time trajectory in Figure 3.3B. These molecules show high and reasonably stable efficiencies, implying that they are mostly in a closed conformation. A second subpopulation of molecules, represented by the middle trajectory in Figure 3.3B, frequently fluctuates between low and

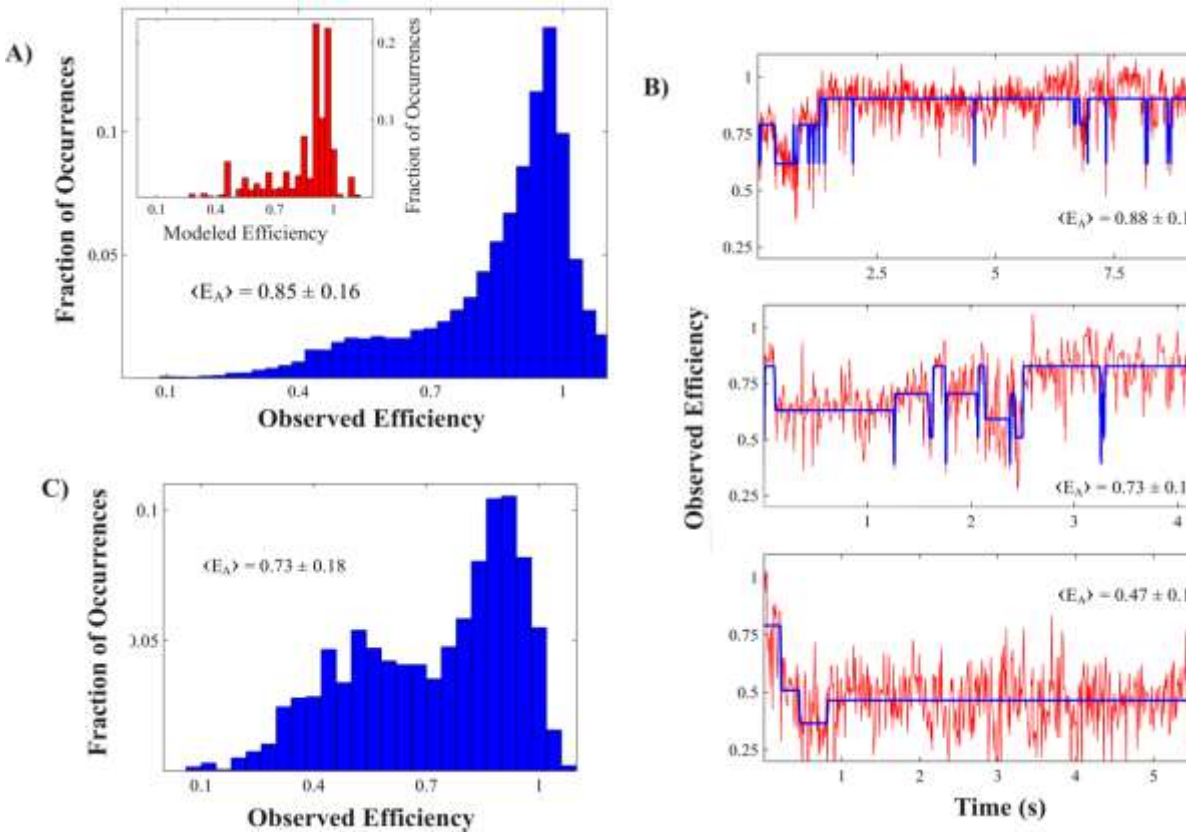


Figure 3.3. The aV aptamer in the presence of excess VEGF. **A)** A global ensemble of observed FRET efficiencies. This histogram was compiled at 145 mM Na^{+}, 2 mM Mg^{2+}, and 2.2 µM VEGF. HMM values are displayed in the *Inset*. Selected single molecule E_A trajectories (red) in the presence of excess VEGF along with their Markov-fitted trajectories (blue) are shown in **B)**. The histogram in **C)** is compiled from the trajectories in **B)**.

intermediate efficiencies. The third family is represented by the lower trajectory in Figure 3.3B. These molecules show low and reasonably constant efficiencies. The trajectories in Figure 3.3B were compiled with no modification to produce the histogram presented in Figure 3.3C. Visual inspection of this distribution shows qualitative agreement with the global distribution in Figure 3.3A, suggesting that the selected

subpopulations are accurate representations of aptamer dynamics that occur in the presence of VEGF.

3.3.3 Transition Lifetime Analyses without the Presence of VEGF

Kinetics of fluctuation in the aV aptamer were determined using lifetime analyses both in and out of the presence of VEGF. State-to-state transitions and reactant state lifetimes were identified from hidden-Markov modeled trajectories. A three-state model was assumed for these analyses. Details of this assumption are based on the three subpopulations of molecules discussed above and on further analysis described in supplementary material. Approximations of the efficiency ranges of each state were taken to be greater than 0.90 for a closed state (A), representing the base-paired forms of aV, 0.78 – 0.90 for a state (B) that represents intermediately base-paired forms of the aptamer, and less than 0.78 for a state (C) that has no base pairs in the structure.

Transition	*k without VEGF (s^{-1})*	*k with VEGF (s^{-1})*
A → B	3.2	4.5
A → C	1.7	2.2
B → A	9.3	13.1
B → C	8.0	5.7
C → A	26.9	23.7, 1.5
C → B	25.1	15.1

Table 3.1. Rate constants obtained by fitting each lifetime histogram to a single or double exponential decay with a bin time of 10ms. State **A** corresponds to the closed conformation of aV, state **B** to an intermediately base-paired conformation, and state **C** to a conformation having no base pairs.

The pseudo-first order rate constants obtained from these analyses are presented in Table 3.1. There are 404 total transitions in the hidden-Markov modeled trajectories at 2 mM Mg^{2+} without the presence of VEGF. The lifetime distributions and the acquired rate constants confirm that, without the presence of VEGF at 2 mM Mg^{2+}, the preferred conformation of the aV aptamer is a base-paired conformation, although there are frequent transitions out of this state. Looking more closely at transitions from state A reveals that transition to state B is favored over transition to state C. Transitions from state B reveal that transition to state A is slightly favored over transition to state C. Lifetime distributions of state C show that this state is quite unstable with much shorter lifetimes than the aforementioned states. These data signify that intermediately base-paired forms of the aptamer exist largely as intermediates in transitions between states A and C, but that transition to an intermediate conformation during a folding or unfolding process is not a requirement.

3.3.4 *Analysis in the Presence of VEGF*

The same analyses were performed on the aV trajectories obtained in the presence of VEGF, the rate constants are also reported in Table 3.1. These data show 439 total transitions in the modeled trajectories. It is suggested by an increase in the rates of transition out of state A that the aptamer's initial interaction with VEGF is from a base-paired conformation. Most strikingly, it is also found that the C → A transition lifetime is appreciably longer in the presence of VEGF. This suggests that interaction with VEGF results in a stabilization of the aV aptamer's fully open conformation.

Lifetime distributions and exponential fits for the C → A transition are shown in Figures 3.4A and 3.4B. It can be seen from the lifetime distribution in Figure 3.4A that state C is short-lived when VEGF is not present, and that the transition is well represented by an exponential decay. In contrast, the distribution in the presence of VEGF shows a much longer transition lifetime. IN addition, the distribution does not fit to a single exponential decay, but to a sum of two exponentials. This suggests that we observe two separate processes. Furthermore, by fitting to a sum of two exponential decays, we reveal a fast process with $k_1 = 23.7\ s^{-1}$, and a slower process with $k_2 = 1.5\ s^{-1}$. The former is comparable to $k = 26.9\ s^{-1}$ from the VEGF-free case, therefore it is reasonable to conclude that the latter process may represent the dissociation of the protein-aptamer complex and the aptamer's subsequent return to the closed conformation.

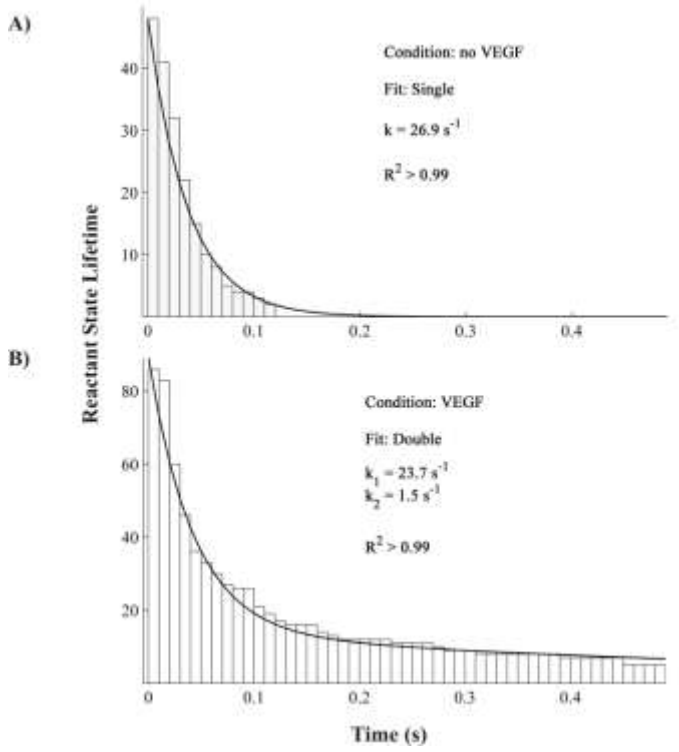


Figure 3.4. Lifetime analyses. Comparison of the C → A transition in the absence of VEGF is shown in **A)**. The same transition in the presence of VEGF in **B)**. Transitions shown in **A)** are fit to a single exponential decay. Transitions shown in **B)** are fit with a sum of two exponential decays.

3.4 Conclusions

We have used a series of SMFRET experiments to show that, without the presence of binding target VEGF, the aV aptamer favors a closed conformation. Despite this preference, conformational fluctuations occur quite often. Global analyses also showed an appreciable increase in the occurrence of less-favored conformations when Mg^{2+} concentration was lowered by an order of magnitude. The addition of VEGF causes distinctive changes to the aptamer's conformational equilibrium, and interaction of the aptamer with the VEGF protein is observed by the appearance of a stable, but reversible, conformation at low SMFRET efficiency. Furthermore, by examining the hidden-Markov modeled trajectories in detail, we determined that interaction of the aV aptamer with the VEGF protein results in an additional pathway for the return of the aptamer from a fully open state to its most stable, closed conformation. This is likely due to the release of VEGF by the complexed aptamer molecule.

3.5 Appendix

3.5.1 *Lowest Energy Conformations*

The lowest free energy structure at experimentally relevant Mg^{2+} conditions is predicted by the Mfold DNA folding server [68] to have three base pairs as in Figure 3.5A while the lowest energy structure at Mg^{2+} concentrations higher than experimental conditions is predicted to have four base pairs as shown in Figure 3.5B. The Mg^{2+}-dependence of the free energy both structures is shown in Figure 3.5C for experimentally relevant conditions.

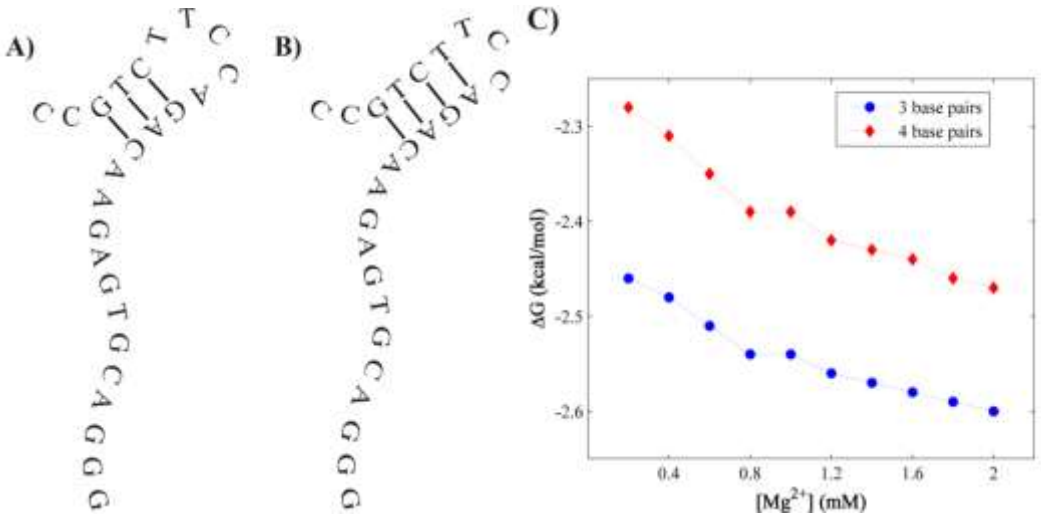


Figure 3.5. The free energy dependences of the two conformations shown in **A)** and **B)** are depicted in **C)**, as calculated by the Mfold DNA folding server.

3.5.2 *The Three-State Model*

The assumption of a state model is necessary for lifetime analyses, so the distribution of modeled efficiencies at 2 mM Mg^{2+} (displayed in the inset of Figure 3.6A) was fit to Gaussian distributions of various orders. The accuracy of each fit was considered along with the relative legitimacy the number of states in each Gaussian distribution, and it was determined that a three state model, with Gaussian efficiency means of 0.58, 0.83, and 0.96, describes the system with enough precision to carry out these analyses. The results of this nonlinear fit are displayed in figure S2A. The nonlinear fit of observed SMFRET efficiencies (shown in Figure S2B) was obtained by fixing the mean efficiency of each Gaussian distribution to its corresponding value in the modeled efficiency distribution. In order to determine if the assigned efficiency ranges were valid assumptions, the range of state B was systematically adjusted. It was found that variations of the lower efficiency threshold on the interval 0.72 – 0.78, and of the

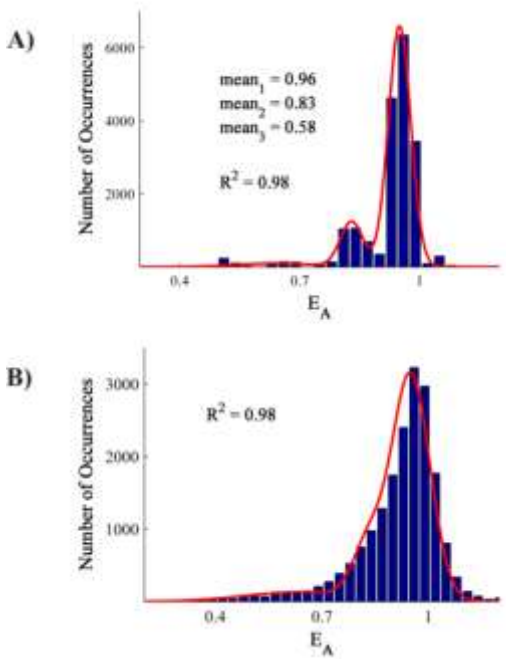


Figure 3.6. Nonlinear fits to triple Gaussian distributions. The HMM-generated histogram at 2mM Mg^{2+} is displayed in **A)**. The mean SMFRET efficiencies of each Gaussian function in this distribution (0.58, 0.83, and 0.96) were used to generate the triple Gaussian fit of the observed SMFRET efficiency histogram at 2mM Mg^{2+} in **B)**.

upper threshold on the interval 0.88 – 0.92, had no effects on the outcome of the analyses. This indicates that intermediately base-paired states of the aptamer have a discrete and well-defined SMFRET efficiency range, and that the assigned efficiency ranges are accurate descriptions of each state.

Chapter 4

Objective Identification of Photoblinks in smFRET Trajectories with Bayesian Inference

The contents of this chapter are adapted from an article originally published in *Biophysical Journal* on January 6, 2010.

Taylor, J. N., Makarov, D. E., Landes, C. F. "Denoising Single-Molecule FRET Trajectories with Wavelets and Bayesian Inference." *Biophys. J.* ***2010.*** 98(1):164-173.

ABSTRACT

Photoblinks are a considerable source of uncertainty in smFRET experiments. Here, we present a method to objectively identify photoblinks are identified in single-molecule fluorescence resonance energy (smFRET) trajectories using Bayesian inference. Bayesian methods are developed to identify fluorophore photoblinks in the time trajectories. Simulated data are used to quantify the improvement in static and dynamic data analysis. Application of the method to experimental smFRET data shows that it distinguishes photoblinks from large shifts in smFRET efficiency while maintaining the important advantage of an unbiased approach.

4.1 Introduction

Reversible photoblinks that result in the fluorophore's occupation of a non-absorbing and non-emitting, or dark, electronic state [69-71] are a problematic source of noise in single-molecule experiments. Photoblinking arises primarily due to the triplet excited state of the fluorophores that are used in single-molecule experiments [28]. It is characterized in the data by a drop in the emission intensity of the fluorophore to zero for some amount of time, and then the return of the emission intensity to normal values. Photoblinking is particularly troublesome to smFRET measurements, where the occurrence of a photoblink causes many undesired photophysical phenomena [28].

Many of the statistical analysis implementations require preprocessed data that is free of photoblinks, but their identification becomes an issue when considering that smFRET experiments are most often designed so that conformational shifts lead to changes in smFRET efficiency [57, 72, 73]. Furthermore, these events are most often removed manually, leading to bias in the smFRET time trajectories. Therefore, an unbiased method of photoblink identification that recognizes photoblinks on all time scales is desirable. Bayesian methods make use of observations in such a way as to provide insight into unknown events based on known properties of a system [74]. Such methods are often used in model [75] and hypothesis testing, and in the case of photoblinking events in smFRET time trajectories, we utilize the power of Bayesian inference to identify these events.

In this chapter, we first define and describe the characteristics of smFRET photon trajectories; then we describe the Bayesian method in terms of these characteristics.

Next, we apply the method to a single photon trajectory; then extend the analysis to a collection of photon trajectories. We then test the method against a simple and commonly used method, and find that the Bayesian method significantly outperforms the simple, subjective method. Lastly, we test the Bayesian method in relation to the equilibrium constant of a binary equilibrium, and find that the Bayesian method performs well throughout a large range of equilibrium constants.

4.2 Parameters in smFRET Trajectories

The acquisition of a two-channel smFRET time trajectory results in two data vectors that contain acceptor and donor photon counts in discrete time steps. We term these numbers of detected photons as N_A for acceptor and N_D for donor. The standard collection window contains both acceptor and donor fluorophore photobleaching events, and results in three distinct regions within each of these vectors: background, crosstalk, and FRET regions, as illustrated in Fig. 2.1. The utilization of photobleaching events on both fluorophores involved in these measurements allows for the collection of local background signals for each single molecule. Poissonian shot-noise in each of the background signals arises, producing a Poisson distribution around each channel's mean number of photons. We define the mean and variance of these distributions of on detection channel α as b_α (α = A, D).

Imperfect separation of donor and acceptor photons by a dichroic mirror leads to the observation of donor-emitted photons on the acceptor detector. This is referred to as crosstalk, and we produce the crosstalk region by observing a molecule after photobleaching of the acceptor but before that of the donor. Due to the absence of

acceptor fluorophore emission in this region, photons detected above background levels on the acceptor channel are assumed to be emitted by the donor fluorophore. The number of crosstalk photons is thus estimated as $p_x = N_A - b_A$. Because the crosstalk photons are emitted by the donor but counted as acceptor photons, the actual number p_D of donor-emitted photons in the crosstalk region is different from the number of observed photons N_D, and is estimated as

$$p_D \equiv N_D - b_D + p_x \qquad (4.1).$$

The ratio of the number of crosstalk photons p_x to the number of photons emitted by the donor fluorophore p_D is calculated at every time step in the region, and the mean of these values is taken to be the crosstalk parameter

$$x = \langle p_x / p_D \rangle \qquad (4.2).$$

To estimate the efficiency of energy transfer in the FRET region of an smFRET trajectory, again we must correct for the difference between the numbers N_D and N_A of detected photons, respectively, on the donor and the acceptor channels, and the numbers n_D and n_A of photons actually emitted by the respective fluorophores. The latter can be estimated as

$$n_D = N_D - b_D + n_x \qquad (4.3a)$$

$$n_A = N_A - b_A - n_x \qquad (4.3b)$$

Here N_α and b_α (α = A or D) are defined previously, and n_x is the number of crosstalk photons calculated at each time step using the relation

$$n_x = x n_D = \frac{x}{1-x}(N_D - b_D) \qquad (4.4).$$

The energy transfer efficiency, E, is calculated at each time step in the FRET region using the standard relationship

$$E = \frac{n_A}{n_A + n_D} \qquad (4.5).$$

In the context of the present discussion, we refer to the efficiency E as "observed efficiency". We determined the correction factor commonly known as γ [64, 76] to be within error of unity for our apparatus using the fluorophores Cy3 and Cy5.

4.3 Bayesian Inference to Detect Photoblinks

Photoblinks involving either fluorophore are characterized by observation of a sharp drop in the detected number of acceptor photons. In the instance of a donor photoblink, photon counts on both channels fall to background levels due to the donor's occupation of a dark electronic state, thereby rendering it unable to transfer energy to the acceptor fluorophore. Similarly, during an acceptor photoblink, donor emission is observed in the absence of energy transfer, and the numbers of detected acceptor photons during fall to levels similar to those observed in the crosstalk region.

As a means to detect photoblinks, Bayes' Law [74] is used to estimate the probability that the detected number of acceptor photons N_A arises due to a photoblink. In order to accomplish this, we need the conditional probability distributions of N_A given two alternatives, the "no blink" hypothesis (NB) and the "blink" hypothesis (B). After we obtain these distributions, we use Bayes' Law to reverse this logic and calculate the probabilities of each hypothesis given the observation of acceptor intensity N_A. This allows us to select those time steps that arise due to a photoblink, and remove them from the time trajectory in an unbiased manner.

Here we describe the Bayesian photoblink detection algorithm in detail. We need the conditional probability distributions of the acceptor photon count N_A given our two alternatives, the "blink" hypothesis, and the "no blink" hypothesis. A Poisson process describes photon emission in each case, but their properties differ. Specifically, in the absence of a blink the distribution of N_A can be approximated as

$$P(N_A|NB) = \frac{1}{(2\pi\mu_{NB})^{\frac{1}{2}}} \exp\left[-\frac{(N_A - \mu_{NB})^2}{2\mu_{NB}}\right] \qquad (4.6),$$

where $\mu_{NB} = \langle N_A \rangle$ is the calculated mean of detected acceptor photons for the trajectory under consideration. In writing Eq. 4.6, we have replaced the Poisson distribution by a Gaussian one whose mean and variance are both equal to μ_{NB}. This approximation is valid assuming $\mu_{NB} >> 1$, which is always the case for typical time steps used (~10 ms).

Similarly, the conditional probability distribution for N_A given the "blink" hypothesis can be written as a normal distribution. To calculate its width and its mean we note that the mean number of detected acceptor counts arising from such a situation will be the same that arises in the crosstalk region. We express this mean as

$$\mu_B = \langle b_A \rangle + x\langle n_D \rangle \qquad (4.7),$$

where $\langle b_A \rangle$ is the mean of the acceptor background signal, x is the crosstalk value as calculated from Eq. 4.7, and $\langle n_D \rangle$ is the mean of donor fluorophore-emitted photons over the FRET region as calculated from Eq. 4.3a. Again, we note that the width of this distribution is induced by shot-noise, and thus assume the distribution's variance to be

equal to its mean, μ_B. These assignments result in the following expression for the prior probability distribution of N_A given the "blink" hypothesis:

$$P(N_A|B) = \frac{1}{(2\pi\mu_B)^{\frac{1}{2}}} \exp\left[-\frac{(N_A - \mu_B)^2}{2\mu_B}\right] \qquad (4.8).$$

The posterior probability of the blink hypothesis B given the observation N_A is now given by Bayes' theorem [74]:

$$P(B|N_A) = \frac{P(N_A|B)P(B)}{P(N_A|B)P(B) + P(N_A|NB)P(NB)} \qquad (4.9),$$

where P(B) is the prior probability of hypothesis B and P(NB) is the prior probability of the no blink hypothesis NB.

To evaluate Eq. 4.9 we need the probabilities P(B) and P(NB). These probabilities represent the probability of the hypothesis H (= B or NB) being true *prior* to taking the observation N_A into consideration. We do not assume to know these values in advance, but instead obtain self-consistency through an iterative process. Initial guesses of P(B) = 0.001 and P(NB) = 0.999 are input, each posterior probability is calculated at each time step, and those having Bayes factors – i.e., the ratio of $P(B|N_A)$ to $P(NB|N_A)$ – greater than 2 are labeled as blinks. The fraction of time steps labeled as photoblinks is defined as the blink fraction, f_B. This value is compared with the value of P(B) at each iteration. Unless a deviation between these values of less than 5% is obtained, the substitutions $P(B) = f_B$, and $P(NB) = 1 - f_B$ are made, and the process repeats until this condition is met.

Fig. 4.1 illustrates the process of the identification and removal of photoblinks from an acceptor fluorophore-emitted time trajectory. Fig. 4.1a depicts a trajectory that

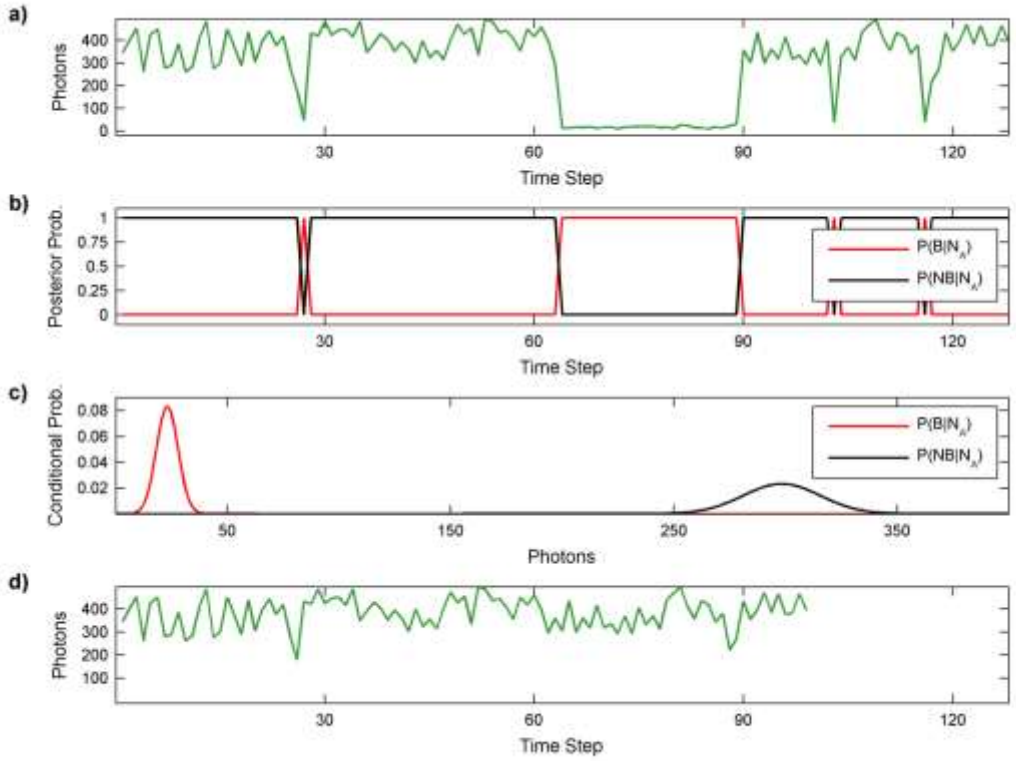


Figure 4.1. Photoblinks in an smFRET time trajectory. a) An acceptor photon count trajectory that contains photoblinks on long and short time-scales. b) The calculated posterior probability of the occurrence of a photoblink [77] and the absence of a photoblink (black) at each time step. c) The prior probability distributions of a photoblink (black, left) and the absence of a photoblink (red, right). d) The same acceptor photon trajectory with photoblinks removed.

contains both long and short time-scale photoblinks. Fig. 4.1b shows the posterior probabilities evaluated at each time step from the probability distribution of the blink hypothesis B. The conditional probability distributions $P(N_A|B)$ and $P(N_A|NB)$ are shown in Fig. 4.1c. Lastly, the acceptor fluorophore-emitted photon counts after removal of

time steps that contain photoblinks are shown in Fig. 4.1d. The time steps that are removed Bayes factors greater than 2, which is chosen because the number of data points removed becomes, to a large extent, constant for Bayes factors larger than 2. Typical Bayes factors obtained for time steps containing photoblinks are on the order of 10^{10}.

A caveat arises in the preceding logic in that, if one is searching for smFRET efficiencies approaching zero, then one cannot distinguish low efficiencies from acceptor photoblinks. However, if experiments are designed such that low efficiencies cannot be 'real' observations, as will be addressed in more detail below, or if photoblinks are typically on a much faster time scale than experimental observations, this caveat can be avoided entirely.

4.4 Photoblink Detection in Simulated Data

In order to assess the strengths and weaknesses of the photoblink detection method, we generate simulated smFRET trajectories using the kinetic Monte Carlo method [78-81], and apply the photoblink detection algorithm to the simulated data. We simulate a three-state system that represents the equilibrium of two efficiency states having central efficiencies of 0.8 and 0.2, respectively, as well as a photoblink state that represents both acceptor and donor photoblinks. An equilibrium constant, K_{eq}, of 0.4 is chosen for the $0.8 \leftrightarrow 0.2$ equilibrium. The average photoblink lifetime is described by exponential kinetics, and is chosen such that realistic photoblinking statistics are obtained [82]. The lifetimes of states 0.2 and 0.8 are also described by exponential kinetics, and are chosen to mimic realistic physical conditions [83]. After

the simulation generates the states that are present at each time step, shot-noise laden acceptor and donor photon trajectories are constructed from the simulated state trajectories. The photoblink detection algorithm is applied to the constructed photon trajectories, and time steps identified as photoblinks by the algorithm are removed. State lifetimes are extracted from the simulated data both before and after photoblink detection as a means to obtain the forward and backward rate constants for transition between the two real states. The equilibrium constant is estimated from the ratio of these rate constants as well as the ratio of the occurrences of each state in the efficiency distribution.

	Before Blink Filter	After Blink Filter
Total Data Points	100,078	74,594
No. of State 0.8	52,665	52,651
No. of State 0.2	23,218	21,882
No. of Blinks	24,195	52
% Blinks Removed	-	99.8 %
No. Identified as Blinks	-	25,484
% State 0.8 Removed	-	< 0.1 %
% State 0.2 Removed	-	5.8 %

Table 4.1. Statistics of simulated data before and after photoblink detection.

As shown in Table 4.1, the photoblink detection algorithm removes 99.8 % of the total number of generated photoblinks. Additionally, the algorithm's selectivity is demonstrated by the removal of only 1.8 % of the actual data points. Even in the presence of shot-noise, state 0.2 is only marginally affected by the removal of photoblinks, as a meager 5.8 % of the data points originally assigned to this state are removed during photoblink detection.

Fig. 4.2 illustrates the data simulation and the application of the photoblink detection algorithm in more detail. A sample acceptor and donor photon trajectory is shown in Fig. 4.2a, demonstrating the following chemical and photophysical transitions: transitions between the two designated FRET states, donor photoblinks, and acceptor photoblinks. Fig. 4.2b contains the efficiency distribution of the simulated data prior to photoblink detection, and Fig. 4.2c shows the efficiency distribution of the simulated data after photoblink detection. This comparison demonstrates that the denoising algorithm effectively removes photoblinks, resulting in an efficiency distribution that accurately reflects the two states of the system, even though the shot-noise broadened signal from state 0.2 overlaps with blink values.

Effective blink removal also improves dynamic analyses. Figs. 4.2d and 4.2e show the lifetime distributions prior to photoblink detection for state 0.8 and state 0.2, respectively, overlaid with their respective fits to single exponential decays. Figs. 4.2f and 4.2g show the same data, after photoblink detection and removal. The simulated data shown in Fig. 4.2 demonstrate that the removal of photoblinks from the simulated data results in more accuracy in the extracted kinetic rates. The ratio of forward to backward rate constants prior to photoblink detection and removal extracted from Fig. 4.2d and 4.2e, is 0.68, showing poor agreement with the equilibrium constant of 0.4. However, the corresponding ratio obtained after the removal of photoblinks extracted from Fig. 4.2f and 4.2e, is 0.41, thus showing excellent agreement with the equilibrium constant of 0.4. It is therefore shown that performing Bayesian photoblink removal on the simulated data results in a fitted equilibrium constant that differs by only 2.5% from

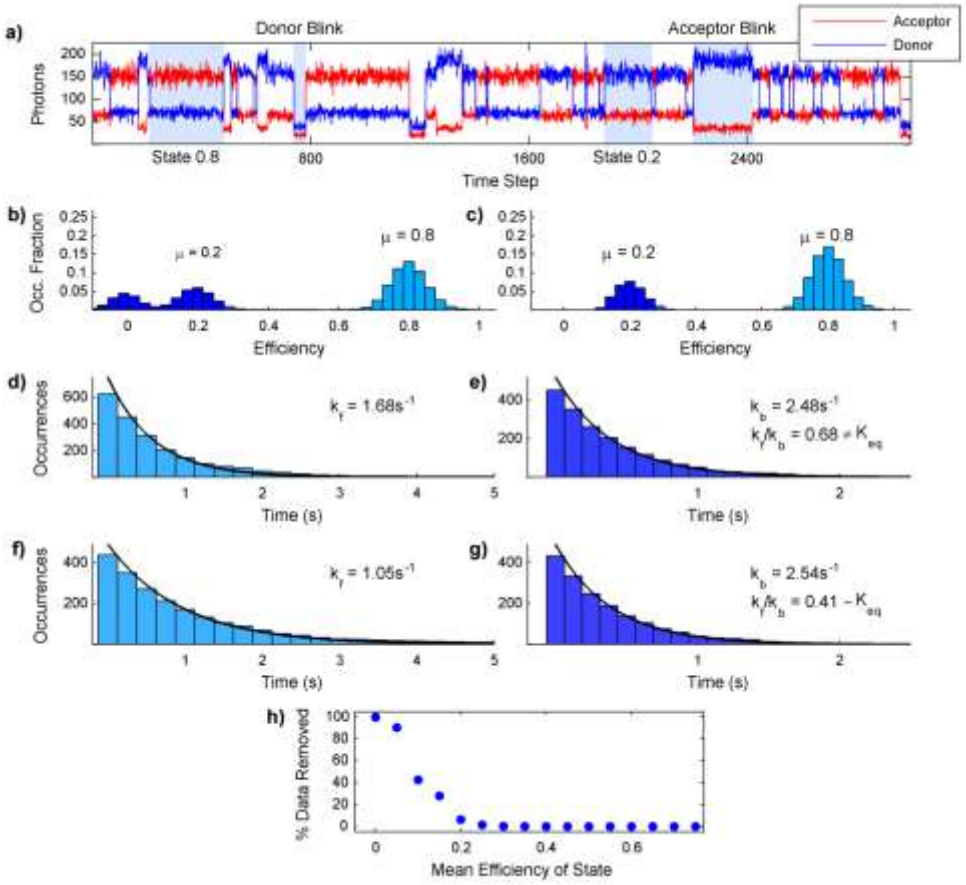


Figure 4.2. Applying photoblink detection to simulated smFRET trajectories. a) Sample acceptor (*red*) and donor (*blue*) photon trajectories. The mean of the sum of acceptor and donor photon counts at each time step was held constant at 220. b) Efficiency distribution of the model system prior to photoblink removal. c) Efficiency distribution of the model system after photoblink removal showing K_{eq} to be 0.4. d) The lifetime distribution of State 0.8 prior to photoblink detection overlaid by a fit to a single exponential decay. e) The lifetime distribution of State 0.2 prior to photoblink detection overlaid with its fit to an exponential decay. f) The lifetime distribution of state 0.8 after photoblink detection overlaid with a fit to an

exponential decay. g) The lifetime distribution of state 0.2 after photoblink detection overlaid with its fit to an exponential decay. h) The fraction of total data points removed from a state's efficiency distribution versus the mean efficiency of the state.

the actual value. In comparison, the error prior to photoblink removal is 70%. These results confirm that more accurate dynamic information can be extracted from smFRET trajectories after the removal of photoblinks.

In order to address the resolution of photoblinks from states having low central efficiencies, we perform a series of simulations as a function of mean state efficiency as a means to determine a lower limit for this method. The results of these simulations are shown in Fig. 4.2h. We find that, in order to distinguish photoblinks from actual data, a state's mean efficiency needs to be higher than a lower bound of approximately 0.2. It is of note that this lower bound is a function of total acquired photons per time step, and will move toward zero as the total number of acquired photons increases. In the context of the current discussion, the mean number of total photons per time step is 220, and at this value the simulations confirm that the algorithm, while removing all but a negligible amount of photoblinks, leaves occurrences of states with mean efficiencies higher than approximately 0.2 essentially unaltered.

It is also important to note that the effects of intermediate time scale photoblinks that are less than one time step in duration. These photoblinks limit the Bayesian method presented here in that, relative to the length of the event, intensity

falls, but does not fall low enough to be designated a photoblink. As such, the time step remains. Such events fall into the unquantifiable noise contribution discussed above.

4.5 Comparison to Simple Thresholding

To compare the effects of the Bayesian photoblink detection algorithm to a more traditional method of photoblink removal, simulated trajectories were constructed using kinetic Monte Carlo simulation. Three states were included in this simulation, the observable state with central efficiency of 0.15 and average lifetime of ≈1 s, an acceptor photoblinking state, and a donor photoblinking state. The lifetimes used for the photoblinking states were allowed to vary, producing photoblinks with lifetimes ranging from 10 ms (1 time step) to 5 seconds. By inspection of the sample trajectory shown in Fig. 4.3a, one can clearly see that, in the case of these trajectories, manual photoblink removal would most likely have a negative effect on the data. Human bias would arise due to the slight difference between what is actual data and what is a photoblink.

Photoblinks were filtered from the simulated trajectories by 2 different methods: (1) the Bayesian method presented here, and (2) a simpler method that removed time steps with acceptor intensities below a predefined threshold τ_B, given by

$$\tau_B = \mu_B + 2(\mu_B)^{\frac{1}{2}} \qquad (4.10)$$

where, μ_B is as defined in Eq. 4.7.

Fig. 4.3 shows the results of this comparison. A sample trajectory containing unfiltered acceptor [77] and donor (blue) counts is shown (Fig. 4.3a) to have photoblinks of both the acceptor and donor varieties on both long and short time scales, as well as the presence of the observable state with a central efficiency of 0.15. Figs. 4.3b and

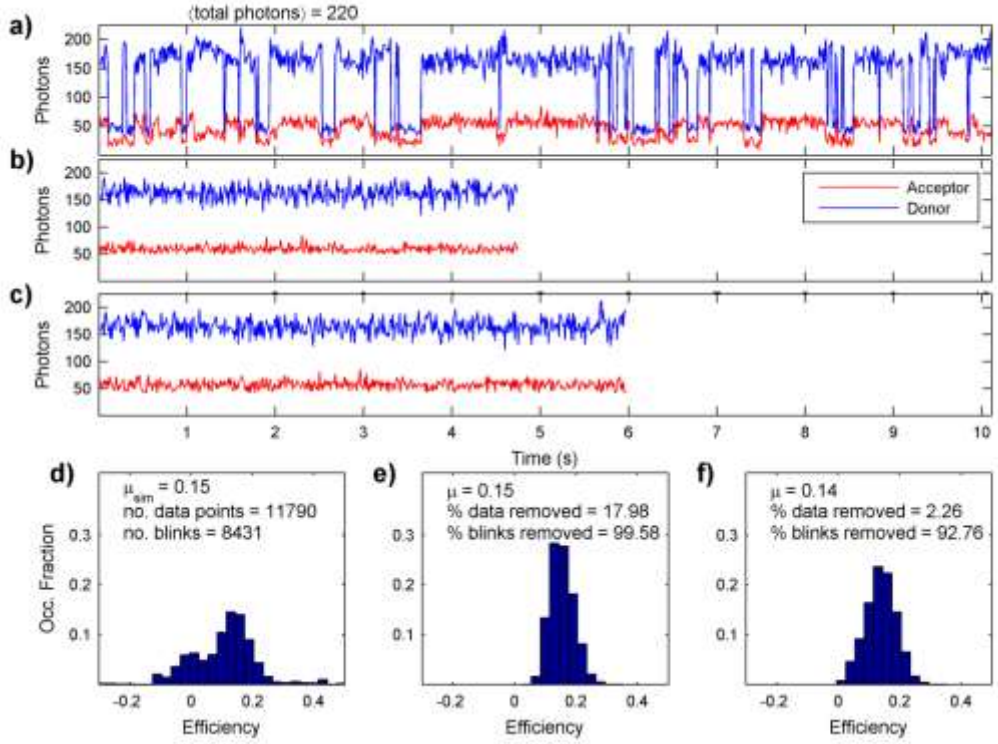


Figure 4.3. Comparison of photoblink removal methods. a) A sample trajectory from the simulated data. b) The sample trajectory after removal of photoblinks by the thresholding method. c) The sample trajectory after removal of photoblinks by the Bayesian method. d) The distribution of efficiencies produced by the unfiltered, simulated data. e) The efficiency distribution after photoblink removal using the thresholding method. f) The efficiency distribution after photoblink removal using the Bayesian method.

4.3c show the same trajectory after photoblink filtering by the thresholding method and by the Bayesian method, respectively. It is clear in these trajectories that a larger fraction (~ 15% larger) of the original data is retained by the Bayesian operation, but it

remains unclear if the remaining data is of the observable variety or of the photoblink variety.

Fig. 4.3d shows a distribution of the unfiltered data's calculated efficiencies. While we can see the presence of the observable state at a its central efficiency of 0.15, we also see that the distribution is marred by the noise arising from the presence of photoblinks. Fig. 4.3e shows the corresponding distribution after photoblink removal using the thresholding method, and Fig. 4.3f shows the distribution after photoblink removal using the Bayesian method.

It is clear from the statistics shown in Fig. 4.3e that the thresholding method successfully removes photoblinks from the trajectories (with > 99% success), but it is also clear that it removes a large portion of the actual data as well (~ 18%). This is a less than desirable quality, in that, if nearly 20% of the observable data is removed, then nearly 20% of the information acquired about this system is lost. In comparison, the Bayesian method also successfully removes photoblinks (with ~ 93% success), and removes a much smaller fraction of the observable data (~ 2.3%). While a portion of the original photoblinks remain, their contribution to the overall outcome is small, as is shown by the distribution in Fig. 4.3f. Perhaps more importantly, the human element of photoblink detection, whether it is in the manual selection of photoblink regions, or in the selection of an intensity threshold, has been completely eliminated.

4.6 The Performance of Bayesian Photoblink Detection in Relation to K_{eq}

Given the substantial variation of equilibrium constants in various biological interactions, the performance of the Bayesian photoblink filter was tested on simulated,

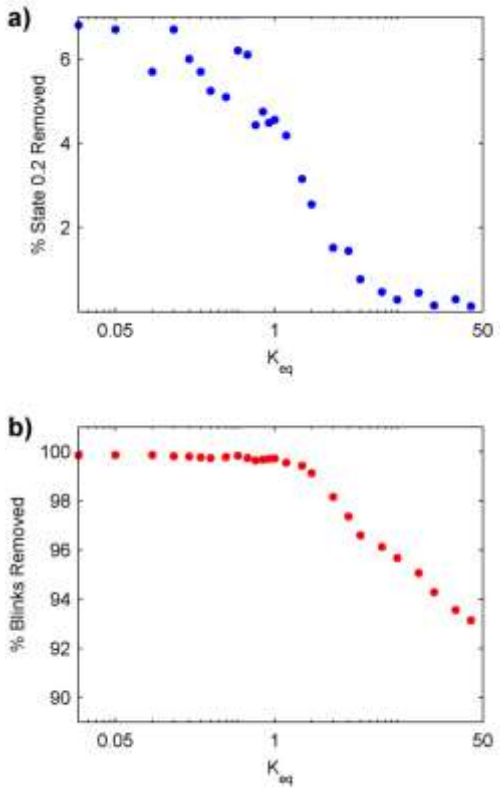


Figure 4.4. The performance of the Bayesian photoblink filter vs. K_{eq}. a) The percentage of state 0.2 data points removed vs. K_{eq}. b) The percentage of photoblinks removed vs. K_{eq}.

two-state equilibria with equilibrium constants ranging over 4 orders of magnitude. The smFRET trajectories for each equilibrium were simulated using kinetic Monte Carlo simulations with the mean number of total photons at each time step being 220. Each simulated equilibrium contained over 100,000 data points prior to photoblink removal, and consisted of two efficiency states, one with a central efficiency of 0.2 (state 0.2), and the other with central efficiency of 0.8 (state 0.8). The forward transition is taken to be the transition from state 0.8 to state 0.2, such that a $K_{eq} < 1$ corresponds to state 0.8 being favored in the equilibrium. The average photoblinking lifetime was ~ 0.4 s in all cases, and photoblinks were removed as described above.

Fig. 4.4 shows the results of these simulations. In Fig. 4.4a, the percentage of state 0.2 data points that were removed in each simulation are plotted logarithmically versus K_{eq}. It is shown by Fig. 4.4a that the percentage of state 0.2 removed never

exceeds 6.8%, and that, as the equilibrium shifts toward this state, the percentage of this state's data points that are removed trends toward zero. Fig. 4.4b shows the percentage of photoblinks removed versus K_{eq}, again plotted logarithmically. As shown in Fig. 4.4b, the percentage of photoblinks removed by the Bayesian photoblink filter only begins to fall after state 0.2 becomes favored in the equilibrium. This is simply a result of the mean acceptor intensity falling as the mean efficiency of the trajectory falls, and of the prior probability of a photoblink being less than that of the absence of a photoblink. We therefore conclude that, regardless of which state is favored in the equilibrium, as well as the extent to which a that state is favored, the Bayesian photoblink filter performs capably under all tested circumstances.

4.7 Conclusions

In conclusion, we have developed methods to identify, quantify, and remove a considerable source of uncertainty in smFRET time trajectories – photoblinks. The development of an unbiased method of photoblink detection eliminates the need to manually preprocess the trajectories, and perhaps more importantly, removes bias introduced into the measurement by manual selection of photoblink regions.

The algorithm's efficacy has been tested using simulated data. Acceptor and donor photon trajectories containing photoblinks were generated, and photoblink detection in these trajectories resulted in nearly complete elimination of photoblinks with little effect on the actual data. We have also shown that application of the Bayesian photoblink detection method in combination with the application of the wavelet denoising algorithm significantly improves the quality of experimental smFRET

data. This improvement is observed both in the ensemble analysis of structural distributions, and in kinetic analysis of dwell times. Although there are caveats involved with the method of photoblink detection, we have also shown that the caveats can be avoided through establishment of a lower efficiency bound.

Chapter 5

Wavelets and Shot Noise in smFRET Trajectories

The contents of this chapter are adapted from an article originally published in *Biophysical Journal* on January 6, 2010.

Taylor, J. N., Makarov, D. E., Landes, C. F. "Denoising Single-Molecule FRET Trajectories with Wavelets and Bayesian Inference." *Biophys. J.* ***2010***. 98(1):164-173.

ABSTRACT

A method to denoise single-molecule fluorescence resonance energy (smFRET) trajectories using wavelet detail thresholding is presented. Known sources of experimental noise are examined and quantified as a means to remove their contributions via soft thresholding of wavelet coefficients. A wavelet decomposition algorithm is described, and thresholds are produced through the knowledge of noise parameters in the discrete-time photon signals. Reconstruction of the signals from thresholded coefficients produces signals that contain noise arising only from unquantifiable parameters. The method is applied to simulated and observed smFRET

data, and it is found that the denoised data retain their underlying dynamical properties, but with increased resolution.

5.1 Introduction

The effects of experimental noise in single-molecule studies often limit their scope. Low signal-to-noise ratios are inherent to these experiments [84], and various statistical implementations have been applied in attempt to reduce the effects of experimental noise [64, 76, 85-93]. These implementations include the information theoretical methods [17] [18], statistical correlation functions to reveal single-molecule kinetic heterogeneities [19], and hidden-Markov models to extract the most likely sequence of events from smFRET time trajectories [20]. Most recently, statistical correlation is combined with wavelet decomposition in attempt to describe kinetic heterogeneities in single-molecule systems [22]. Despite the relative success of these implementations, much is still left to be desired from the resolution of single-molecule experiments.

Many analyses also rely on the assumption that the system's states are well-defined, and that transitions among these states are purely Markovian in nature [20, 21, 62]. However, the observation of memory effects in single-molecule enzymatic turnover [94], large variations in the folding kinetics of a ribozyme [95], and the occurrence of overlapping efficiency states in single DNA aptamer molecules [14] all offer recent experimental results that violate these assumptions. As such, a means of processing single-molecule data that provides a more accurate representation of a physical setting is still a pressing need.

A dual-component interpretation of noise in smFRET photon distributions results in a quantifiable component arising from known sources such as shot-noise and photoblinks, and an unquantifiable component arising from molecular phenomena like conformational fluctuations. Methods that discriminate the former component from the latter are known in signal processing, and wavelet-based approaches are directly applicable to time-series data [96-99]. Similar to Fourier transforms, wavelet transforms are mathematical constructs that convert a time-series signal into a representation in another domain. Wavelet transforms, however, offer the advantage of localization in both time and frequency [100].

The first and simplest of all wavelets was presented by Alfred Haar [77]. Since its invention a century ago, this wavelet and more sophisticated varieties have evolved into important tools in the fields of data compression and signal processing. Contributions by Mallat [101], Daubechies [102], and others [103-105] have extended the impact of such analyses to nearly all subdivisions of these fields. Wavelet-based analyses now enjoy a broad range of applicability, and have supplanted the use of the traditional Fourier transform in many areas [100, 106].

A framework that paved the way for the use of wavelets in signal processing was introduced as multiresolution analysis in the late 1980s [101, 102]. The basic scheme decomposes the signal into two components: an approximation component containing coefficients that multiply a scaling function, and a detail component containing coefficients that multiply the wavelet function. Thresholding the detail components of a signal's wavelet decomposition that are smaller than a certain value, a threshold,

effectively removes noise components from a noisy signal [26]. There are many elegant thresholding methods [100, 103, 104, 106, 107], but our aim in thresholding smFRET detail components is simple: we wish to discard the noise components we can quantify while keeping those we cannot.

Herein we present an algorithm for quantifiable noise suppression in smFRET time trajectories. We then use the Haar wavelet to decompose each of the two photon signals acquired in dual-channel smFRET experiments. Noise parameters in each signal are quantified as a means of generating a universal threshold, and quantifiable noise is removed from each photon signal via soft thresholding of the detail components. Signals are then reconstructed from the highest level approximations and thresholded details, producing denoised signals that contain noise artifacts arising only from unquantifiable sources.

Section 5.2: Denoising with Haar Wavelet

Denoising methods are generally designed to separate the essential component of the signal from the random noise generated by experimental error. The simplest example of denoising is the removal of high frequency noise via the application of a low pass filter to the original signal. Mathematically, this is accomplished by suppressing the high frequency Fourier components of the signal, which is comprised of (1) applying the Fourier transform to the signal, (2) modifying the high frequency components according to a certain rule, and (3) applying the inverse Fourier transform to obtain the denoised signal.

From this example it is clear that there are two ingredients in a denoising method: (i) the choice of the basis set used to represent the signal (e.g., the sine and cosine functions are chosen as the basis for the Fourier transform), and (ii) the rule according to which certain components of the signal – which are presumably associated with noise – are suppressed. For example, the above smoothing method implicitly assumes that the signal is contained in the low frequency part of the Fourier spectrum while the noise is associated with high frequency components of the signal. A successful method takes advantage of existing knowledge about the noise. Furthermore, if the basis functions that are chosen to represent the signal do so poorly (under the inevitable constraint of using a finite basis set), the method will not be successful. The keys then, to a successful denoising method, lie in making the proper choices regarding basis sets and noise suppression rules.

The orthonormal basis set used in the denoising method presented here is comprised of the Haar [77] wavelet and scaling functions. In general, wavelets offer the advantage over more conventional basis sets that they are localized in both the frequency and time domains. In contrast, the sines and cosines of Fourier transforms are localized only in the frequency domain. This time locality is particularly suitable for non-stationary time series, as in the famous example of the wavelet-denoised recording of Brahms at the piano [108].

Returning to the context of smFRET time trajectories, we recall that the trajectory consists of two data vectors containing detected numbers of photons in

discrete time steps. In the present discussion we consider only the acceptor photon trajectory $\mathbf{N}_A$ (= $N_A(0)$, $N_A(\Delta t)$,…), in discrete time steps Δt, which is written in the form

$$\mathbf{N}_A = \mathbf{S}_A + \sigma\mathbf{Z} \qquad (5.1).$$

Here, at each time step Δt, **Z** is a Gaussian white noise component, and each element of **Z** is independently and identically distributed on a normal distribution with mean 0 and variance 1, σ is a known noise level, and $\mathbf{S}_A$ is the "true" signal that we wish to recover. Similarly to the smoothing method described above, we accomplish the recovery of the true signal $\mathbf{S}_A$ in three steps: (1) transform the observed data $\mathbf{N}_A$ into the wavelet domain, (2) suppress the presumed noise component of the signal, and (3) invert the wavelet transform to obtain the denoised signal. A key assumption of most denoising schemes is that the noise is additive – that the strength of the noise is independent of the signal. In the case of smFRET experiments, the shot-noise strength is dependent on the brightness of the fluorophores. Here however, we approximate the noise strength with its average value.

The first step in this procedure is accomplished by the multiresolution approximation of Mallat [101]. It is a pyramidal algorithm that consists of multiple decomposition levels, each of which reduces the resolution of the signal by a factor of 2. Transformation of the signal to the first decomposition level produces two components, one containing information about the low frequency part of the Fourier spectrum, and the other containing information about the high frequency part of the Fourier spectrum.

The bases of each orthogonal complement are built by dilating and translating a unique function. In the case of the low frequency complement, this function is known

as the scaling function, and in the case of the high frequency complement, it is known as the wavelet function. The digital filter transforming the signal to the low frequency basis is a low pass filter, $\mathbf{W}_{\mathrm{lo}}$, and the counterpart filter transforming the signal to the high frequency complement is a high pass filter, $\mathbf{W}_{\mathrm{hi}}$. The filters used in the Haar wavelet transform [109] are described by the vectors

$$\mathbf{W}_{\mathrm{lo}} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \dots, 0\right) \tag{5.2a}$$

$$\mathbf{W}_{\mathrm{hi}} = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \dots, 0\right) \tag{5.2b}.$$

By applying the low pass filter, $\mathbf{W}_{\mathrm{lo}}$ to the time series $\mathbf{N}_{\mathrm{A}}$, one obtains the approximation $\mathbf{A}_1$ containing the signal at a reduced resolution

$$\mathbf{A}_1 = (\mathbf{W}_{\mathrm{lo}} * \mathbf{N}_{\mathrm{A}})_{\downarrow 2} \tag{5.3a}.$$

Here, the symbol "$*$" denotes a convolution, and $\downarrow 2$ represents the downsampling operation, where alternating elements of the convolution's output are removed. The high frequency, or detail component $\mathbf{D}_1$ is similarly produced by the convolution of the signal with the high pass filter

$$\mathbf{D}_1 = (\mathbf{W}_{\mathrm{hi}} * \mathbf{N}_{\mathrm{A}})_{\downarrow 2} \tag{5.3b}.$$

The approximation and detail signals at subsequent decomposition levels are obtained by substituting the approximation from the previous level for the signal $\mathbf{N}_{\mathrm{A}}$ in Eqs. 5.3.

The second step in the denoising procedure is the suppression of the noise component. As in the example above, the noise in a discrete-time photon signal is in the high frequency part of the Fourier spectrum, and as such, is contained within the high frequency component of the wavelet decomposition, the detail signal. The

simplest way to reduce the noise is to set any detail component that exceeds a certain tolerance value to zero. An improved procedure utilizes "soft thresholding" [110], where noise suppression is achieved by shrinking any detail component Y of a decomposition according to:

$$Y \rightarrow \mathrm{sgn}(Y)\max(0, |Y| - \tau) \equiv T(Y, \tau) \qquad (5.4).$$

Here, τ is the universal threshold value given by Donoho [26]

$$\tau = \sigma[2 \log n]^{\frac{1}{2}} \qquad (5.5).$$

where, n is the number of data points in the original time series (i.e., the dimensionality of $\mathbf{N}_A$). Given that this method is designed to remove shot-noise, we approximate the noise parameter σ as the fluctuation about the mean intensity induced by shot-noise. For the acceptor photon trajectory $\mathbf{N}_A$ with mean intensity μ_A, we have $\sigma = [\mu_A]^{1/2}$. Likewise, for a donor photon trajectory with mean intensity μ_D, we have $\sigma = [\mu_D]^{1/2}$. Inserting these values in Eq. 5.5, we generate thresholds τ_A and τ_D, respectively, for each of the acceptor and donor photon trajectories

$$\tau_A = [2\mu_A \log n]^{\frac{1}{2}} \qquad (5.6a)$$

$$\tau_D = [2\mu_D \log n]^{\frac{1}{2}} \qquad (5.6b).$$

Returning to the detail signal obtained in Eq. 5.3b, we now apply, element-wise with threshold calculated from Eq. 16a, the thresholding operation of Eq. 5.4, and obtain thresholded detail signal $\mathbf{D}_1^T$

$$\mathbf{D}_1^T = T(\mathbf{D}_1, \tau_A) \qquad (5.7).$$

Thresholded details at subsequent decomposition levels are obtained by applying the thresholding operator with the same threshold as in Eq. 5.6a.

The denoised signal $\mathbf{S}_A$ is obtained by inverting the decomposition procedure described above. Firstly, we define the low pass and high pass reconstruction filters, $\mathbf{W}_{lo}^{-1}$ and $\mathbf{W}_{hi}^{-1}$ respectively, as the reverse of their decomposition counterparts [109]

$$\mathbf{W}_{\mathrm{lo}}{}^{-1} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \ldots, 0\right) \tag{5.8a}$$

$$\mathbf{W}_{\mathrm{hi}}{}^{-1} = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, \ldots, 0\right) \tag{5.8b}.$$

Next, we dyadically upsample the highest level approximation and thresholded detail components by inserting zeros between each element of each vector. After upsampling, the approximation vector is convolved with $\mathbf{W}_{lo}^{-1}$, and the detail vector with $\mathbf{W}_{hi}^{-1}$. The sum of the output of each convolution is then the reconstructed approximation of the next resolution level. This reconstructed approximation is then upsampled, convolved, and combined with the next level's thresholded details. This procedure is summarized by Eq. 5.9 for the reconstruction of the denoised signal $\mathbf{S}_A$ from a level 1 approximation and thresholded details, where the superscript $\uparrow 2$ represents the upsampling operation

$$\mathbf{S}_{\mathrm{A}} = \left(\mathbf{W}_{\mathrm{lo}}{}^{-1} * (\mathbf{A}_1)^{\uparrow 2}\right) + \left(\mathbf{W}_{\mathrm{hi}}{}^{-1} * (\mathbf{D}_1^{\mathrm{T}})^{\uparrow 2}\right) \tag{5.9}.$$

Evaluation of the mean-squared error of simulated signals decomposed to various resolution levels revealed that this quantity is largely minimized for signals reconstructed from the third resolution level. As such, all decompositions were processed to the third level. Due to the nature of the decomposition, the number of elements in the time series must be a power of 2. Each signal was extended with a vector containing m elements, where the value of each element is the mean intensity of the signal, and where m is chosen such that $\mathrm{n} + \mathrm{m} = 2^{\mathrm{j}}$, with j being an integer. The dimensionality of the signal, n, remained unchanged in Eqs. 5.6. This extension was

discarded after reconstruction of the denoised signal. Each of the acceptor and donor photon signals in a particular smFRET trajectory were decomposed separately using the digital filters in Eqs. 5.2. Thresholds for each signal were generated using Eqs. 5.6, and the detail signals of each decomposition were thresholded via the soft thresholding operator in Eq. 5.4. Each of the acceptor and donor signals were then reconstructed as in Eq. 5.9.

Fig. 5.1 illustrates the denoising of an acceptor photon trajectory. The fluorophore-emitted acceptor photon signal is shown in Fig. 5.1a. Convolution of the signal vector with the low pass filter results in the first level approximation shown in Fig. 5.1b after downsampling. The second and third level approximations are generated by convolution of the previous level's approximation with the low pass filter. The detail coefficient vectors shown in Fig. 5.1c are similarly produced by convolution of the signal and approximation vectors with the high pass filter. The universal threshold as calculated from Eq. 5.6a is shown as dashed lines in Fig. 5.1c. Detail coefficients whose magnitudes are smaller than the threshold are set to zero, and the threshold is subtracted from detail coefficients having magnitudes larger than the threshold with retention of these coefficients' signs.

Fig. 5.1d shows the original acceptor signal in red overlaid with its denoised representation in black. The denoised signal was reconstructed by upsampling the third level approximation and thresholded details, convolution of the upsampled coefficients with their respective low or high pass reconstruction filters, and addition of the two upsampled vectors, producing the second level reconstruction. This reconstruction

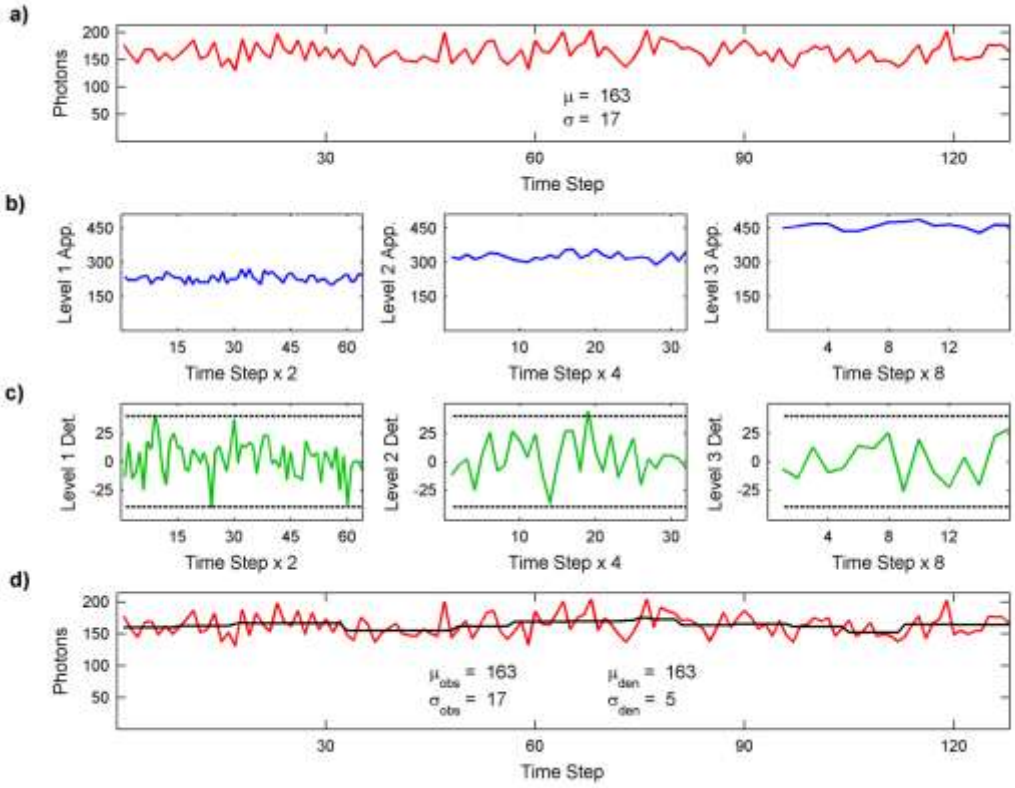


Figure 5.1. Denoising an acceptor photon trajectory. a) The original trajectory. b) The first, second, and third level approximation coefficients shown left to right. c) The first, second, and third level detail coefficients shown left to right, along with the detail threshold is shown as dotted lines. d) The original signal [77] overlaid with the denoised signal (black). The standard deviation of the original signal is reduced by a factor of 3, illustrating that the denoising process has been a success.

vector was then combined with thresholded second level details in the same manner, and the resulting vector combined with the thresholded first level details to produce the reconstructed signal shown in black in Fig. 5.1d. It is seen that, while the mean of the denoised signal remains the same as its noisy counterpart, comparison of the standard

deviations of each trajectory shows the noise level to be reduced by approximately 300% for this particular photon trajectory.

Section 5.3: Denoising an Oscillatory System

As a means to quantify the effects of the wavelet denoising algorithm to smFRET trajectories, simulated trajectories were generated for two types of systems. The first, a two state equilibrium, was simulated using kinetic Monte Carlo methods. Each of the simulated trajectories was denoised by the wavelet denoising algorithm as well as the hidden-Markov model (HaMMy) described by McKinney, *et al* [20]. Both methods are effective at denoising trajectories comprised of well-defined FRET states. A figure showing this comparison of the two methods as discussed below.

The next simulation was performed for a system without defined FRET states. The wavelet denoising algorithm does not make use of Markovian and/or distinct-state assumptions. Given that these assumptions are not valid in all cases, wavelet denoising offers a significant advantage. Examples include the wormlike multi-dT chains discussed by Murphy, *et al* [111], the aV aptamer [14], or any system that undergoes breathing dynamics with a continuously changing conformation. An example of such behavior is shown by green fluorescent protein (GFP), which recently been observed to exhibit periodic oscillation between two conformational extremes during the unfolding process [112].

As an extreme example of a system without well-defined states, we simulate a system showing conformational oscillation. Assuming the efficiency E of such a system

oscillates around a central value E_c with amplitude E_0 according to the equation $E(t) = E_0\cos(\omega t) + E_c$, the probability distribution p(E) of the efficiency is given by:

$$p(E) = \frac{1}{\pi[E_0^2 - (E - E_c)^2]^{1/2}} \qquad (5.10).$$

Although p(E) is weakly singular at $E = E_c \pm E_0$, the singularities are readily removed when a discrete probability distribution is used. The constructed trajectories are analyzed by the wavelet-denoising algorithm, and again compared with analyses produced by HaMMy.

Fig. 5.2a shows a typical trajectory generated for these analyses. The original, noisy efficiency trajectory (cyan) is overlaid with the results produced by the wavelet denoising algorithm (red) and HaMMy (black). Fig. 5.2b depicts the efficiency distribution of the noisy trajectory, and Fig. 5.2c shows that of the wavelet-denoised trajectory. Lastly, Fig. 5.2d shows the efficiency distribution as predicted by HaMMy. Each of the efficiency distributions in Figs. 5.2b, 5.2c, and 5.2d are overlaid with the probability distribution p(E) in blue, depicted in discrete steps.

While the period of oscillation is identified nicely by HaMMy, it is obvious in Fig. 5.2a that the assumption of distinct states in the trajectory poses a major hindrance to the hidden-Markov analysis. In fact, this is a system that does not possess “states” and lifetimes, but a system that merely oscillates between two efficiency extremes. This is illustrated in the efficiency distributions shown in Fig. 2b and 2d as well. While the efficiency distribution in Fig. 5.2b shows the occupation of a broad range of efficiencies,

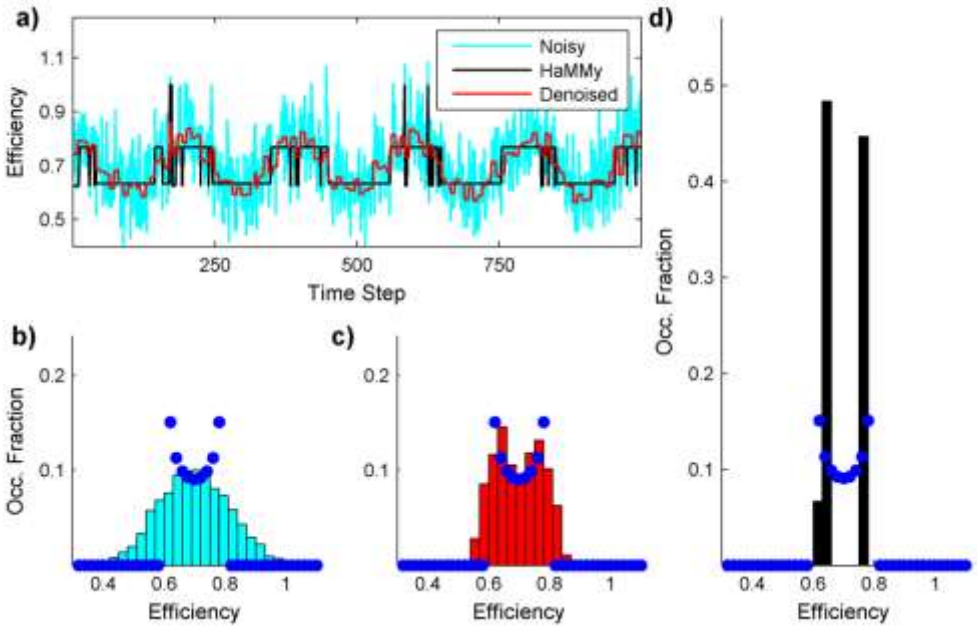


Figure 5.2. Denoising an oscillatory system. a) The original, shot-noise laden efficiency trajectory (*cyan*) is overlaid with the denoised efficiency trajectory (*red*) and the efficiency trajectory generated by HaMMy (*black*). b) The efficiency distribution of the original data. c) That of the denoised data. d) The distribution of efficiencies generated by HaMMy. The efficiency probability distribution p(E) is overlaid in blue on each efficiency distribution.

that which is produced by HaMMy in Fig. 5.2d shows the molecule to occupy two major conformations.

In contrast, both the denoised trajectory and the denoised efficiency distribution show improved agreement with the noisy data. It is seen in the trajectories shown in Fig. 5.2a that the denoised data constitutes a better representation of this system's dynamics than does that produced by the hidden-Markov analysis. In comparing the efficiency distributions, one can also see that, although there is a slight discrepancy that

arises at the efficiency extremes of $E_c \pm E_0$ due to a small amount of remaining noise in the trajectories, there is good agreement between the efficiency probability distribution p(E) and the denoised distribution. The distribution produced by the wavelet denoising algorithm in Fig. 5.2c is, therefore, a more accurate representation of the system's actual properties. Thus, the comparison shown in Fig. 2 demonstrates the value of the wavelet denoising algorithm when applied to data derived from systems that exhibit non-Markovian kinetics, and/or do not possess distinct conformational states.

Section 5.4: Denoising a System with Indistinguishable States

Fig. 5.3a depicts the efficiency distribution of a two-state equilibrium that was simulated using kinetic Monte Carlo methods. As shown by this efficiency distribution, the two states, having mean efficiencies of 0.89 and 0.81, are indistinguishable in the presence of shot-noise. However, it is clearly shown by Fig. 5.3b that the states in the underlying equilibrium are distinguishable after the trajectories are denoised by the wavelet denoising algorithm.

To demonstrate the wavelet denoising algorithm's value as a companion to other methods, the hidden-Markov model of HaMMy [20] was used to further identify the central efficiency the states as well their relative populations, and the statistical correlation method described by Schenter, *et al* [19] was used to extract the kinetics that underlie the equilibrium. This method requires that a state be assigned to each time step in the trajectories, and HaMMy was also used to accomplish this task.

Acting on the shot-noise laden trajectories of the distribution shown in Fig. 5.3a produces the idealized efficiency distribution in Fig. 5.3c. Performing the same operation on the denoised trajectories produces the distribution shown in Fig. 5.3d. It is seen that, while HaMMy ably identifies the central efficiency of each state, the equilibrium constant, K_{eq}(obs), that is produced by this operation differs from the actual equilibrium constant, K_{eq}, by 58%. In contrast, the equilibrium constant produced by acting on the denoised trajectories with the hidden-Markov model, K_{eq}(den), differs from the actual value of K_{eq} by only 4.4%. In addition, the central efficiencies that are produced by acting on the denoised trajectories with the hidden-Markov model differ only trivially from the actual efficiencies. It is quite obvious from this comparison that the denoised data produces an accurate representation of the thermodynamics that underlie the states this equilibrium.

The autocorrelation curves shown in Fig. 5.3e were produced from idealized trajectories produced by HaMMy. Fitting these curves to exponential decays allows for the extraction of rate constants, and thus for the extraction of mean lifetimes of each state in the equilibrium. These lifetimes are also reported in Fig. 5.3e. Inspection of each of the autocorrelation curves in Fig. 5.3e reveals good agreement between the denoised and simulated curves, but poor agreement between the observed and simulated curves. Also, as seen in Fig. 5.3e, the lifetimes of each state, as calculated from the autocorrelation of the noisy trajectories, are 0.632 s and 0.547 s. These values differ from the simulated lifetimes by 34.6% and 3.4%, respectively. The lifetimes produced by the autocorrelation of the denoised data are 0.91 s and 0.519 s,

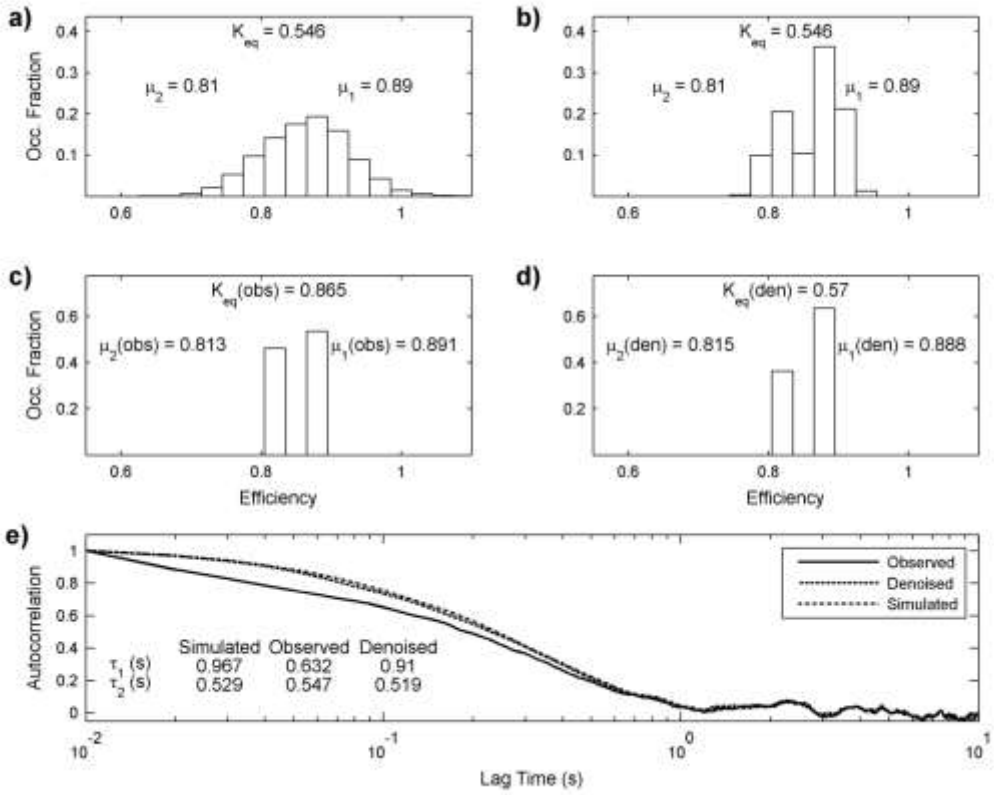


Figure 5.3. Denoising a system with indistinguishable states. a) The efficiency distribution of the simulated equilibrium showing the central efficiency of each state, μ_1 and μ_2, as well as the simulated equilibrium constant, K_{eq}. b) The efficiency distribution produced after denoising the trajectories with the wavelet denoising algorithm. c) The distribution of efficiencies produced by acting on the noisy trajectories with HaMMy, showing the central efficiencies of each state, μ_1(obs) and μ_2(obs), as well as the equilibrium constant, K_{eq}(obs), produced by this operation. d) The distribution of efficiencies produced by acting on the denoised data with HaMMy, showing central efficiencies of each state, μ_1(den) and μ_2(den), as well as the equilibrium constant, K_{eq}(den), produced by this operation. e) Autocorrelation curves

produced from the trajectories generated by HaMMy acting on the noisy data (*solid*), the denoised data (*dotted*), and the simulated state trajectories (*dot-dash*). The average lifetimes of each state, as extracted from the autocorrelation curves, are shown in the inset table for each of the simulated, observed, and denoised data.

respectively, and these values differ from the simulated lifetimes by 5.9% and 1.9%, respectively.

Given the accuracies of each state's central efficiency and of the extracted equilibrium constant, we conclude that denoising the trajectories of this simulated system with the wavelet denoising algorithm successfully removes noise while retaining the actual data. Furthermore, acting on the denoised trajectories with the hidden-Markov model allows for the extraction accurate kinetic data, thereby completely characterizing two states in an equilibrium that were, prior to denoising, indistinguishable.

Section 5.5: Denoising a Simulated System with Well-Defined States and Dynamics

As a means to quantify the effects of the wavelet denoising algorithm to smFRET trajectories, a two state equilibrium having an equilibrium constant of 0.33 was simulated using kinetic Monte Carlo methods. The states were assigned mean efficiencies of 0.2 and 0.8, respectively, with the state having mean efficiency of 0.8 being favored in the equilibrium. Acceptor and donor photon trajectories that include shot-noise were constructed from the simulated trajectories. Each of the simulated

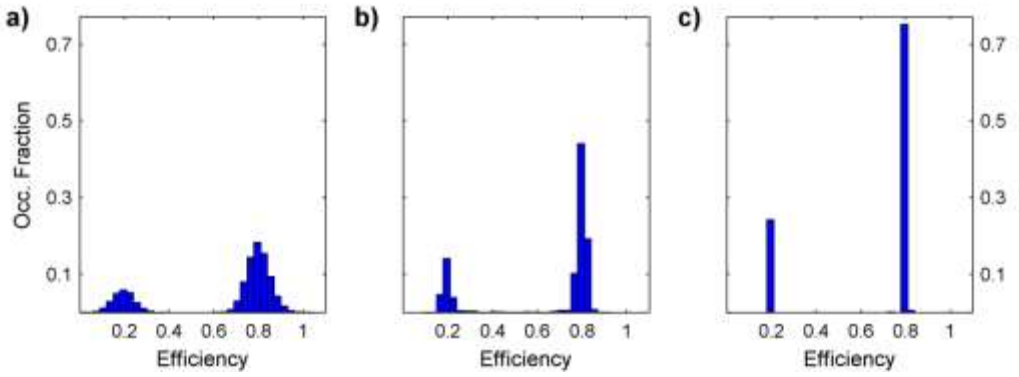


Figure 5.4. Efficiency distributions of a simulated two state system. a) The shot-noise induced efficiency distribution of a simulated two state system having states with mean efficiencies of 0.2 and 0.8. b) The efficiency distribution of the system after denoising. c) The efficiency distribution as generated by the hidden-Markov model HaMMy [20]**.**

trajectories was denoised by the wavelet denoising algorithm as well as the hidden-Markov model (HaMMy) described by McKinney, *et al* [20].

Fig. 5.4a shows the efficiency distribution generated by the simulated photon trajectories. Each state's standard deviation was calculated as measurement error and found to be approximately 0.06 efficiency units for both states at a mean of 220 total photons per time step. The standard deviation of each state's distribution was also found by least-squares regression to normal distributions, and these values show good agreement as the standard deviations of both states were again found to be approximately 0.06 efficiency units.

Figure 5.4b shows the efficiency distribution produced after wavelet denoising. The standard deviation of each state's efficiency distribution was again estimated by

least-squares regression to normal distributions. These values were again found to be approximately equal at 0.02 efficiency units, translating to a reduction in the width of each state's efficiency distribution of 300%. It is shown in Figure 5.4c that HaMMy identifies each state nicely by collapsing the broad efficiency distribution of each of the states nearly to a single efficiency. Although the denoising algorithm does not reduce the width of each distribution as radically as HaMMy, the wavelet denoising algorithm is not, as is HaMMy, constructed to do so. The comparison does show, however, that the denoising algorithm narrows each state's efficiency distribution by 300%, thereby proving its viability in the case of a system having well-defined states and kinetics while offering the advantage of making no *a priori* assumptions regarding the state's thermodynamic or kinetic properties.

Section 5.6: Denoising an Experimental smFRET Trajectory

Here we examine a single smFRET trajectory obtained in the aV aptamer studies described previously [14]. Fig. 5.5a depicts the fluorophore-emitted photon signals in 10 ms time steps, with the acceptor signal shown in red and the donor signal shown in blue. The trajectory includes photoblink anomalies of both the acceptor and donor varieties on both short and long time-scales. Additionally, intensity variation in of each of the signals is observed, and arises both from energy transfer efficiency fluctuations and shot-noise. Fig. 5.5b shows the photoblink-filtered and wavelet-denoised versions of the same acceptor and donor trajectories. Time steps identified as photoblinks were removed from the acceptor [77] and donor (cyan) trajectories shown, and each is overlaid with its wavelet-denoised complement (denoised acceptor in black and

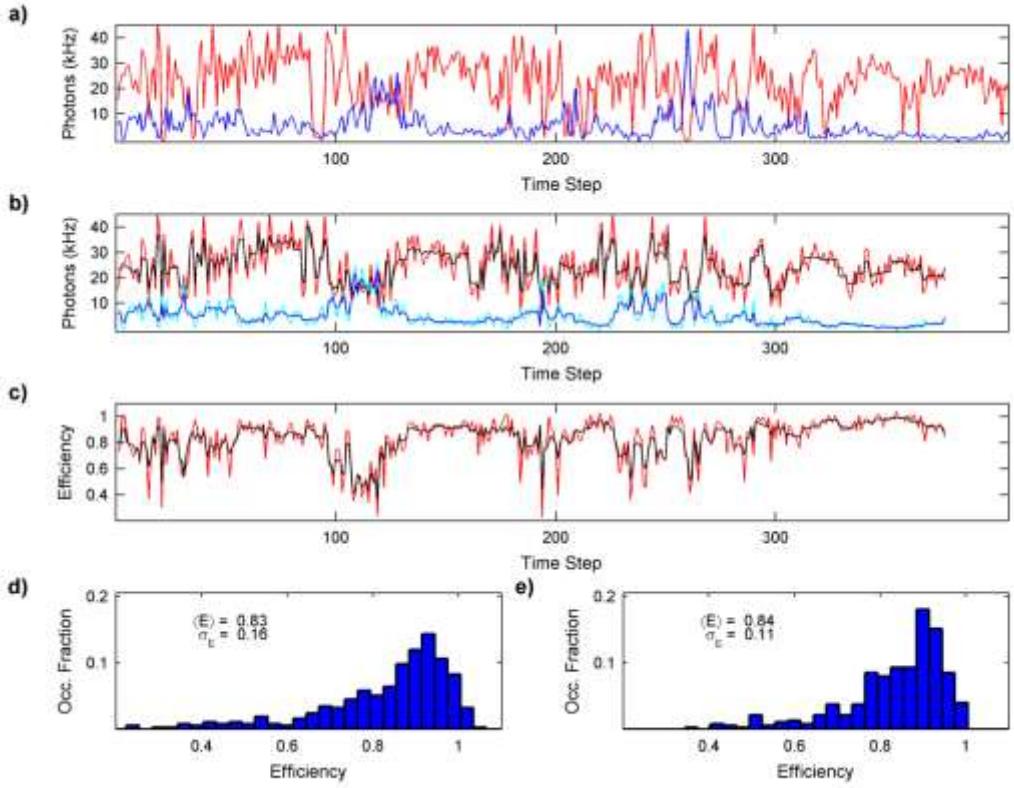


Figure 5.5. Denoising an experimental smFRET trajectory. a) The original fluorophore-emitted acceptor [77] and donor (blue) photon trajectories. b) The original acceptor photons [77] are overlaid with their denoised counterparts (black), and the original donor photons (cyan) are overlaid with their denoised counterparts (blue). c) The smFRET efficiency calculated from the original acceptor and donor photon counts [77] is overlaid by that calculated from the denoised acceptor and donor photon counts (black). d) The efficiency histogram generated by the noisy data in b). e) The efficiency histogram generated by the denoised data shown in b).

denoised donor in blue). Fig. 5.5c shows the energy transfer efficiency calculated from each complementary pair of signals in Fig. 5.5b using Eq. S5. The observed efficiency

trajectory was calculated at each time step using the photoblink-filtered pair of acceptor and donor signals shown in Fig. 5.5b, and this trajectory is overlaid in black by the denoised efficiency calculated at each time step using the wavelet-denoised acceptor and donor signals shown in Fig. 5.5b. Fig. 5.5d shows a histogram of the observed efficiencies in Fig. 5.5c, and Fig. 5.5e shows that of the denoised efficiencies shown in Fig. 5.5c.

Also, as shown by Fig. 5.5b, the wavelet-denoising algorithm proves to be quite effective in its application to experimental data. While fluctuations in the signals' intensities due to experimental considerations such as apparatus limitations and fluorophore orientations are unavoidable, it is shown in the denoised signals that small fluctuations arising from quantifiable sources are virtually eliminated. Of equal, or perhaps greater, importance, it is also shown that large intensity fluctuations that are induced by conformational changes in the system are allowed to remain.

The significance of these aspects is clear, and is illustrated by the efficiency trajectories in Fig. 5.5c. The observed trajectory of calculated efficiencies fluctuates wildly from time step to time step due to small, insignificant changes in the signal intensity on one or both of the detection channels. The denoised complement (black) to this trajectory, however, does not exhibit such excessive fluctuation. We still observe major fluctuations in efficiency, and these are virtually unchanged from the major fluctuations we see in the observed efficiency trajectory. Anomalously large and fast efficiency fluctuations have been reduced, and we observe a smoother trajectory as well as more accurate representation of our physical system.

Figs. 5.5d and 5.5e further validate this point by showing that, while the shape of each efficiency distribution is approximately the same, and their mean efficiencies are virtually unchanged, the denoised standard deviation is reduced by approximately 30%. This validates that occurrences of anomalously large or small efficiencies have been reduced. On the whole, the application of the photoblink detection and wavelet-denoising algorithms is shown by Figure 5.5 to improve the quality of this experimental smFRET trajectory.

Section 5.7: HIV-1 TAR DNA: Denoising a Single State Experimental System

Here we apply the photoblink detection and wavelet-denoising algorithms to a collection of smFRET trajectories. The trajectories are collected from studies, described previously [83], on the transactivation response (TAR) region of HIV-1 viral DNA. The collection contains nearly 16,000 data points, and was chosen to represent a predominantly single state system. The secondary structure of the TAR hairpin is shown in Fig. 5.6a, and the depicted conformation is expected to yield smFRET efficiencies approaching unity.

Application of the photoblink detection algorithm to this collection of trajectories yields, after computation of energy transfer efficiency at each of the individual data points, the global efficiency histogram that is shown in Fig. 5.6b. As expected, the system does produce smFRET efficiencies approaching unity, and the normal distribution about the mean energy transfer efficiency of 0.96 signifies that we do observe a predominantly single state in this collection of trajectories. The width of

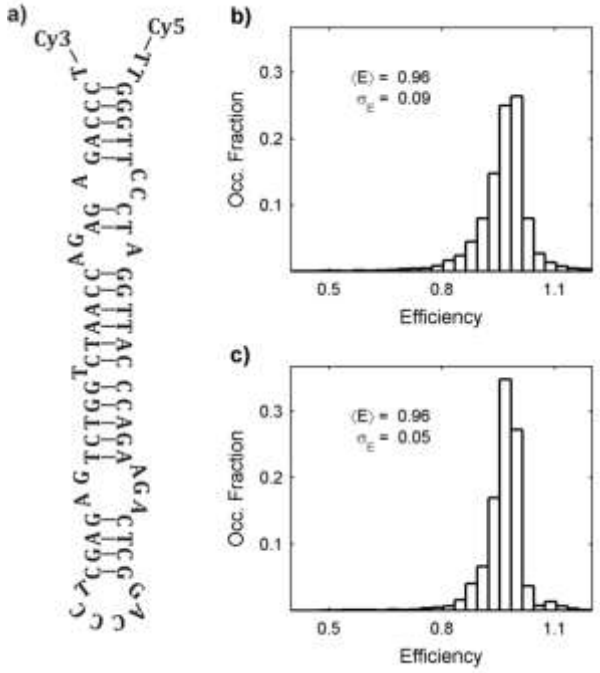


Figure 5.6. Single state TAR DNA. a) The secondary structure of TAR. B) The efficiency distribution of observed and blink-filtered data acquired from experiments involving TAR in 2 mM Mg^{2+} buffer solution. c) The corresponding denoised data.

this distribution is small, again signifying a single state, and we observe this value to be 0.09 efficiency units.

Calculating and compiling each data point's energy transfer efficiency from wavelet-denoised complements produces the global efficiency histogram shown in Fig. 5.6c. While we do not observe a shift in the mean smFRET efficiency, we do observe a sizeable reduction in the magnitude of the distribution's standard deviation. The width of this distribution, while small to begin with, is reduced by approximately 45% by the denoising algorithm. The significance is clear – in the characterization of a single state system, eliminating the artifacts of shot-noise and photoblinks results in a more precise representation of the state's structure and energetics.

Section 5.8: Acceptor Photobleaching: Denoising a 2-State Experimental System

As a model two state system, irreversible acceptor photobleaching from a high efficiency state is chosen. This collection of trajectories was also chosen from studies on

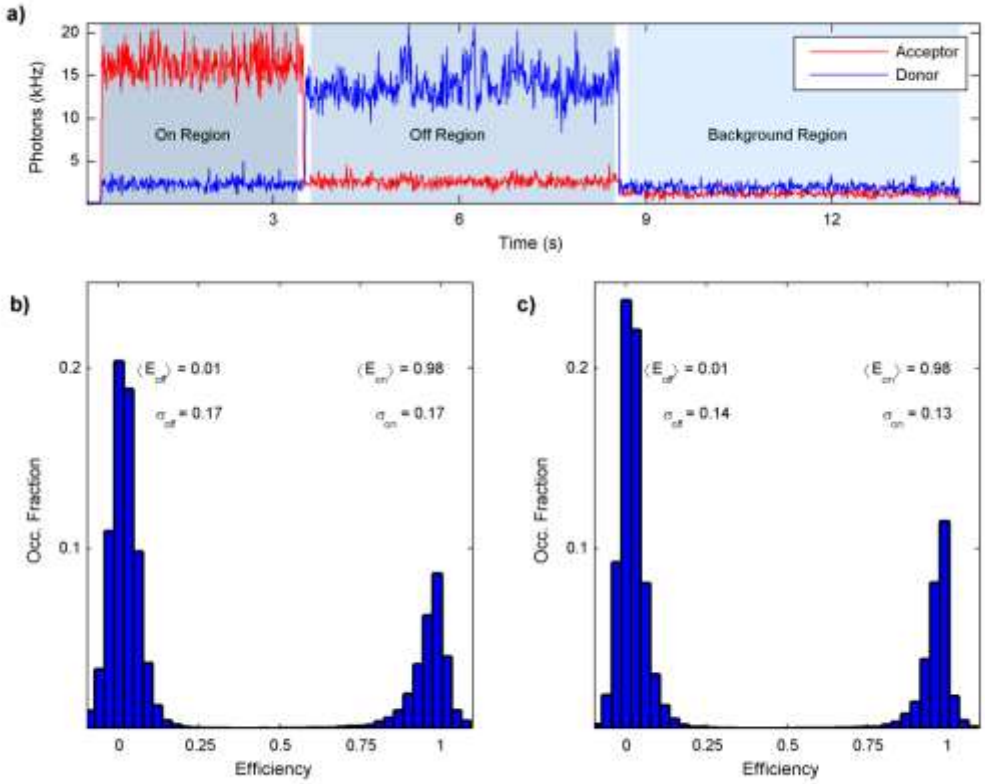


Figure 5.7. Irreversible acceptor photobleaching as a purely two state system. a) A model trajectory where the "on" state is represented by the region before the acceptor photobleach, and the "off" state is represented by the region after the acceptor photobleach. b) The efficiency distribution compiled from blink-filtered trajectories. c) The corresponding denoised efficiency distribution.

the TAR region of HIV-1 viral DNA described previously [83]. This system is treated as a purely two state system: energy transfer is either "on", as shown to the left in Fig. 5.7a, or "off", as shown in the center of Figure 5.7a. As the crosstalk region of this collection of trajectories is being considered as part of the FRET region, the crosstalk value x is fixed to a characteristic value of 0.11. Additionally, given that one of the states in our

model has an efficiency that is expected to approach zero, the previously discussed caveat arises. We circumvent this caveat by simply marking the time step at which the acceptor photobleach occurs. The photoblink detection method is applied to time steps prior to this time step as previously described, and for time steps after the photobleach we substitute the donor photon signal for the acceptor and proceed in the same manner.

Fig. 5.7b shows a global efficiency histogram of approximately 18,000 blink-filtered data points in this collection of smFRET trajectories. The "on" state is represented to the right, and shows a mean efficiency of 0.98 with a standard deviation of 0.17. The "off" state is represented to the left, and this distribution shows a mean efficiency of 0.01 with equal width of 0.17. To avoid bias, instead of selecting the limits of each state manually, each mean and standard deviation is produced by a least-squares regression to the sum of two normal distributions.

Figure 5.7c shows the wavelet-denoised complements of Figure 5.7b. Again we obtain the characteristics of each state's efficiency distribution by a least-squares regression, and again we find dramatic improvement. While each state's mean value has remained unchanged, their distributions have narrowed significantly. In the case of the "off" state, we see that the distribution has narrowed by approximately 20%, and in that of the "on" state, we see a larger narrowing of just under 25%. As such, we conclude that the wavelet-denoising algorithm is capable of removing quantifiable noise components of each state's efficiency distribution, resulting in more precise description of each state.

Section 5.9: The aV Aptamer: A Multi-State Experimental System

Trajectories acquired from studies previously reported [14] on the aV aptamer and its interaction with its binding target, vascular endothelial growth factor (VEGF), are chosen to represent a multiple state system. These experimental studies revealed a highly dynamic secondary structure that ranged from the closed hairpin, illustrated in Fig. 5.8a, to an irresolvable continuum of open states with lower smFRET efficiencies. To complicate matters, the aptamer's interaction with VEGF was found to be similarly dynamic. Although the smFRET studies suggested that the VEGF-bound aptamer structure was the open state, quantitative analysis was hampered by contributions of both shot-noise and structural fluctuations to the measured smFRET distributions.

The global efficiency histogram containing approximately 15,000 data points of the aV aptamer in 2 mM Mg^{2+} buffer solution is shown in Fig. 5.8b. The distribution shows a skewed mean with an anomalously large standard deviation that is a result of the trajectories containing photoblinks. Application of the Bayesian photoblink filter to the trajectories results in the distribution shown in Fig. 4c. This efficiency distribution has a mean efficiency of 0.9 with standard deviation of 0.13. Application of the wavelet-denoising algorithm to this collection of trajectories results in the efficiency distribution shown in Figure 5.8d. We observe that the mean efficiency is unaffected, and that the standard deviation has been reduced by 25%. As such, we conclude that the algorithm has the capability to simultaneously refine the distributions of multiple, efficiency states, even if the efficiency state distributions have significant overlap.

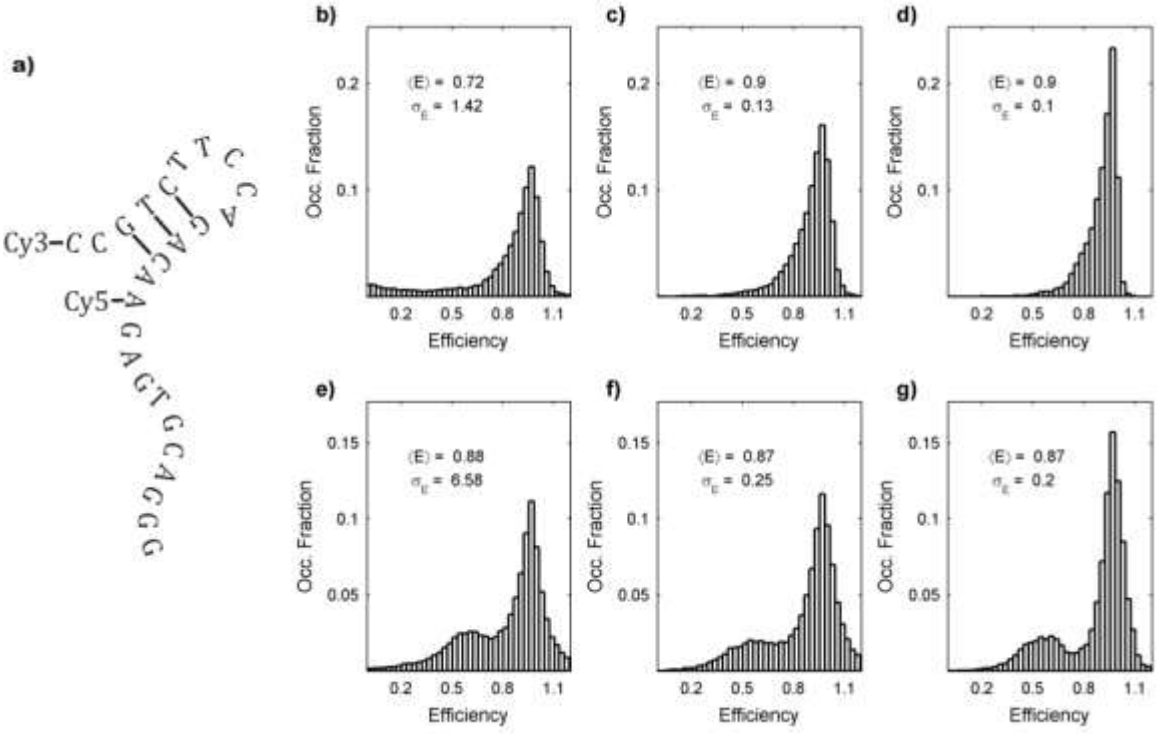


Figure 5.8. The aV aptamer as a multiple state system. a) The 2° of the aV aptamer. b) The observed efficiency distribution of the aV aptamer at Mg^{2+} concentration of 2mM, prior to blink-filtering. c) The blink-filtered efficiency distribution of the aV aptamer. d) The denoised distribution corresponding to b). e) The observed efficiency distribution, prior to blink-filtering, resulting from the addition of 2 μM VEGF. f) The blink-filtered efficiency distribution corresponding to e). g) The denoised distribution corresponding to d).

Figure 5.8d shows a global efficiency histogram of the aV aptamer while in the presence of VEGF that contains approximately 26,000 data points. Again, due to the presence of photoblinks in the trajectories, the distribution shows an anomalously large standard deviation. Despite photoblinks, the VEGF-induced shift in the aptamer's conformational equilibrium shown previously [14] is seen quite clearly in Figure 5.8e. It is not, however,

clear that this shift arises due to a shift in the conformational equilibrium until the application of the Bayesian photoblink filter, which results in the distribution shown in Fig. 5.8f. This collection of trajectories shows an overall mean efficiency of 0.87 with a large standard deviation of 0.25 efficiency units, and shows that a shift in the aptamer's equilibrium is indeed observed in the presence of VEGF.

The wavelet-denoised complement to this collection is shown in Figure 5.8g. Again, the effects of the denoising algorithm are quite clear. While the mean efficiency once again remains constant, the standard deviation is decreased by 20%. More importantly, the shape of the distribution is visibly refined. Although the distribution is broadened, presumably by effects of the fluorophores' respective orientations [113-115], efficiencies representative of a conformation yielding lower efficiencies are noticeably increased, in good agreement with the presumed interaction with VEGF [14]. As a result of Figs. 5.8d and 5.8g we conclude that, while improving the finer aspects of the analysis, the application of the wavelet-denoising algorithm does not affect the overall outcome of the analysis of a system containing a complex combination of multiple and overlapping efficiency states. Furthermore, we conclude that the wavelet denoising algorithm enhances the analysis of this system by confirming the presence of a continuum of irresolvable conformations in the aptamer's conformational equilibrium, as in Fig. 5.8c, as well as improving the visibility of the presumed aV-VEGF interaction as in Fig. 5.8g.

5.10 Conclusions

In conclusion, this chapter has shown that using a two component interpretation of noise observed in smFRET signals allows us to remove the component we can quantify, thereby enhancing the accuracy of these measurements. Wavelet denoising was applied to simulated acceptor and donor trajectories, and significantly decreased the width of a state's efficiency distribution. Additionally, trajectories representing a system showing oscillatory behavior were simulated as a means to demonstrate the efficacy of denoising in complex systems. These simulations showed that the denoised data formed a most accurate representation of the system at hand.

We expect that the methods presented here will have immediate impact on the smFRET community. We also expect the method to have a broad scope of applicability because the wavelet denoising algorithm is not strictly limited to smFRET measurements. Many wavelet-based applications have already been realized, and this particular method requires only slight adjustment for application to other types of time-series photon measurements, single-molecule or otherwise.

An executable application containing wavelet denoising algorithm is available at http://www.lrg.rice.edu.

Chapter 6

Improved Resolution of Complex Single-Molecule FRET Systems via Wavelet Shrinkage

The contents of this chapter are adapted from an article originally published in *Journal of Physical Chemistry B* on January 9, 2011.

Taylor, J. N.; Landes, C. F. "Improved Resolution of Complex Single-Molecule FRET Systems via Wavelet Shrinkage." *. J . Phys. Chem. B.* ***2011****.* 115(5):1105-1114.

ABSTRACT

The resolution of complex interactions found in single-molecule fluorescence resonance energy transfer (smFRET) experiments is hindered by noise. Wavelet shrinkage is proven to reduce noise, but traditional methods introduce artifacts when acting on discontinuous signals, such as those acquired in smFRET experiments. Modifications to the basic method that are specific to smFRET are developed and tested on simulated systems. Use of the Haar wavelet basis produces the most optimally denoised estimates. We also assess various thresholding methods, develop a time-localized noise estimator, and implement a translation-invariant wavelet transformation to reduce

artifacts associated with discontinuities and inadequate distinction of noise. The time-local estimator enhances noise reduction by 5-20 %, and translation-invariant transformation nearly eliminates the aforementioned artifacts. Kinetic parameters extracted from denoised estimates are accurate to within 5 % of the simulated values. Overall, the improved resolution results in the complete and accurate characterization of both simple and complex smFRET systems.

6.1 Introduction

Rapid advancement of single-molecule spectroscopic techniques has led to the recent discovery of complex energy landscapes and heterogeneous mechanistic pathways that drive many biological interactions [116-120]. Among these techniques, single-molecule fluorescence resonance energy transfer (smFRET) has proven to be a useful tool in the recognition of complexity and mechanistic heterogeneities [14, 121-123]. Despite its usefulness however, various forms of noise limit the abilities of smFRET [5, 22, 124-127]. Specifically, shot-noise is a random error that contributes uncertainty to the photon measurements [5, 126], affecting the resolution, i.e., the ability to distinguish distinct states in an efficiency distribution, of smFRET experiments. Several methods, including hidden-Markov models [21, 125], applications of information theory [128, 129], photon statistics in the context of two colors [16], maximum likelihood and Bayesian inferential [130] estimation of change points using Poisson statistics, wavelet correlation [22], and wavelet shrinkage [5] have been developed and applied to single-molecule data as a means of extracting more accurate information about the system under observation. The focus of this work is wavelet shrinkage, which

is a signal processing tool whose primary resource is the use of wavelet transformation [5, 97, 100, 103, 110, 131-138]. Primary applications of wavelet transformation include data compression and signal processing, but analysis applications such as density estimation [139] and change-point identification [136] are known as well. Wavelet shrinkage is known to reduce noise contributions in various forms of analytical data while minimally affecting its overall integrity [5, 97, 138, 140].

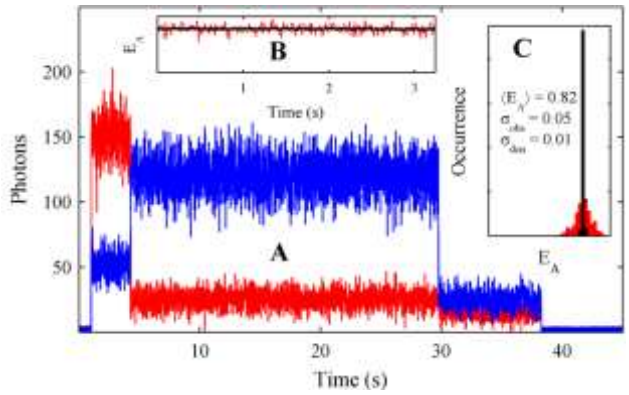


Figure 6.1. Decreased efficiency broadening after wavelet shrinkage. A) A full smFRET trajectory containing acceptor (red) and donor (blue) photon trajectories. B) The observed efficiency trajectory (red), calculated from the acceptor and donor trajectories in A), is overlaid with the denoised efficiency trajectory (black). C) The denoised efficiency distribution (black) overlays the observed distribution (red), and is shown to collapse to a narrow peak representing the single efficiency state.

Fig. 6.1 illustrates the application of wavelet shrinkage, as described previously [5], to an experimental smFRET trajectory that represents a single efficiency state. Fig. 6.1A shows the entirety of the acceptor (red) and donor (blue) photon trajectories. Fig. 6.1B shows the observed efficiency trajectory (red) overlaid with the denoised efficiency trajectory (black), and Fig. 6.1C shows the observed (red) and denoised (black) efficiency distributions. The ability of wavelet shrinkage to reduce shot-noise-induced

broadening in smFRET trajectories is clearly demonstrated in Fig. 6.1 by the collapse of the denoised distribution to a nearly singular value. Our previous work [5] also demonstrates the ability of wavelet shrinkage to extract previously inaccessible information from complicated systems. Despite the successful application of wavelet shrinkage to smFRET trajectories, it is possible, and therefore a sensible aim, to do better still.

There are two main targets for improvement. The first is the introduction of previously absent artifacts to the denoised estimates produced by wavelet shrinkage. These artifacts are introduced near discontinuities in smFRET trajectories, and arise primarily as a result of the translation dependence of the wavelet transformation itself [141]. The second target is inadequate estimation of the noise level. Transitions (i.e., discontinuities) in the system under observation result in a nonstationary noise level, and the use of its average value throughout a trajectory yields underestimation of the noise level when the efficiency is above its mean value and overestimation when the efficiency is below its mean value. There are, however, methods available that will reduce the impact of these issues [135, 141, 142].

Contributions of the artifacts introduced by translation dependence of the wavelet transform can be reduced by forcing the transformation to be translation-invariant. This can be accomplished by a method known as cycle-spinning [141], in which the original photon trajectory is subjected to a range of time-translations, and wavelet shrinkage is independently performed on each translation. The translations are then removed, and the results are averaged to produce a denoised estimate without

translation dependence and a reduced presence of translation dependent artifacts. The error introduced by suboptimal estimation of the noise level can be reduced through variations in the application of wavelet shrinkage, and through the use of the time-local noise estimator introduced below.

In this chapter, these limitations are addressed by first tackling the most important question: which wavelet basis produces the most optimally denoised estimates? Different bases are defined by the algorithms by which they separate the signal into orthogonal components, and thus it is expected that one or more basis algorithms might be better suited for smFRET applications. In this work, we apply various combinations of wavelet basis, shrinkage method, noise estimator, and cycle-spinning to the model systems as a means of selecting the combinations that produce the most optimally denoised estimates. A noise estimator that is specific to smFRET photon trajectories is introduced, tested, and compared to the results produced by the universal estimator [110]. The variations producing the most optimally denoised estimates were further analyzed, and found to show improved agreement with the simulated model systems. Parameters containing information about kinetic aspects and state efficiency distributions were also extracted from the denoised estimates and compared with those of the simulated values.

6.2 Methods

6.2.1 *Data Simulation*

The model systems were simulated via kinetic Monte Carlo methods. Each data set is comprised of 30 individual trajectories whose minimum length was selected randomly from a normal distribution with a mean of 7 seconds and a variance of 4 seconds. Initial states for each trajectory were selected uniformly with each available state having equal selection probability. Lifetimes were generated sequentially from the input escape rate constants, and final states for each transition were selected with probability proportional to the transition rate constant from the current state to the final state. The central efficiencies at each time step were used in combination with mean total photons per time step $\langle \mathrm{N_{tot}} \rangle$ to generate acceptor and donor photon trajectories. Total photons per time step included fluorophore-emitted photons from both fluorophores as well as background photons on both simulated detection channels. The signal to background ratio for $\langle \mathrm{N_{tot}} \rangle$ was 2.3, and the donor to acceptor background ratio was 2. The sum of fluorophore-emitted photons and background photons on each channel was selected randomly at each time step from a normal distribution with equal mean and variance, as calculated from the central efficiency at the time step. Shot-noise induced efficiencies were calculated from the simulated acceptor and donor photon trajectories, after correction for background and crosstalk as described in [5].

The characteristics of the two state model system are shown in Fig. 6.2. Sample acceptor and donor photon trajectories are shown in Fig. 6.2A, and the efficiency trajectory generated from these photon trajectories is shown in Fig. 6.2B. Fig. 6.2C shows the shot-noise induced efficiency distribution of the two state model system, and simulation parameters are described in Fig 6.2D. The system is comprised of two states,

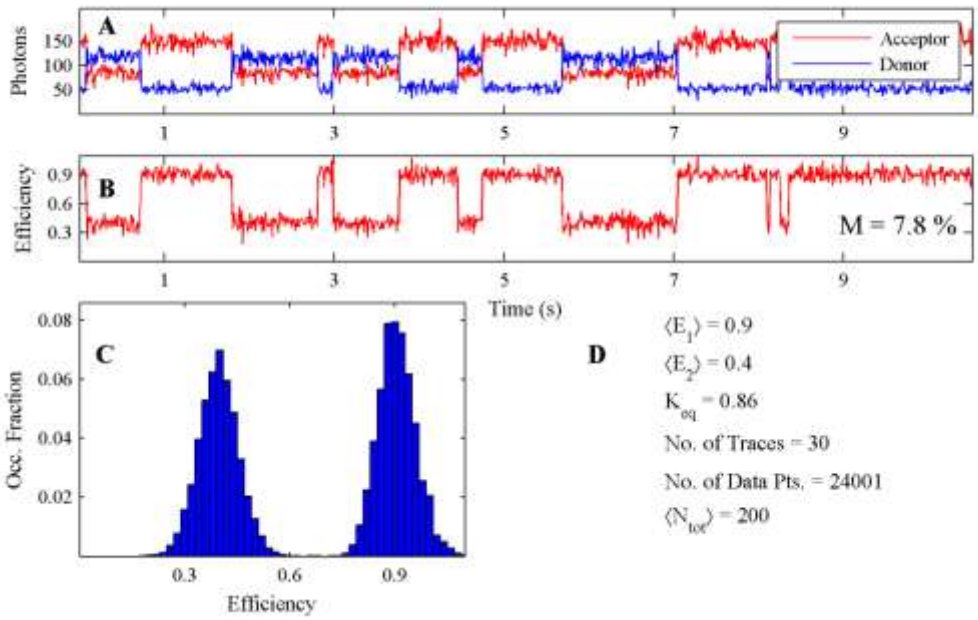


Figure 6.2. Details of two state simulation. A) shows representative acceptor [77], and donor (blue) photon trajectories. B) shows the efficiency trajectory generated by the photon trajectories in A), along with the mean absolute percent error, M, for the entire simulated data set. C) shows the distribution of efficiencies generated by compilation of all 30 trajectories in the model system, and D) reports simulation parameters.

having mean efficiencies of 0.9 and 0.4, and the equilibrium constant K_{eq} resulting from the simulation is 0.86, favoring the higher efficiency state. Input lifetimes of each state were 0.5 s, corresponding to input transition rate constants of 2 Hz.

Likewise, the characteristics of the four state model system are shown in Fig. 6.3. This system was simulated with four states having central efficiencies of 0.9, 0.75, 0.6, and 0.5. The states obey a sequential equilibrium, with each state being allowed to transition only to its neighboring state(s). Total escape rate constants from each state

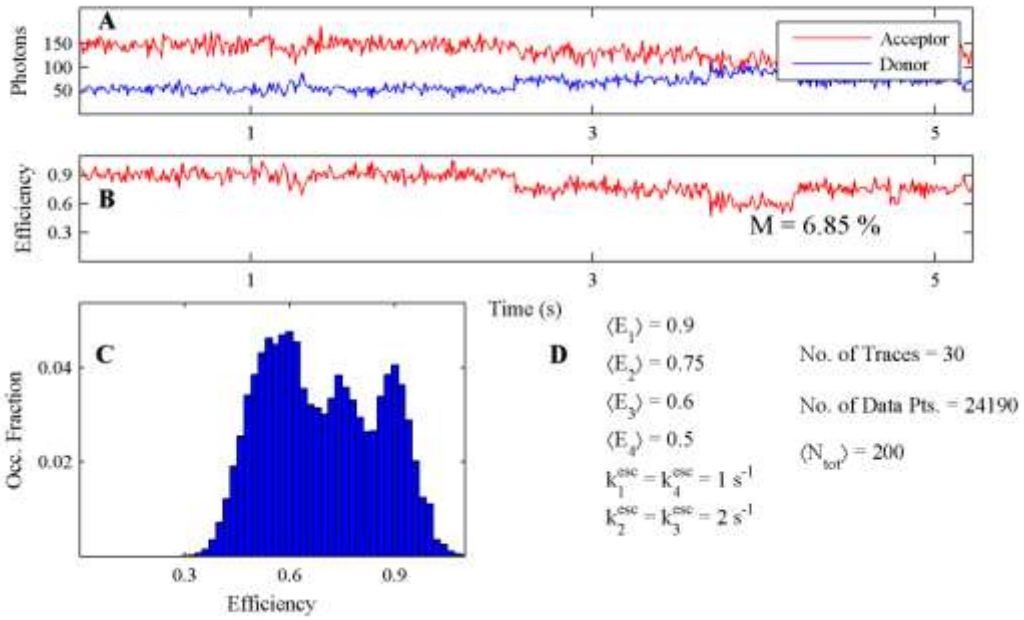


Figure 6.3 Details of four state simulation. A) representative acceptor and donor photon trajectories. B) the efficiency trajectory generated by the photon trajectories in A), along with the mean absolute percent error, M, of the entire model system. C) Efficiency distribution compiled from all trajectories in the system, and D) reports simulation parameters.

were 1, 2, 2, and 1 Hz, respectively. States having two allowed transitions had equal probability of transition to either the lower or higher efficiency states, with transition rate constants of 1 Hz.

6.2.2 *Wavelet Shrinkage*

Wavelet shrinkage is described in detail in our previous work.[5] Shrinking wavelet coefficients to obtain a denoised estimate combines multiresolution analysis[101] and an additive white Gaussian noise approximation.[110] Wavelet

shrinkage consists of three main steps: 1) transform the signal to the wavelet domain, 2) shrink the wavelet coefficients to reduce noise, and 3) invert the wavelet transform to obtain the denoised estimate. Photon signals $\boldsymbol{N}$ ($= \{N(t), N(t + \Delta t), \dots\}$) containing shot-noise are assumed to have the form

$$\boldsymbol{N} = \boldsymbol{S} + \sigma \boldsymbol{Z} \ . \qquad (6.1)$$

Here, $\boldsymbol{S}$ is the "true" signal we wish to recover, σ is the shot-noise magnitude, and $\boldsymbol{Z}$ are independently and identically distributed (i.i.d.) noise components on the standard normal distribution. Signals are transformed to the wavelet domain via digital filtering, in which the signal is convolved with a pair of digital filters to obtain two orthogonal components at reduced time resolution. Once transformed, the shot-noise portion of the signal, assumed to lie within the high frequency component of the transformation, is reduced by the shrinkage procedure. With reduced magnitudes, the high frequency components are recombined with the low frequency information from the other orthogonal component, and the signal is reconstructed via inversion of the wavelet transformation.

The careful reader may contend that the statistics of photon collection are not normal but Poissonian. This discrepancy can be resolved by considering the properties of Poisson processes. More often than not, smFRET measurements are time series photon measurements that may be described as non-homogenous Poisson processes. The mean number of detected photons is a time-dependent random variable in the non-homogeneous case, which leads to a time-dependent variance as well. However, if the

underlying process is stationary, the mean number of photons λ and the variance (λ) also become stationary as the number of samples increases to infinity.

The source of abnormality in the distributions of non-homogeneous Poisson processes lies in the higher order moments of the distribution, skewness and kurtosis. For low numbers of events, higher order moments skewness and kurtosis add asymmetry and irregularity, respectively, to the distribution. However, these moments trend to zero quickly[143] – skewness as $\lambda^{-1/2}$ and kurtosis as λ^{-1} – as the mean number of events per unit time increases, causing them to become negligible with respect to the mean and variance.

As a result, Poisson-distributed photon measurements become well-approximated by normal distributions. Applying the central limit theorem to the normal approximation yields an estimate of the underlying noise distribution, and it is easy to show that the normal approximation of a stationary Poisson process reduces to a standard normal distribution under such conditions. The additive white Gaussian noise approximation is therefore a sufficient approximation of the underlying noise distribution in a Poisson-distributed process contaminated by a stationary white noise. For a non-stationary process, the differences that arise in the noise distribution may only be related to the magnitude of the noise

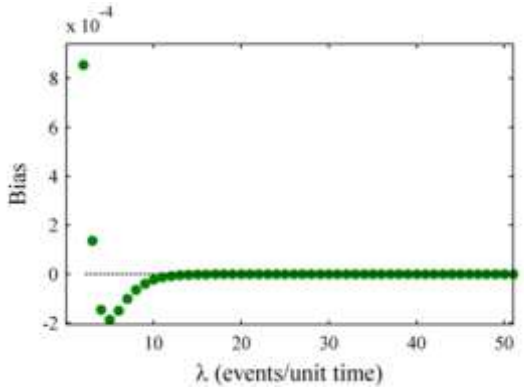


Figure 6.4 Bias of the normal approximation to a Poisson distribution.

fluctuations. That is, fluctuation in the number of detected events, in the form shot-noise, remains a zero-mean process, and it is only the width of the distribution that changes. These criteria are satisfied if fluctuation in the system under observation is sufficiently slow and the mean number of detected photons remains sufficiently large. Fig. 6.4 illustrates this point.

6.3 Options in Wavelet Shrinkage

Wavelet shrinkage requires that several choices be made prior to its application, and these choices are the determining factors in its effectiveness. Among these choices, the most important is, as discussed above, undoubtedly that of wavelet basis. Scores of wavelet families, such as the Daubechies [132], Coiflet [100, 138], and biorthogonal [100, 138] families, to name a few, are available for the choosing, and once the choice of family has been made, there are still more options to consider within each family. Furthermore, after a wavelet basis has been chosen, there is a variety of shrinkage methods to consider. The most commonly used methods are hard and soft thresholding [103, 110] but other methods, such as SUREShrink [133], based on Stein's unbiased estimator of risk, non-negative garrote [134], and firm thresholding [135], offer differing properties that are advantageous in various circumstances. In addition to these choices, there are other options and modifications, such as cycle spinning and noise level estimation, which will also be addressed below.

6.3.1 Wavelet Basis

Firstly, we address the choice of wavelet basis. This choice depends most heavily on the objective of the wavelet transformation, which, in this case, is wavelet shrinkage as a means to remove shot-noise from discrete-time photon trajectories. As such, we consider only those wavelet families that are capable of discrete wavelet transformation, such as those mentioned above. Other aspects to consider are the properties of the photon trajectories acquired in smFRET experiments, as well as the properties of shot-noise.

Assuming the Poisson process under observation is stationary, the mean and variance λ of the Poisson distribution is also stationary. Transforming to the frequency domain, we find that a stationary white noise is uniformly distributed over all frequencies. For a non-stationary process that satisfies the criteria above, the noise spectrum does not change in uniformity, but only in magnitude.

A wavelet basis may be represented in frequency space by its frequency response, which, in the case of a wavelet filter, is the Fourier transformation of the digital filters representing the wavelet and scaling functions. Details are provided in the appendix to this chapter.

6.3.2 *Shrinkage Methods*

We now consider the choice of shrinkage, or thresholding, method. As mentioned above, the most common methods are soft and hard thresholding.[103, 110] These thresholding operators are nonlinear shrinkage operators that use an established noise level, along with a coefficient defined by the risk associated with the magnitude of

a Gaussian white noise, to generate a threshold below which a wavelet coefficient is considered to be purely noise. The hard thresholding operator is the simplest of these, in which thresholded wavelet coefficient D_i^{τ} is obtained via comparison of the raw wavelet coefficient D_i with the threshold τ as in Eq. 6.2.

$$D_i^{\tau} = \begin{cases} 0 & |D_i| \leq \tau \\ D_i & |D_i| > \tau \end{cases} \tag{6.2}$$

Similarly to hard thresholding, soft thresholding sets wavelet coefficients with magnitude smaller than the threshold to zero, but instead of keeping the coefficients with magnitude larger than the threshold as they are, these coefficients are reduced by the magnitude of the threshold as in Eq. 6.3.

$$D_i^{\tau} = \begin{cases} 0 & |D_i| \leq \tau \\ sgn(D_i)(|D_i| - \tau) & |D_i| > \tau \end{cases} \tag{6.3}$$

Other methods include SUREShrink, also introduced by Donoho[133], which is a decomposition level-dependent technique that considers the density of nonzero wavelet coefficients at a certain decomposition level, and uses Stein's unbiased estimator of risk to calculate the minimax threshold. Still other methods include the non-negative garrote[134] and firm thresholding[135] operators of Gao and Bruce. These methods attempt to mediate between hard and soft thresholding. The non-negative garrote, or simply garrote, operator is a scaled version of soft thresholding as shown by Eq. 6.4.

$$D_i^{\tau} = \begin{cases} 0 & |D_i| \leq \tau \\ D_i - \tau^2/D_i & |D_i| > \tau \end{cases} \tag{6.4}$$

Firm thresholding uses two threshold values, a lower threshold, below which the magnitude of a wavelet coefficient is considered to be definitely noise, and an upper threshold, above which a wavelet coefficient is considered to be definitely non-noise. Gao and Bruce generate the two thresholds in [135] by fixing the magnitude of either the lower or upper threshold and minimizing the mean-squared risk over the remaining threshold. This results in a thresholding operator, shown in Eq. 5 where $\tau_1 < \tau_2$, that has lower risk than both the hard and soft thresholding operators shown above.

$$D_i^{\tau} = \begin{cases} 0 & |D_i| \le \tau_1 \\ sgn(D_i) \times \tau_2(|D_i| - \tau_1)/(\tau_2 - \tau_1) & \tau_1 < |D_i| \le \tau_2 \\ D_i & |D_i| > \tau_2 \end{cases} \qquad (6.5)$$

6.3.3 *Estimation of the Noise Level*

We now turn our attention to estimation of the noise level, which involves the value of the threshold itself. The so-called "universal" threshold developed by Donoho and Johnstone[103, 110] assumes an additive noise model as in Eq. 6.1, and generates a shrinkage coefficient via the risk associated with a set of i.i.d. Gaussian white noise components on the standard normal distribution. In combination with a known noise level σ, the universal threshold is given by

$$\tau_u = \sigma\sqrt{2\log n}\,, \qquad (6.6)$$

where n is the length of the signal to be denoised. The universal threshold is quite general, as shown by its application to smFRET data in our prior publication[5], as well as to various other types of data as shown in[97, 136, 140]. In the context of this publication, the universal threshold is used in all thresholding operators. In order to

generate the shrinkage coefficient of the upper threshold used in firm thresholding, we use the method of Gao and Bruce [135] by fixing the lower threshold to the universal value and minimizing the mean-squared risk over the upper threshold. The shrinkage coefficient is multiplied by the noise level estimation to generate the threshold.

In the context of smFRET photon trajectories, we use the average noise level for the entire trajectory as the noise level σ in Eq. 6.6 to generate the universal threshold. For a discretely binned photon trajectory with mean number of photons per time step $\langle N \rangle$, we obtain, via the properties of a stationary Poisson process, noise level $\sigma^2 = \langle N \rangle$. While this approach is perfectly appropriate for use in photon signals for which the mean number of photons, and thus the strength of the signal's shot-noise, is relatively constant, trajectories that show changes in smFRET efficiency are comprised of individual acceptor and donor photon trajectories that show changes as well.

Consider an efficiency trajectory exhibiting a transition, occurring over a single time step, from a high efficiency state to a lower efficiency state. An acceptor intensity change from high to lower intensity and a donor intensity change from low to higher intensity accompany the transition. As a result, at the time step containing the transition, the acceptor shot-noise magnitude decreases and the donor shot noise magnitude increases. The universal noise estimator does not account for the intensity change in each photon trajectory. Depending on the magnitude of the transition in relation to the shot-noise magnitude, the universal estimator may act on the wavelet coefficients containing the transition and introduce error to the denoised estimate. This motivates the use of a time-localized noise level to more appropriately define the noise

level in relation to changes in smFRET efficiency. A time-localized approach to thresholding is a natural extension of wavelet shrinkage in the context of smFRET trajectories in that, instead of using the average noise level throughout the entire trajectory, the time-local noise strength is used instead. A better estimate of the noise level in the context of a nonstationary Poisson process is therefore obtained.

Wavelet shrinkage is no stranger to data-dependent, or data-adaptive, thresholds. For example, Bayesian statistics[131, 137] and Brownian-bridge processes[136] have been used in attempt to define thresholds and remove noise more effectively. To calculate the time-local noise estimate, we consider a discretely binned photon signal $\boldsymbol{N}$, as in Eq. 6.1, collected in discrete time steps of length Δt. Using the Haar wavelet [77], the signal's wavelet coefficients at decomposition at level J contain information from 2^J time steps of the signal at its original resolution. The noise level for one wavelet coefficient D_i at decomposition level J is given by the average noise level spanning that particular time window. For all time windows i (= {0,1,…}) at decomposition level J, we have noise level

$$\sigma_i{}^2 = 2^{-J} \sum_{t=2^J i+1}^{2^J(i+1)} \boldsymbol{N}(t). \qquad (6.7)$$

Taking into consideration that the universal threshold was developed using infinite statistics, and that smFRET signals are of finite and various length, we assume normal statistics in calculation of the time-localized thresholds, and calculate thresholds for each time window i at decomposition level J to be

$$\tau_i = 3\sigma_i \,. \qquad (6.8)$$

Here we have used the heuristic $\tau = 3\sigma$ arising from the assumption of i.i.d. Gaussian white noise components, yielding a probability that a single noise component is larger than its corresponding threshold of < 1 %.

6.3.4 *Translation-Dependent Artifacts*

Lastly, we consider the effects of discontinuities in smFRET photon trajectories. We assume that transitions in the system under observation lead to abrupt changes in smFRET efficiency, and that the abrupt transitions lead to discontinuities in the individual acceptor and donor photon trajectories. Such discontinuities result in artifacts in the denoised estimates. These artifacts are analogous to, although local instead of global, the Gibbs phenomena of Fourier-based denoising, and consist of a rough oscillation around the target values in the neighborhood surrounding a discontinuity. Termed pseudo-Gibbs phenomena[141], these artifacts are mainly a consequence of the discrete wavelet transform being translation-dependent. Translating a signal containing discontinuities in time, denoising it, then removing the translation to obtain a denoised estimate is not always equivalent to the signal that is denoised with no translations. The preceding discussion forms the basis for the translation-invariant discrete wavelet transform (TIDWT)[141].

TIDWT, introduced by Coifman and Donoho and also known as cycle-spinning[141], attempts to reduce the presence of pseudo-Gibbs phenomena in the denoised signal by averaging the estimates produced by a range of translations. The basic premise is to produce a range of circular time shifts, decompose each shift

independently to the wavelet domain, threshold the wavelet coefficients of each shift, reconstruct each shift to the original resolution, remove the time shifts, and average the denoised estimates produced by all of the shifts. The result is a translation-invariant denoised estimate in which pseudo-Gibbs artifacts have largely been smoothed by averaging the shifted estimates. Given the nature of smFRET trajectories and their inherent discontinuities, cycle spinning offers an advantageous approach to obtaining more accurately denoised estimates.

6.4 Results and Discussion

6.4.1 *Wavelet Basis*

Fig. 6.5 compares the denoised estimates produced by different wavelet bases acting on both the two and four state model systems. Each of the Haar, second-order Daubechies (db2), biorthogonal 1.3 (bior1.3), as well as the first order Coiflet (coif1) bases use the universal soft thresholding method described in [110] to produce their denoised estimates. Fig. 6.5A plots the mean absolute percent error (*M*) produced by each wavelet basis acting on the two (solid circles) and four (open circles) state systems, respectively, versus the highest decomposition level used to produce the denoised estimate. The mean absolute percent error (*M*) is calculated as

$$M = \frac{1}{N}\sum_{i=1}^{N}\frac{|E_i - \varepsilon_i|}{\varepsilon_i}, \qquad (6.9)$$

where E_i is the denoised efficiency at time step i, ε_i is the true, simulated efficiency at time step i, the total number of time steps N is taken over all time steps in all

trajectories of the simulated data set, and *M* is expressed as a percentage. Fig. 6.5B shows the percent reduction of *M* with respect to that induced by shot-noise in the simulated two and four state systems, respectively.

It is clear in Fig. 6.5C that the Haar basis significantly outperforms the other bases in the comparison. Specifically, the Haar basis outperforms the other bases by approximately a factor of 2 for the two state system, and results in a 10-20 % improvement relative to the other bases in the case of the four state system. The better performance of the Haar basis can be attributed to its frequency response. This is shown in more detail in Fig. S4, in which the frequency response of the Daubechies series[132] of wavelet and scaling functions, of which the Haar basis is the first and simplest. Considering the uniform distribution of shot-noise in the

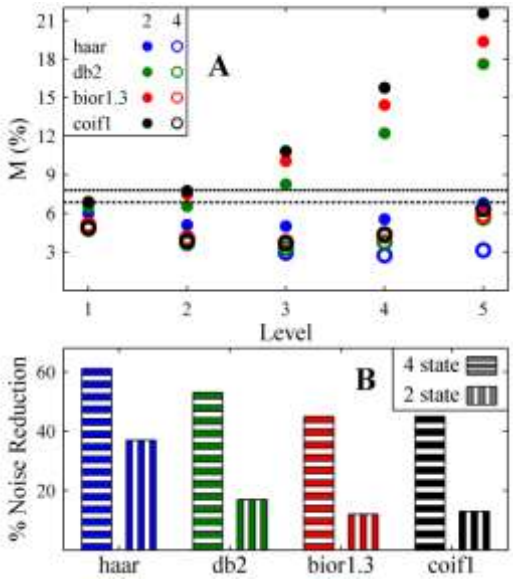


Figure 6.5. Comparison of wavelet bases using the universal soft threshold. A) plots mean absolute percent error (M) of each wavelet basis acting on the 2 state (solid) and 4 state (open) systems vs. highest decomposition level. Shot-noise induced M is shown for the 2 state system (upper dashed) and the 4 state system (lower dashed). B) shows the percent reduction of the most optimally denoised estimate, with respect to M of the noisy data, for the 2 state (vertically hatched), and 4 state (horizontally hatched) systems.

frequency domain, we see that the Haar basis outperforms the other bases because of its distribution in frequency space. Notice that the Haar wavelet in Fig. S4A amplifies a wider range of frequencies in a more uniform manner than the rest of the Daubechies series. Considering that the wavelet coefficients are subject to shrinkage and that the scaling coefficients are not, the main contributor to the Haar wavelet's success in comparison to other wavelet bases, with regards to denoising a stationary or sufficiently slow process in the context of smFRET, is its failure to separate frequency components of the signal as well as its competitors. Based on these properties, it can be concluded that, for most systems studied by smFRET (i.e. those with dynamics slower than the collection frequency), the Haar basis is the best choice of the many wavelet bases for denoising purposes. Given the superior performance of the Haar basis with respect to the removal of shot-noise in smFRET trajectories, all forthcoming denoised estimates were produced with the Haar wavelet basis.

6.4.2 *Comparison of Variations in Wavelet Shrinkage*

Fig. 6.6 compares the results of the application of wavelet shrinkage variations, or variants, consisting of various combinations of shrinkage method, noise estimator, and standard or cycle-spun transformation, to each of the two and four state systems. Each variant uses the Haar wavelet basis, and is named with a three-letter acronym, the first letter being "U" or "A" for universal or adaptive (time-local) threshold, the second letter being "S", "H", "G", or "F" for soft, hard, garrote, or firm thresholding, respectively, and the last letter being "S" or "C" for standard or cycle-spun wavelet transforms. Figs. 6.6A-D show the mean absolute percent errors (M), calculated from

Eq. 6.9, of denoised efficiency signals plotted against the highest level of decomposition for the two state system, and Figs. 6.6E-H show that of the four state system. The dashed lines in Figs. 6.6A-H show the mean absolute percent error induced by simulated shot-noise. Figs. 6.6I and 6.6J show the percent reduction, with respect to the shot-noise induced values, of the mean absolute percent error resulting from the application of each denoising variant to the two and four state systems, respectively.

It is important to note that, as shown by both Figs. 6.5 and 6.6, increasing the highest decomposition level does not necessarily result in improvement to the denoised estimate. In fact, in the case of the two state system, the soft, hard, and garrote thresholding methods cause error to increase with respect to the shot-noise induced value at higher levels of decomposition. This effect is the result of decreasing time resolution in the wavelet transformation. As the decomposition level increases, the time resolution of the wavelet components decreases, and transitions between states may be downsampled out of the wavelet approximation altogether. As a result, transitions may be poorly reconstructed when the wavelet components are upsampled during signal reconstruction. This effect is also observed in the four state system, but its magnitude is smaller due to the proximity of the four states and the construction of a sequential equilibrium between each of the neighboring states.

The reduction of this effect when firm thresholding is used lies in its use of two thresholds. The lower threshold is designed to remove wavelet coefficients that definitely arise from shot-noise. Oppositely, the upper threshold is designed to keep wavelet coefficients that definitely do not arise from shot-noise. Lastly, it is assumed

that wavelet coefficients lying in between the two thresholds are part noise and part non-noise, and these coefficients are reduced as a means to remove the noise part and keep the non-noise part. The overall result is a shrinkage method that is very selective as to what part of each coefficient is noise and what is not, and a denoised estimate that is relatively insensitive to the highest decomposition level. This effect is also reduced in hard thresholding by the use of the time-local threshold that was introduced above. The cause of this reduction is similar to that of firm thresholding – a more appropriate distinction of the barrier between what is noise and what is not noise.

The advantage of time-local noise estimation is made obvious in Figs. 6.6B and 6.6F, where the accuracies of the universally denoised estimates deviate from the accuracies of the adaptively denoised estimates. The deviation arises as a result of the properties of hard thresholding. Hard thresholding is the simplest of all shrinkage methods; it keeps wavelet coefficients with magnitude larger than the threshold, and sets to zero wavelet coefficients lower than the threshold. It is, therefore, of the utmost importance in the use of this method to estimate the noise level with great care. Over or underestimating this value leads to error in the denoised estimate, as shown in Figs. 6.6B ad 6.6F. Time windows that span one or more state changes do not have a stationary noise level, and these changes are under-represented by the universal threshold. This under-representation results in shrinkage of a suboptimal number of wavelet coefficients, and therefore increased error in the denoised estimate. In contrast, the time-local threshold more accurately selects the barrier between noise and non-noise by using time-localized noise strength, thereby resulting in improvement to

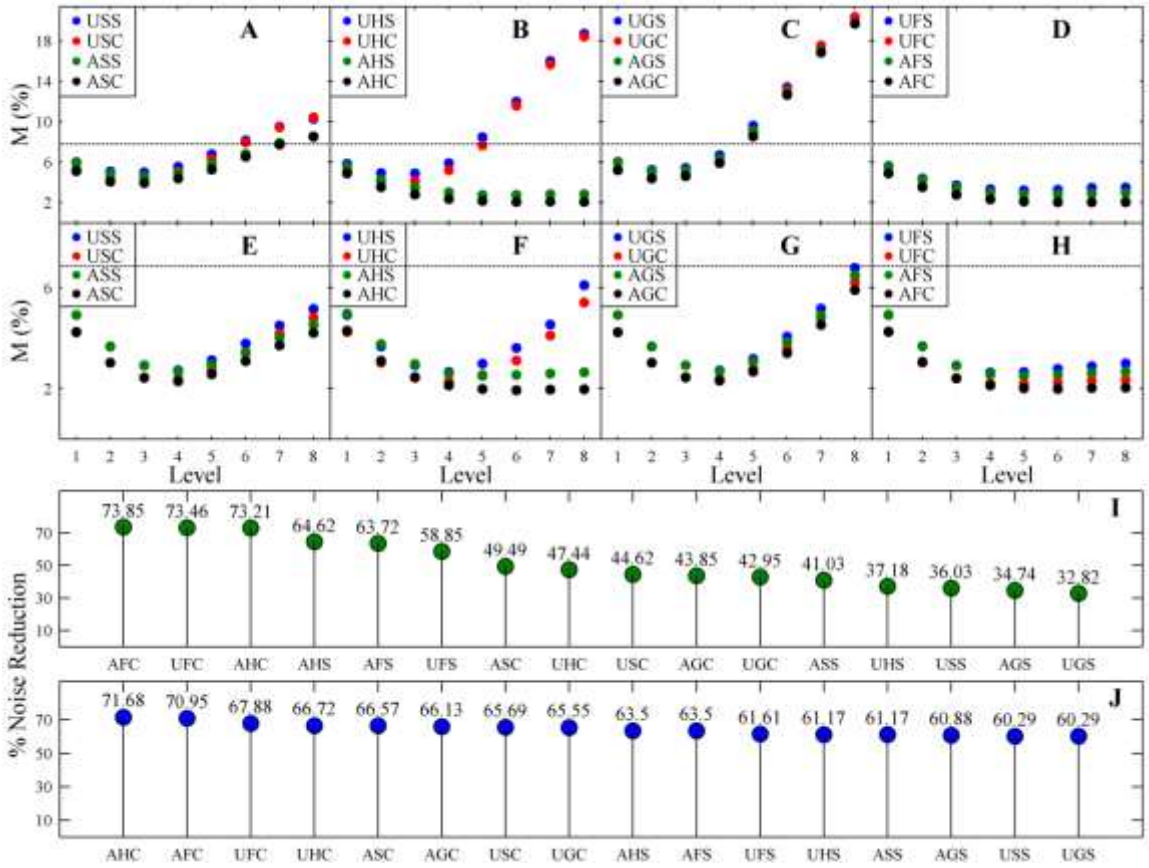


Figure 6.6. Comparison of denoising variants with the Haar basis. A) plots the mean absolute percent error (M), calculated from Eq. 6.9, of soft variants acting on the 2 state system vs. the highest decomposition level used to produce the denoised estimate. B) plots M vs. level for hard variants acting on the 2 state system, C) plots that of garrote variants, and D) shows that of firm variants. E)-H) correspond to the variants shown in A)-D), respectively, acting on the 4 state system. The percent reduction of each variant acting on the 2 state and 4 state systems are shown in I) and J), respectively. Each denoised estimate is labeled on the x-axis with its acronym (see main text), and the percent reductions plotted in descending order.

the denoised estimate. Although the largest improvement is seen with the use of the time-local threshold in combination with hard thresholding, Fig. 6.6 shows that, when the same shrinkage method is used, denoised estimates produced by the time-local threshold are, on average, approximately 5% more accurate than their universally thresholded counterparts.

It is also shown in Fig. 6.6 that denoised estimates produced by cycle-spinning the wavelet transformation have approximately 8.1 % improvement over their counterparts that were transformed in the standard fashion. This is again a result of transitions in the smFRET trajectories and translation dependence of the wavelet transformation. For example, a wavelet coefficient containing a transition that occurs over a single-time step may be downsampled out of the wavelet transformation. When reintroduced during reconstruction, this transition will spread over more than one time step, thereby increasing the error in the denoised estimate. Cycle-spinning reduces the impact of this pitfall by introducing time shifts to the signal and denoising the multiple representations generated by the various time shifts. These time shifts are advantageous in that, if a state-to-state transition is downsampled out of the unshifted wavelet transformation, it is not downsampled out of at least one of the time-shifted transformations. This allows the transition to be present in the wavelet transformation, and thereby reduces the effect of translation dependence.

Despite the differences in our two test systems, Figs. 6.6I-J show that the variants producing the most optimally denoised estimates (i.e., those having the lowest mean absolute percent error as calculated from Eq. 6.9) are consistent across the two

systems. Figs. 6.6I and 6.6J were generated by selecting the decomposition level of the application of each denoising variant with the highest percent improvement relative to the noisy efficiency signal. These data show that the percent improvement (i.e., the ratio of mean absolute percent error of the denoised estimate to that of the noisy data) of the best denoised estimate exceeds 70 % in both the two and four state systems, corresponding to approximately 2% mean absolute error. Moreover, these data show that the same three variants, AFC, AHC, and UFC, produce the most optimally denoised estimates in both cases. These variants are further compared via analyses of state efficiency distributions in the next subsection

6.4.3 *Evaluation of Optimally Denoised Variants*

The three variants, AFC, AHC, and UFC, producing the most optimally denoised estimates were further compared via efficiency distributions and the extraction of the escape rates (i.e., the sum of transition rates from one particular state to all other states). Fig. 6.7 compares the shot-noise induced efficiency distributions of both systems to the denoised distributions produced by these variants. The simulated efficiency distribution of the two state system along with its shot-noise induced percent error *M* is shown in Fig. 6.7A. The denoised efficiency distributions produced by the AFC, AHC, and UFC variants are shown in Figs 6.7B, 6.7C, and 6.7D, respectively. Likewise, the shot-noise induced efficiency distribution for the four state system is shown in Fig 6.7E, and the AFC, AHC, and UFC denoised distributions are shown in Figs. 6.7F-H, respectively.

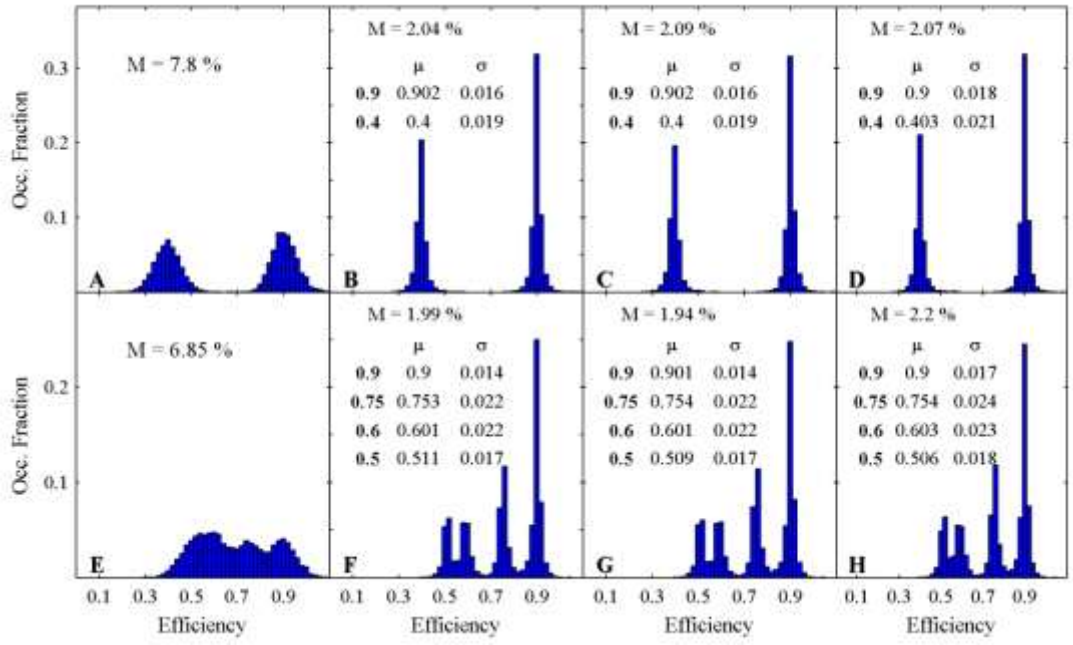


Figure 6.7. smFRET efficiency distributions generated by denoising variants. A) shows the efficiency distribution of the two state model system. B) shows the corresponding denoised estimate produced by the AFC variant, C) shows that produced by the AHC variant, and D) that of the UFC variant. E) contains the efficiency distribution of the four state model system, while F)-H) show the denoised estimates produced by the AFC, AHC, and UFC variants, respectively. Each distribution is accompanied by its mean absolute deviation, *M*, and each denoised estimate is also accompanied by a table inset showing the central efficiency and standard deviation of the data points that contain each simulated state.

There is little difference in the percent error *M* produced by each of the three variants in each of the two model systems. Subtle differences exist in the central efficiencies and standard deviations of each state (table insets), and are results of the differing properties of each of the three methods. Most notably, the AFC and AHC

variants produce the highest resolution between states for both systems, and slightly broader state efficiency distributions are observed in the case of the UFC variant. This result is expected, as discussed above, and arises from better distinction of noise from non-noise by the adaptive threshold. Despite the broader distributions, in the case of the two state system the mean percent error of the UFC variant remains smaller than that produced by the AHC variant. This is simply a result of large shot-noise fluctuations. Regardless of the care that is taken to select the appropriate threshold value, for a finite signal length, there always exists a nonzero probability that a fluctuation produced by shot-noise exceeds the threshold. As such, firm thresholding is more equipped to handle such a situation than is hard thresholding via its shrinkage operation and use of two thresholds. Fluctuations larger than the lower, but smaller than the upper, threshold are reduced in magnitude by the shrinkage operation, thereby reducing the magnitude of the fluctuation in the denoised estimate. This effect seems to disappear in the four state distributions, but its presence is simply masked behind the proximity of the states in the model.

6.4.4 *Comparison of Noise Estimators*

Fig. 6.8 demonstrates the advantages of time-local noise estimation over universal noise estimation. A simulated acceptor photon trajectory (upper panel) shows distinct transitions between photon levels that represent 0.9 and 0.4 efficiency states. The first and second level wavelet coefficients of the high frequency component of the photon trajectory's standard wavelet transformation are shown in the middle and lower panels, respectively. The universal (dashed lines) and time-local (solid lines) thresholds

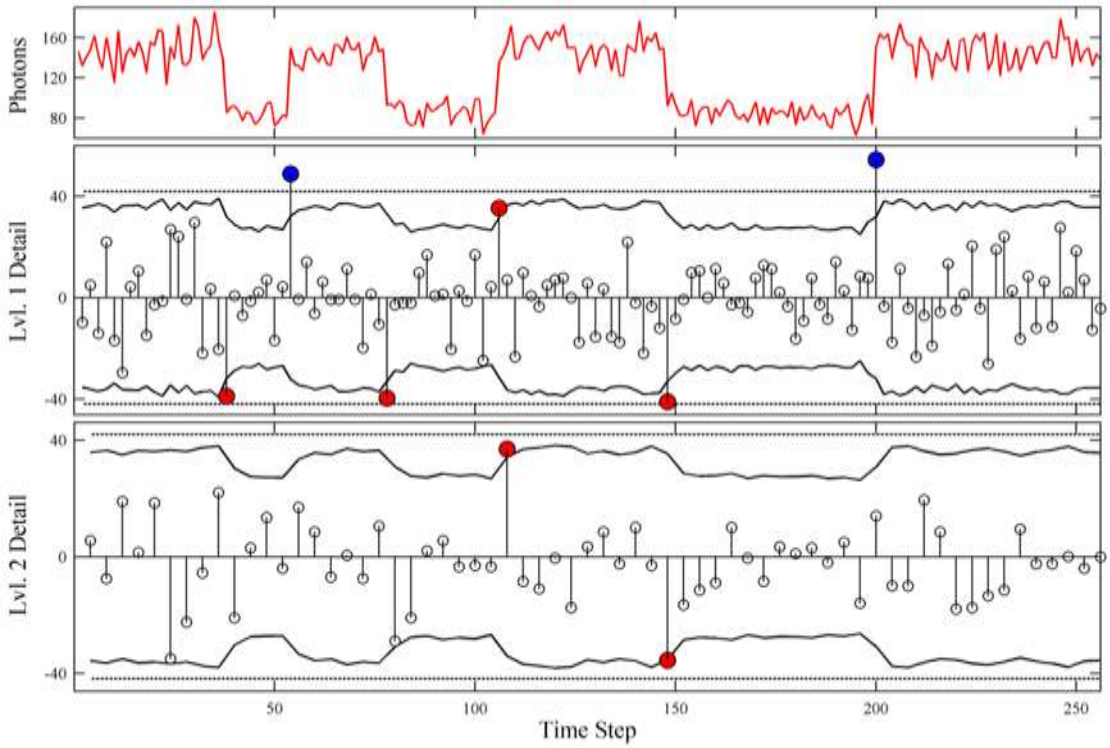


Figure 6.8. A simulated acceptor photon trajectory (upper panel) represents states having efficiencies of 0.9 and 0.4. The trajectory is transformed, resulting in the high frequency wavelet coefficients of the first decomposition level (middle panel) and the second decomposition level (lower panel). The universal (dashed lines) and time-local (solid lines) noise estimators are also shown in each of the lower two panels. Wavelet coefficients resulting from transitions in the photon trajectory that exceed the time-local threshold but do not exceed the universal threshold are emphasized by red markers, and coefficients exceeding both thresholds are emphasized by blue markers.

are also shown in each of the lower two panels. Wavelet coefficients resulting from transitions in the photon trajectory that exceed the time-local threshold but do not exceed the universal threshold are emphasized by red markers, and those exceeding

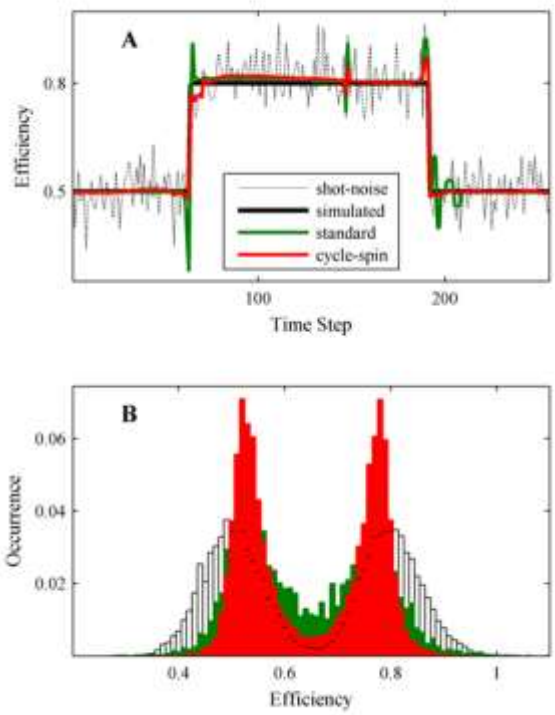


Figure 6.9. Pseudo-Gibbs phenomena and cycle spinning. A) Pseudo-Gibbs phenomena are introduced due to the translation dependence of the standard wavelet transformation (green). Cycle-spinning the wavelet transformation (red) reduces the magnitude of these artifacts. B) The efficiency distribution produced by the noisy data (white) is narrowed by both the standard (green) and cycle-spun (red) transformation, but it is clear that pseudo-Gibbs-induced broadening is more prevalent in the standard denoised estimate.

both thresholds are emphasized by blue markers. Of the 6 transitions in this particular trajectory, only 2 are recognized by the universal threshold. In contrast, time-local adjustment of the noise level estimation results in the recognition of all 6 transitions in the photon trajectory as well as their distinction from shot-noise at both efficiencies. This therefore demonstrates that the time-local noise estimator more ably differentiates between shot-noise and transitions than does the universal estimator.

6.4.5 *Translation-Dependent Artifacts*

The deleterious effects of pseudo-Gibbs phenomena and the subsequent reduction of these effects via cycle-spinning are illustrated in Fig. 6.9. In Fig. 6.9A, a shot-noise induced efficiency trajectory (grey, dotted) is overlaid with the simulated efficiency trajectory (black, and the denoised estimates produced by the standard wavelet transformation (green), and by the translation-invariant[141] wavelet transformation (red). Near the discontinuities in the trajectory, the artifacts are obvious in the denoised estimate produced by the standard transform. In comparison, these artifacts are largely eliminated by the cycle-spun transform, leading to a more accurately denoised estimate. The unwanted effects of pseudo-Gibbs phenomena can be illustrated more clearly with a simulated ensemble smFRET histogram, in which the smFRET distributions of many molecules are combined. Fig. 6.9B displays the original shot-noise affected efficiency distribution of the two-state equilibrium (white) overlaid with the denoised distribution produced without cycle-spinning (green) and the denoised distribution produced with cycle-spinning (blue). While the distribution of each state is narrowed noticeably by the standard wavelet shrinkage method, unwanted macroscale broadening of the distribution by pseudo-Gibbs phenomena are clearly reduced by the translation-invariant approach.

6.4.6 *Population and Kinetic Considerations*

The question remains as to whether wavelet shrinkage improves the ability of common state-finding algorithms to resolve equilibrium or kinetic parameters in a noisy multi-state distribution. In addition to our previous work, Fig. 6.10 addresses this

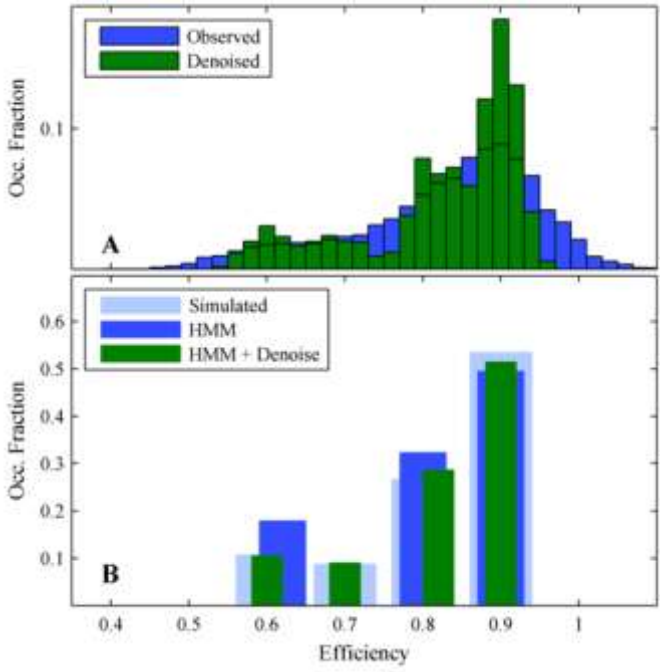


Figure 6.10. State identification in a four state system. A) The shot-noise contaminated efficiency distribution of a simulated 4 state system (blue) is overlaid by the corresponding denoised distribution (greed). B) The simulated fractional population (light blue) is overlaid with the fractional populations extracted by the HMM of [144] from the noisy data (blue) and the denoised data (green).

question. Kinetic Monte Carlo methods were used to simulate a 4 state system with a low occurrence of low-FRET states. Fig. 6.10A shows the simulated experimental distribution of smFRET states from an ensemble of smFRET trajectories (blue) overlain with the denoised distribution (green). Fig. 6.10B shows the underlying states and distributions used in the simulations (light blue). Also included in Fig. 6.10B are the HMM-extracted [144] states (and distributions) acquired from the original data (blue), as well as from the denoised data (green). It is quite obvious in Fig. 6.10B that the state-finding algorithm fails both at identifying states as well as at extracting accurate distributions from the original data. In contrast, not only are all four states found by the HMM in the denoised data, but their populations and efficiencies are

accurate with respect to the simulated values. We therefore conclude that this system is sufficiently complex to provide for the failure of the HMM without the use of wavelet shrinkage, but is accurately resolved with the use of wavelet shrinkage.

For extractable kinetics, we rely on the two and four state systems that are profiled above. Dwell time histograms of each state in the two standard models were constructed and fit to single exponential decays as a means to extract rate constants for the escape from each state. The hidden-Markov model of [144] was used to assign states in each trajectory. The extracted rates, for the four state system, are reported in Table 6.1 along with those extracted from the shot-noise laden data and the true values as extracted from the simulated data. The escape rates for the two state system are not shown due to the fact that the extracted values showed little difference (< 5 %) among the data sets, as well as little difference from the simulated values.

$\langle E \rangle$	Actual	Shot-noise	AFC	AHC	UFC
0.9	0.98	1.21	0.97	1.12	0.94
0.75	1.95	3.16	1.70	1.98	1.66
0.6	1.91	6.49	1.72	1.87	1.58
0.5	1.32	4.13	1.21	1.32	1.03

Table 6.1. Escape rates (in Hz) for the four state system, extracted by dwell time analysis.

Table 6.1 demonstrates that kinetics are not uniformly extracted from the various representations of the more complex model system. The escape rates extracted from the noise-laden data show an average deviation of 132 %, demonstrating that the 4 states in the model along with shot-noise result in a sufficiently complex

system that cannot be accurately distinguished by HMM. The rates extracted from the denoised estimates show large improvement over this value. The extracted escape rates from the three variants, UFC, AFC, and AHC, show average deviations of 14.5 %, 8 %, and 4.4 %, respectively. The improved performance of the AHC variant in comparison to the firmly thresholded variants is a result of their differing behavior near transitions. A transition results in a number of nonzero wavelet coefficients in the region of the transition at different resolution levels. Firm thresholding shrinks the wavelet coefficients with moderate magnitude while hard thresholding leaves them as they are. Shrinkage of these wavelet coefficients results in a smoother transition with the use of firm thresholding, and therefore, slightly less accurate kinetics in the model system. In the overall picture however, the kinetic parameters extracted from each of the three denoised estimates are remarkably accurate in comparison to those extracted from the shot-noise laden data.

6.5 Conclusions

We have presented modifications to the wavelet denoising method presented in our previous publication [5] that enhance ability of wavelet shrinkage to remove the contribution of shot-noise from smFRET photon trajectories by 200 %. We have tested our methods on two separate simulated data sets, a simple two state equilibrium, and a more complex four state sequential equilibrium, and found that the algorithm's ability to remove shot-noise and retain data remains intact. Furthermore, we have demonstrated that, with the inclusion of the presented smFRET-specific modifications, wavelet shrinkage is of general use in smFRET systems ranging from simple, two state

systems, to more complex, multiple state systems in which the resolution, defined by the ability to distinguish distinct states in the efficiency distribution, is marred by shot-noise.

We have also shown, through tests on the simulated data sets, that the use of a more advanced wavelet basis does not result in improvements to the denoised estimates. In fact, if a basis other than the simplest of all wavelet bases, the Haar basis, is used, the percent error obtained from the denoised estimate is far from optimal. This is mainly a result of the frequency space properties of shot-noise and of the digital filters comprising the wavelet basis. Due to the fact that shot-noise is uniformly distributed in frequency space, the Haar wavelet's failure to separate frequencies as well as its competitors becomes an advantage in the removal of shot-noise from a non-homogeneous Poisson process.

We have also introduced a time-localized noise estimator that is custom-designed for smFRET trajectories. Problems with the barrier between noise and non-noise arise with the use of the universal threshold[110] due to its use of the average noise strength over the entirety of a photon trajectory. Given that the observation of changes is most often the goal of smFRET measurements, the noise strength present in a photon trajectory is not stationary, but is ably described by a non-homogeneous Poisson process. The recognition of this limitation gives rise to a time-localized approach to wavelet shrinkage, allowing for more accurate estimation the time-local barrier between noise and non-noise, and therefore improvement to the denoised estimates.

We have also addressed a variety of available shrinkage methods, and have found that the most versatile method in application to smFRET trajectories is the firm thresholding method[135]. Firm thresholding uses two thresholds, which allows for more accurate distinction between noise and non-noise, and therefore gives rise to improved estimates. We have also found that use of the time-localized threshold significantly increases the accuracy of the denoised estimates obtained with the hard thresholding method. This is also due to the more accurate distinction of the barrier between noise and non-noise via the time-localized estimation.

Also, we have addressed the presence of artifacts known as pseudo-Gibbs oscillations that arise in the neighborhood of discontinuities in smFRET trajectories. These artifacts are a direct result of the position of the discontinuity in the signal, and arise due to translation-dependence in the discrete wavelet transformation. We apply a method known as cycle-spinning[141] to reduce the effect of these artifacts, and find that the method results in more accurately denoised estimates.

Lastly, properties concerning resolution of states and state-to-state kinetics of each system were extracted from the most optimally denoised estimates, and were compared to the true, simulated values. Through these comparisons, we find that the combination of the time-local noise estimator, firm shrinkage method, and cycle-spun wavelet transform (AFC) consistently produces optimal resolution between states in the denoised estimates across the two systems. We also find that, although the resolution between states is identical for the two variants acting on the more complex system, the kinetics extracted by this combination are nominally less accurate than the kinetics

extracted by the time-local, hard shrinkage, cycle-spun (AHC) variant due to fast transitions over small efficiency differences between states. The firmly thresholded estimates are smoother than the estimates produced with hard thresholding, leading to slightly decreased accuracy in the extracted kinetics.

Overall, through the application of these methods to our two model systems, we show that, with respect to the process introduced in our earlier work, the modifications to wavelet shrinkage presented here can improve the accuracy of denoised estimates by more than 200 %. We also demonstrate that wavelet shrinkage used with the combination of the Haar wavelet basis, a time-local noise estimator, firm or hard shrinkage method, and a translation-invariant wavelet transformation decreases the shot-noise uncertainty to smFRET efficiency by over 70 %, and allows for accurate characterization of both simple and more complex smFRET systems.

An executable application as well as a MATLAB package for wavelet denoising of smFRET time trajectories can be found at www.lrg.rice.edu .

6.6 Appendix: Frequency Response of a Wavelet Filter $\psi(m)$

A wavelet filter is a finite impulse response (FIR) filter because its output in response to an impulse input decays to zero in finite time. FIR filters are a subgroup of a larger class of digital filters known as linear, time-invariant (LTI) digital filters. Frequency response, a measure of the filter's output spectrum in response to an input signal, is formally defined as the ratio of the output spectrum to the input spectrum. For FIR

digital filters, the frequency response is equivalent to the discrete Fourier transform (DFT) of the filter itself.

As a result of linearity and time-invariance, the output of any LTI digital filter may be computed by the convolution of the input signal with the filter's impulse response. The impulse response, *h(n)*, is calculated by convolution of the filter, *ψ(m)*, with an impulse input

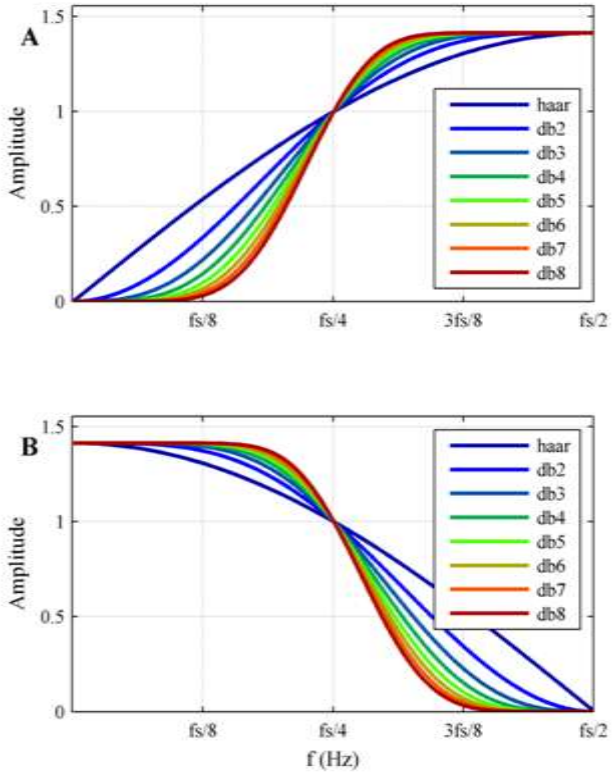


Figure 6.11 The Daubechies series of A) wavelet functions and B) scaling functions in frequency space.

$$h(n) = \sum_{m=0}^{M-1} \psi(m)\, \delta(n-m) = \psi(n) \qquad (6.10).$$

Here *δ(n - m)* is a Kroenecker *δ* that is 1 if *n = m* and 0 otherwise. The transfer function, i.e., the mathematical relationship between the filter's input and output, of any LTI filter is the z-transform of its impulse response. With the Z-transform of output *Y(z)* and input *X(z)*, we have

$$H(z) = \frac{Y(z)}{X(z)} = \sum_{n=0}^{M-1} h(n) z^{-n} \quad (6.11).$$

The frequency response of an LTI digital filter is the transfer function evaluated on the unit circle in the complex plane, i.e., $z = e^{i\omega T}$. Upon inspection, we find that $H(e^{i\omega T})$ is the DFT of the impulse response $h(n)$ in terms of ωT. Furthermore, because $h(n) = \psi(n)$, we find the frequency response of the filter $\psi(m)$ is equivalent to its frequency domain representation

$$H\left(e^{-i\omega T}\right) = \sum_{m=0}^{M-1} \psi(m) e^{-i\omega T n} \quad (6.12).$$

Figs. 6.11A and 6.11B shows the Daubechies series of wavelet and scaling filters, respectively, in frequency space. Notice that the Haar wavelet in Fig. 6.11A amplifies a wider range of frequencies at a more uniform manner than the rest of the Daubechies series. Considering that the wavelet coefficients are subject to shrinkage and that the scaling coefficients are not, the main contributor to the Haar wavelet's success in comparison to other wavelet bases, with regards to denoising a stationary or sufficiently slow process, is its failure to separate frequency components of the signal as well as its competitors.

Chapter 7

Structural landscape of isolated agonist-binding domains of the AMPA receptor studied by single-molecule FRET

The contents of this chapter are adapted from an article originally published in *Nature Chemical Biology* on February 6, 2011.

Landes,C.F.; Rambhadran, A.; Taylor, J.N.; Salatan, F.; Jayaraman, V. "Structural landscape of the isolated ligand binding domain of the AMPA receptor studied by single molecule FRET" *Nat. Chem. Biol.* ***2011.*** 7:168-173.

ABSTRACT

Single molecule fluorescence resonance energy transfer (smFRET) was used to examine the conformational landscapes explored by the agonist binding domain of the α-amino-3-hydroxy-5-methyl-4-isoxazole propionate receptor in its apo and agonist bound forms for wild type and T686 mutant proteins. Although the average conformation for each receptor form was found to be similar to those reported by ensemble measurements, the smFRET data reveal several new points. First, the glutamate-bound form explores a

wide range of conformations. Also, each form was determined to comprise multi-state, sequential equilibria. Finally, rate constants were extracted for the equilibrium conformational interconversions. These results illustrate that the extent of activation is dependent not on a rigid closed cleft, but instead on the probability that a given subunit will occupy a closed cleft conformation, which in turn is not only determined by the lowest energy state but by the range of states that the protein explores.

7.1 Introduction

α-amino-3-hydroxy-5-methyl-4-isoxazole propionate (AMPA) receptors are members of a larger family known as ionotropic glutamate receptors[145]. AMPA receptors are membrane proteins that mediate neurotransmission in the central nervous system [146, 147], and their chemical biology is important to understand considering its role in memory and learning [148]. The structure of these neurological receptors is one containing four subunits in a dimer of dimers arrangement [149, 150]. Each subunit contains two extracellular domains, one being the agonist binding domain (ABD), a transmembrane segment enclosing a central ion channel, and an intracellular domain [150]. When the extracellular agonist binding domain binds the agonist glutamate, a series of conformational changes occurs, ultimately resulting in the formation of a cation permeable transmembrane channel. The ion channel has shown to close (desensitize) even in the continued presence of agonist [151-153]. This allosterically controlled mechanism, in which the agonist-bound protein facilitates both activation and desensitization, continues to be a primary uncertainty in the research of AMPA receptors.

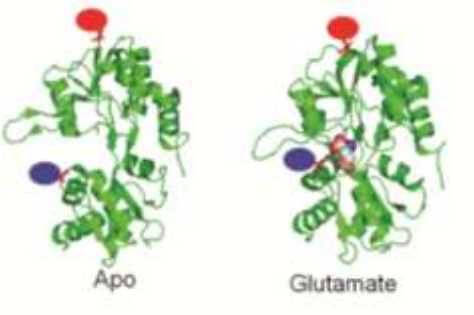


Figure 7.1. Crystal structure of GluR2-ABD showing the sites labeled for the smFRET investigations and distances between the sites in the apo and glutamate bound form.

Differing mechanisms have been suggested for this allosteric process. Initial X-ray structures of the GluR2 subunit (Figure 1) show a bilobed ABD with differing degrees of binding cleft closure when bound to agonists of varying efficacy [151-156]. It has therefore been suggested that the degree to which the cleft is closed is the mechanism by which activation is controlled by the agonist[153]. However, exceptions to this correlation have been recently observed. For example, there is no substantial difference among the degrees of cleft closure in solution based NMR structures of the partial agonists known as the willardiines [157]. The L650T and T686S mutants also show deviations from this correlation; the degree of closure in the binding cleft is larger when the mutants are bound to AMPA and glutamate, respectively, but exhibit only partial agonism [155, 158, 159]. While ensemble studies, such as X-ray structures and ensemble FRET measurements, offer insights into allosteric processes within in the ABD, they are limited by ensemble average and provide information only about the structure of the lowest energy state or the average state, respectively, of the protein. To gain a complete understanding, we must characterize the collection of states that the protein occupies. Here, we observe such collections of states in the GluR2 agonist binding domain (GluR2-ABD), via smFRET, in the apo and glutamate-bound

forms, and in the wild-type and T686S mutant. Using such an approach, we are able to connect the ensemble data with single-molecule data and describe causes of disagreement in the suggested allosteric mechanism.

Molecular dynamics simulation of the GluR2-ABD suggest that the T686S mutants deviate from the hypothesis that cleft closure controls activation by demonstrating that the protein explores conformations that do and do not facilitate the ion channel's activation. These theoretical studies also agree well with electrophysiological measurements of channel activation within the mutant protein [2]. The single-molecule data we present here offer the first experimental proof of the theoretically proposed conformational landscape.

A common issue in smFRET experiments is that the data often yield broad efficiency distributions that are contaminated by experimental noise, thus limiting the interpretation of such distributions to a range of possible conformations [73, 160-162]. There are several sources of such contamination. Firstly, we must determine if there are contributions from experimental parameters such as immobilization protocol or dye-protein interactions that artificially spread the data [163]. Further challenges lie in the characterization of the system via the acquired data. Most often, we must make assumptions regarding the system as a means to extract underlying states and rate constants [20, 164], which introduces bias unless these states are readily obvious. Such an example would be the physically meaningless task of fitting a featureless distribution with multiple normal distributions without knowing whether the observable spreads normally or not. There exist, however, tools that reduce or eliminate the need for

biased assumptions. Specifically, our method to denoise smFRET photon trajectories using wavelet decomposition has proven to be useful in the reduction of electronic shot-noise, making it possible to better resolve underlying states in broad smFRET distributions [5].

In this work, we use this wavelet denoising technique to locate and quantify the states that underlie the noisy, featureless smFRET distributions obtained for single GluR2-ABDs. We use a hidden-Markov model (HMM) to identify four conformational states within the wavelet denoised data of the glutamate-bound ABD, which produces the first experimental evidence for the multi-state equilibria that are predicted by atomistic molecular dynamics simulations [2]. We perform analysis of the waiting time distributions to extract state-to-state transition rate constants, which reveal that these are sequential equilibria. Analysis of the apo form of the ABD indicates an additional resolvable conformation, and that the equilibrium shifts away from the closed-cleft form of the agonist-bound ABD. Lastly, similar analysis of the T686S mutant, in which cleft closure is hindered by the absence of hydrogen bonding between the binding cleft and the glutamate agonist, suggest that the more flexible protein traverses a broader range of states, both with more open and more closed binding clefts.

7.2 Methods

The fluorescence signals of the donor and the acceptor were collected until the fluorophores were photobleached. The apparent FRET efficiency was calculated using the background- and crosstalk-corrected intensities (described in Section 4.2) as in Eq. 2.3. Because the cleft closure is floppy in solution, the molecule is in equilibrium

between multiple conformations. Therefore, it is important to determine the dwell time of the molecule in one conformation. This was done by executing an autocorrelation analysis on the FRET efficiency of each molecule, then taking the average of the autocorrelation values to produce an ensemble autocorrelation (see Eq. 2.4), which is presented and analyzed with logarithmic binning to avoid over-fitting at long lag times. The resultant autocorrelation decays were fit with single exponential functions. Details about the wavelet denoising technique, dwell-time analysis, and rate constant extraction can be found our earlier work [5] and chapters 5 and 6. All data analysis was performed with programs written in house using MATLAB (R2008b, The Mathworks, Natick, MA), with the exception of the hidden Markov state-finding analysis, which was performed with a program made available by the Ha group at the University of Illinois at Urbana-Champaign.

7.2.1 *Multiply Labeled Proteins*

There are two inherent cysteine residues (C425 and C436) in the GluR2-ABD that may be inadvertently dye-labeled, as illustrated in Figure 7.2. These residues are largely inaccessible to dye labeling [158], but occasional multi-dye labeling was observed in our results. We characterize such labeling flaws via multi-step photobleaching, and eliminate these trajectories from the data set. Table 7.1 displays the observed fractions of the occurrence of multiply labeled ABDs. We quantify such characterization via the emission trajectory of a molecule

with one donor and one acceptor. We observe a single-step photobleaching event for each of the donor and acceptor fluorophores as shown in Figure 7.3A. The presence of

two acceptors on the same ABD results in a similar two-step photobleach, as in in Figure 7.3B. Furthermore two donors on the same ABD is characterized by a two-step photobleach in the donor emission trajectory, as shown in Figure 7.3C. Note that this trajectory also contains contributions from two acceptor fluorophores. Lastly, we characterize donor-only labeling via the observation of donor-only emission, as shown in Figure 7.3D.

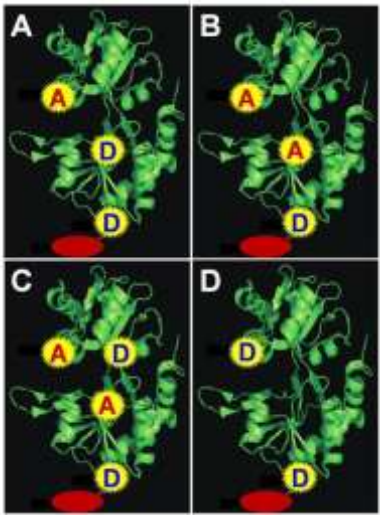


Figure 7.2. GluR2-ABD molecules with multiple labels. (A) Molecule with two donor dyes and one acceptor dye. (B) Molecule with one donor dye and two acceptor dyes. (C) Molecule with two donor dyes and two acceptor dyes. (D) Molecule with two donor dyes.

Table 7.1. Percentages of labeled molecules

Molecule	1D:1A	2D:1A	1D:2A	2D:2A
Apo GluR2-ABD	74%	21%	3%	2%
GluR2-ABD + Glu	78%	17%	4%	1%
GluR2-ABD T686S + Glu	86%	14%	0%	0%

7.2.2 *General Data and Wavelet Analysis*

Each remaining trajectory is then photoblink-filtered with the Bayesian method described in chapter 4 [5]. The resulting trajectories then undergo the statistical tests described in Section 2.2.2, and those not meeting the criteria as defined by the rest of

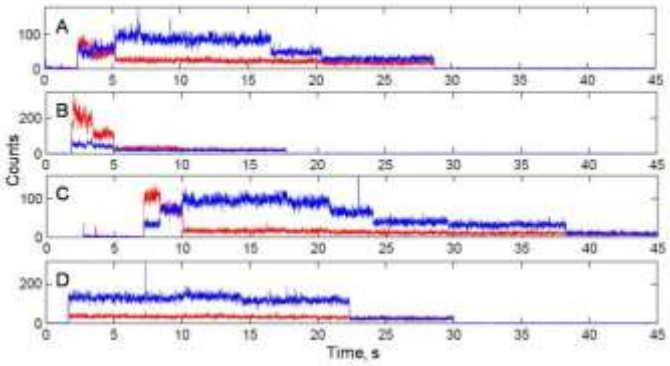
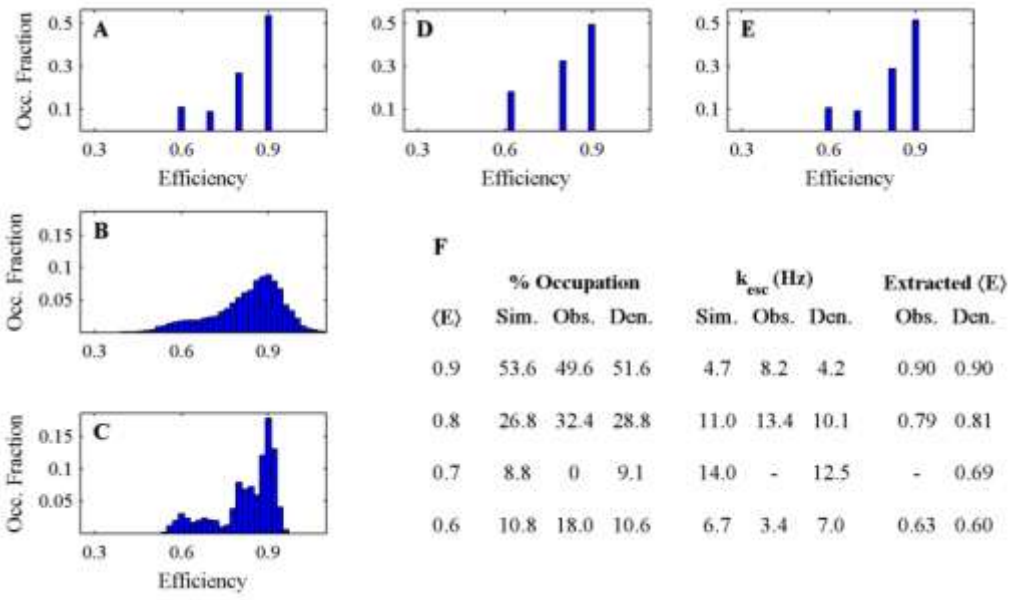


Figure 7.3. Sample emission trajectories of multilabeled GluR2-ABD molecules. The donor emission is blue while the acceptor emission is red. All of these emission trajectories were collected from Glu-bound GluR2-ABD molecules. (A) Molecule with two donor dyes and one acceptor dye. (B) Molecule with one donor dye and two acceptor dyes. (C) Molecule with two donor dyes and two acceptor dyes. (D) Molecule with only donor dye.

the data set are excluded from our analyses. Additionally, donor and acceptor autocorrelation and cross correlation analysis are performed for each trajectory to ensure that the dyes are not rotationally hindered and that the photon count response is anti-correlated for each trajectory, respectively, as should be the case when FRET is occurring. The results of this refinement procedure, in the case of a molecule emitting a stationary number of photons from a stationary conformational state, are trajectories that are up to 97% free of photoblinks and external contamination and broadened only by shot-noise.

After the refinement procedure, the donor and acceptor photon trajectories are subject to the wavelet denoising procedure as described in Chapters 5 and 6 [5], reducing the contribution of shot-noise to the data, thereby increasing both occupational and kinetic resolution. To demonstrate the ability to extract accurate

information from a noisy efficiency distribution obtained from a complicated, multi-state system, we use Kinetic Monte Carlo (KMC) methods to simulate the system shown in Fig. 7.4. The KMC-simulated trajectories consist of 4 states having central efficiencies 0.9, 0.8, 0.7, and 0.6, and each state has an escape rate in the range of 4.5 – 14 Hz, corresponding to average lifetimes in the range of 70 – 200 ms. The relative occupation of each state is shown in Fig. 7.4A. State trajectories are simulated first, with acceptor and donor photon trajectories being constructed from the state trajectories. The sum of acceptor and donor photon counts has a mean of 185 photons per time step for all of the 14,000 time steps in 35 simulated trajectories. The signal to background ratio is 2.3, and the donor to acceptor background ratio is 2. The crosstalk ratio was fixed at 10 %. Shot-noise was simulated at the i^{th} time step as an additive, zero-mean, Gaussian white noise component with magnitude $\sigma_A(i) = \sqrt{N_A(i)}$ for the acceptor signal, and $\sigma_D(i) = \sqrt{N_D(i)}$ for the donor signal. Here, each of the $\sigma_A(i)$ and $\sigma_D(i)$ represent standard deviation, and each is independently and identically distributed for all time steps. After correction for background and crosstalk, efficiencies are calculated and compiled to generate the distribution shown in Fig 7.4B. Each of the 35 acceptor and donor trajectories is then denoised via the procedure described in our published work [5] (Chapters 5, 6), and, after similar background and crosstalk correction, denoised efficiencies are calculated and compiled to generate the denoised efficiency distribution shown in Fig. 7.4C. Hidden-Markov model (HMM) analysis [20] is then used to identify states in each of the noisy and denoised efficiency trajectories. So that the same central efficiencies are identified in each trajectory, efficiency trajectories are concatenated,



F

⟨E⟩	% Occupation Sim.	Obs.	Den.	k_{esc} (Hz) Sim.	Obs.	Den.	Extracted ⟨E⟩ Obs.	Den.
0.9	53.6	49.6	51.6	4.7	8.2	4.2	0.90	0.90
0.8	26.8	32.4	28.8	11.0	13.4	10.1	0.79	0.81
0.7	8.8	0	9.1	14.0	-	12.5	-	0.69
0.6	10.8	18.0	10.6	6.7	3.4	7.0	0.63	0.60

Figure 7.4. Denoising a simulated 4 state system. **A)** shows the relative occupation of each of the 4 simulated states. Adding a zero-mean, Gaussian white noise to each of the acceptor and donor trajectories, subsequent background and crosstalk correction, and calculation of efficiencies leads to the distribution shown in **B)**. The denoised complement to **B)** is shown in **C)**. Applying the HMM to the raw and denoised trajectories results in the distributions shown in **D)**, and **E)**, respectively. Parameters extracted from the hidden-Markov modeled trajectories are reported in **F)**.

and the HMM is instructed to find the most likely states in the trajectories. The relative occupation of the states extracted by the HMM from the noisy data is shown in Fig. 7.4D, and that extracted from the denoised data is shown in Fig. 7.4E. Fig. 7.4F tabulates the information extracted by the HMM analyses. The escape rates shown in Fig. 7.4F are extracted by dwell time analyses, in which the distribution of each state's dwell times is fit to an exponential decay.

The comparison shown in Fig. 7.4 demonstrates that, although the distribution of shot-noise-induced efficiencies of our model system has unremarkable features and indistinguishable states, denoising the trajectories with the procedure described in [5] allows for the accurate extraction of information about the relative occupation of each state, of each state's central efficiency, and of the kinetic aspects of the system. Inspection of the distribution shown in Fig. 7.4D – the relative occupations of each state extracted from the noisy data by the HMM – reveals that the state having central efficiency 0.7 is not identified by the HMM. This is due to the broadening effect

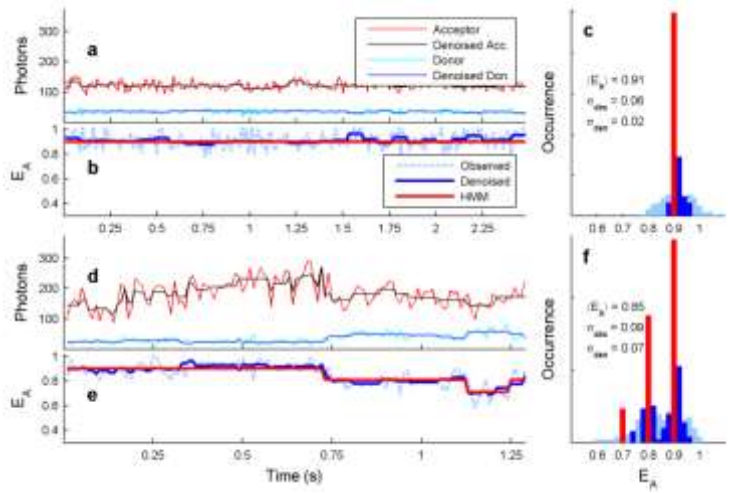


Figure 7.5. Single-molecule trajectory. Upon excitation with 532 nm laser light (A) the energy transfer from the donor dye to the acceptor dye results in high acceptor emission (red) photon count until the acceptor dye is photobleached. The donor dye is more photostable so the donor emission (blue) takes longer to photobleach. (B) The resulting single-molecule FRET trajectory from A is shown in red while the denoised trace is shown in blue. (C) The corresponding denoised FRET histogram shows the average FRET efficiency and the standard deviation of the denoised signal.

of the shot-noise contribution. Furthermore, because of shot-noise broadening and the unidentified state, the kinetics extracted by the HMM from the noisy data are skewed and unreliable at best. In contrast, the central efficiencies extracted from the denoised data have negligible deviation from their true values, relative occupations have average deviation of < 5 %, and the extracted escape rates have an average deviation of only 8.5 %. The precision and accuracy of the system properties extracted after denoising therefore demonstrates the remarkable enhancement to our ability to extract accurate information from complicated efficiency distributions via the application of the denoising procedure.

7.3 Results and Discussion

7.3.1 *Denoising Single smFRET Trajectories*

To characterize the conformational landscape of the GluR2-ABD, we must first increase the resolution in the broad distribution of efficiency values as described in the last section. Figure 7.5A shows a representative smFRET trajectory of a single, donor/acceptor labeled ABD, acquired at 1 ms time resolution and binned to 10 ms for analysis. Each acceptor and donor trajectory was corrected for background and crosstalk [165, 166], passed through the Bayesian photoblink filter, and then denoised via the wavelet procedure described in chapters 5 and 6 [5] to reduce the shot-noise contribution in the photon trajectories. Time-dependent smFRET efficiencies were then calculated from the noisy and denoised trajectories. We show representative results in Fig. 7.5B. These results reveal a sharper smFRET efficiency distribution after denoising because up to 70% of the shot-noise contribution has been removed from the signal.

7.3.2 *Glutamate-Bound GluR2-ABD*

The noisy ensemble distribution compiled from many single-molecule measurements and its denoised counterpart are shown in Fig. 7.6. It is demonstrated in this comparison that the average efficiency does not change after denoising, but the resolution within the distribution is increased by the reduction of the shot-noise contribution. In particular, the average smFRET efficiency of 0.8 corresponds to an average interdye distance (see Eqs. 2.1 and 2.2) of 40.5 Å. Considering the dye placement (see Fig. 7.1), we compare this distance to that determined in solution-based ensemble FRET measurements, and find that the distance we extract corresponds nicely with the ensemble value of 40.8 Å[158]. We expect this result, considering the nature of the two ensemble results are based in FRET, but this result demonstrates that there are no ill effects arising from the immobilization or labeling schemes.

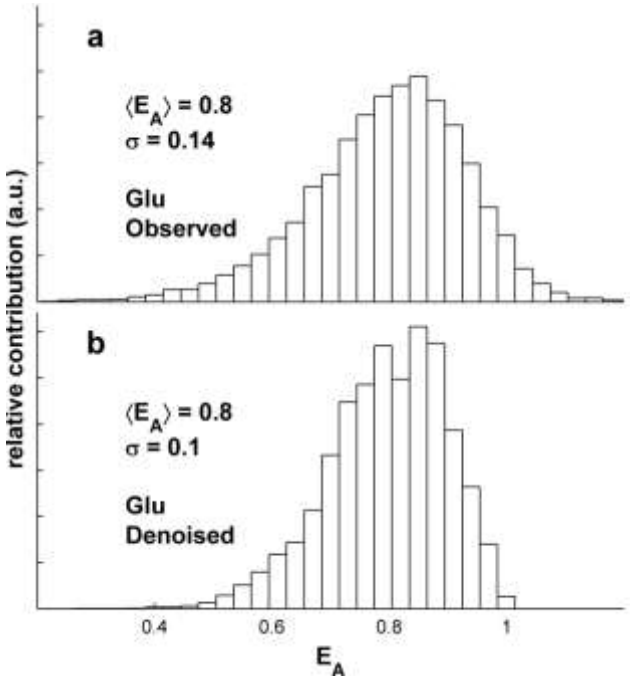


Figure 7.6. (A) Distribution of FRET values from 67 glutamate-bound GluR2-ABD smFRET traces before and (B) after wavelet denoising.

It is important to note the differences between ensemble averages compiled from smFRET experiments to those measured from ensemble experiments. In the

measure of equilibrium distance above, while both types of experiments measure the average over many molecules, only the single-molecule measurement is capable of providing information about the width of a particular population. Furthermore, while X-ray and NMR measurements lead to hypotheses about the source of the spread in the distribution, they are incapable of quantifying the source of the spread. Single-molecule measurements provide information both about the ensemble and about individual molecules, allowing for the quantification of the nature of a broad distribution as well as about the types of transformations that occur within an individual molecular environment.

The results of these experiments provide such information about the GluR2-ABD because additional features are found within the denoised efficiency distribution. The noisy ensemble in Fig. 7.6A has a non-descript standard deviation of 0.14 and no visually resolvable features, thus provides no basis for further occupational or kinetic analyses. But the viability and capability of the wavelet denoising is made evident in the ensemble distribution of denoised efficiencies as shown in Fig. 7.6B. Specifically, we take advantage of the increased accuracy of the denoised efficiencies by performing HMM analysis as described above to extract states from the denoised trajectories. We use the automated algorithm detailed in [20] to find the denoised ensemble trajectories to 2, 3, 4, 5, 6, and 7 conformational states. For each trial, the efficiencies of the extracted states were compared to those of the individual molecules using the smFRET histograms to evaluate the most consistent fit. Results indicate

that four states, centered at efficiencies of 0.59 ± 5.8%, 0.72 ± 4.8% , 0.81 ± 4.2%, and 0.90 ± 3.8% yielded the most accurate depiction of the observed distribution. Fits to more than four states yielded redundant values. This identification of four conformations in the glutamate-bound GluR2-ABD molecules is the first experimental evidence in support of the complex free energy landscape predicted by all atom molecular dynamics simulations [167, 168]. While the distances between the fluorophores are provided by the efficiencies of these states, the source of the distance changes could arise due to side chain rearrangements, changes in backbone orientation due to hydrogen bond changes etc. The precise nature of the conformational change associated with these distances changes cannot be determined from these measurements.

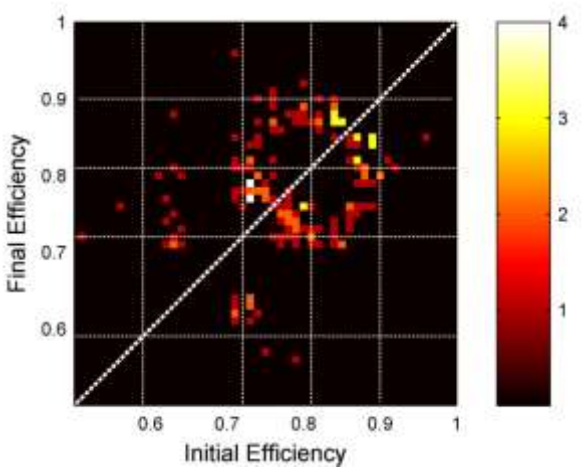


Figure 7.7. Two-dimensional histogram comparing initial and final calculated apparent smFRET efficiencies for all single glutamate-bound ABD protein trajectories. The dashed white line is included to emphasize that, despite the spread in smFRET data values, the transitions overwhelmingly occur between neighboring states.

After the HMM assigns states to each efficiency in the distribution, we analyze the waiting time (or dwell time) distributions to extract the rates of transition as well as

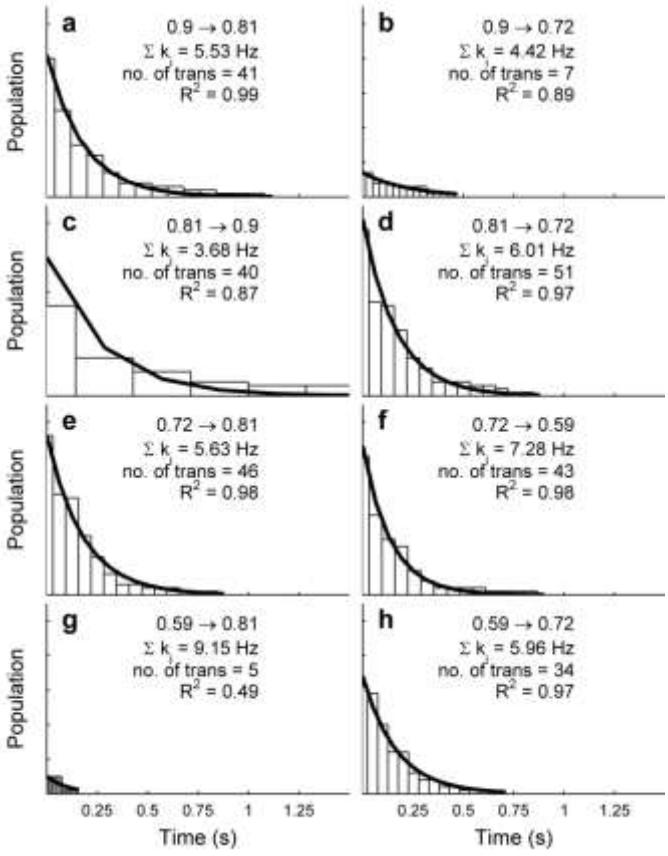


Figure 7.8. Dwell-time histograms for all of the possible transitions between the four states identified from the glutamate-bound GluR2-ABD form. Each histogram was fit to a single exponential decay in order to extract transition rates. 94% of observed transitions occurred between neighboring conformations.

the numbers of transitions between each of the identified states. We extract rate constants by fitting the dwell time distributions of each state-to-state transition to exponential decays. The sequential equilibrium predicted by the theoretical studies of Lau and Roux [2] was inferred from the numbers of transitions between non-adjacent states. Specifically, of the nearly 300 transitions counted within the experimental data, only 6% occurred between non-adjacent states. This is illustrated in the two-dimensional histogram shown in Fig. 7.7. The dwell time distributions for relevant transitions and their respective exponential fits are shown in Fig. 7.8.

We thus draw the general conclusion from our single molecule results in combination with ensemble FRET [158] and NMR [169] analyses on glutamate-bound GluR2-ABD that the agonist-bound form of the GluR2-ABD is not rigidly locked into one active form.

In relation to other studies involving gated ion transport, the smFRET experiments discussed here focus on the structural transformations that occur in the agonist-bound form of single GluR2-ABD. Thus, we compare the rates extracted from our dwell time analyses to those obtained from NMR experiments that studied similar phenomena [170, 171]. These studies found that stabilization of the agonist-bound domain occurs on time scales that are ms or longer, indicating the kinetics describing the agonist-bound conformational landscape occur on slower time scales than the structural changes induced by the agonist binding/dissociation events [172]. Also, channel conductance experiments have measured channel opening and closing dynamics only for full imbedded membrane proteins [173, 174]. The kinetics extracted from the full protein are, naturally, incomparable with our results for isolated ABDs.

Also, we must discuss the extracted rate constants in terms of the experimental time resolution. smFRET trajectories are collected in 1 ms time bins (or 1 kHz frequency) and binned to 10 ms for further data analysis. This implies that the fastest events that can be measured are on the order of 10 ms in lifetime. Thus, in the presence of 10 mM glutamate in the microfluidic chamber, glutamate binding and unbinding events occur 1-2 orders of magnitude faster than our time scale, and consider

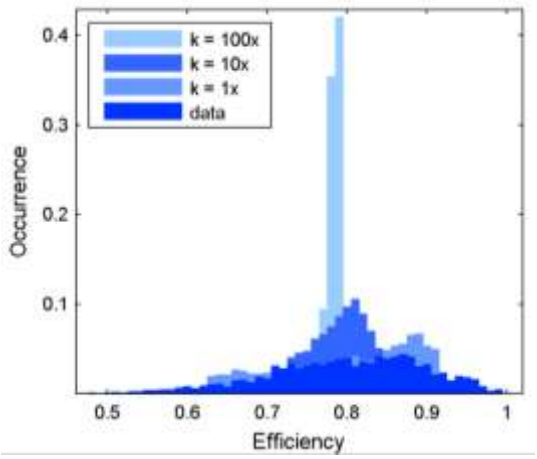


Figure 7.9. Histograms of ensemble smFRET values from experimental GluR2-ABD data (dark blue) are compared to those compiled from simulated data with rate constants on the same order of magnitude, ten times, and 100 times faster than the extracted values, respectively. Only the simulated data with rate constants on the same order of magnitude as the fitted data approximate our observed histogram.

the glutamate/ABD complex to be in its equilibrium bound state for the time window of our measurements.

To confirm, kinetic Monte Carlo simulations were performed for sequential 4-state equilibria with rate constants matching those fit to our experimental data, and with rate constants ten and one thousand times faster, as might be expected if our measurements were monitoring binding/dissociation events. The resulting transitions were incorporated into simulated smFRET histograms and are compared with the actual denoised data in Figure 7.9. This comparison makes is clear that it is not possible we are measuring time-averaged snapshots of faster events. Only the smFRET histogram compiled from transitions with similar rate constants accurately simulates the actual data. It is interesting to note that the only state that is less accurately simulated is the most closed state, which might be associated with a faster docking transition. In this case, the smFRET dwell times would be expected to occur via a convolution between faster

docking events and slower conformational transitions, and the data bear this out. Additionally, we demonstrate, in Fig. 7.10, that two states in an efficiency distribution are only resolvable if their rates of transition are near to or larger than the sampling frequency. If they are faster than the sampling frequency, a bimodal distribution collapses to a unimodal one.

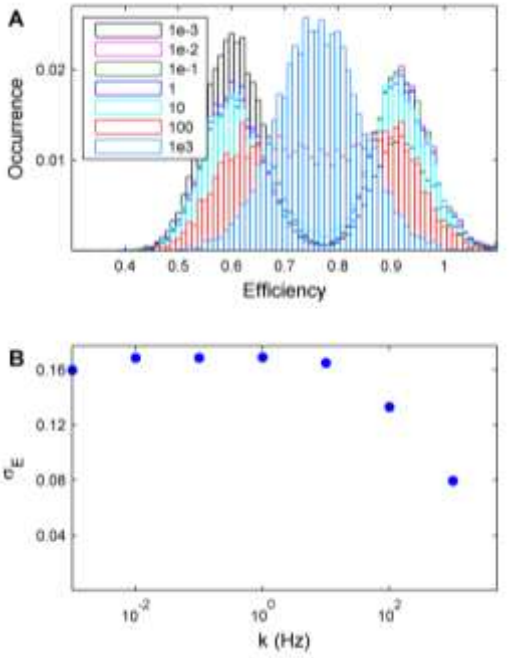


Figure 7.10. Resolution of two states in equilibrium is dependent on the rate of transition between the two states. **A)** The bimodal efficiency distribution of simulated two-state systems collapses to a unimodal distribution as the transition frequency approaches and exceeds the sampling frequency. B**)** The standard deviation of each distribution as a function of transition frequency between the two states in the simulated equilibrium.

7.3.3 *Apo GluR2-ABD*

We apply the same procedure of wavelet denoising, HMM, and dwell time analysis to the apo form of the GluR2 and find three results to be apparent. The first, as expected from other structural studies utilizing crystallography [7,40] and ensemble FRET [158], the binding cleft remains more open overall than the agonist bound ABD as indicated by the decrease in average efficiency in the distribution shown in Fig. 7.11. Furthermore, the width of the distribution is larger than the agonist-bound form, suggesting the apo ABD explores a

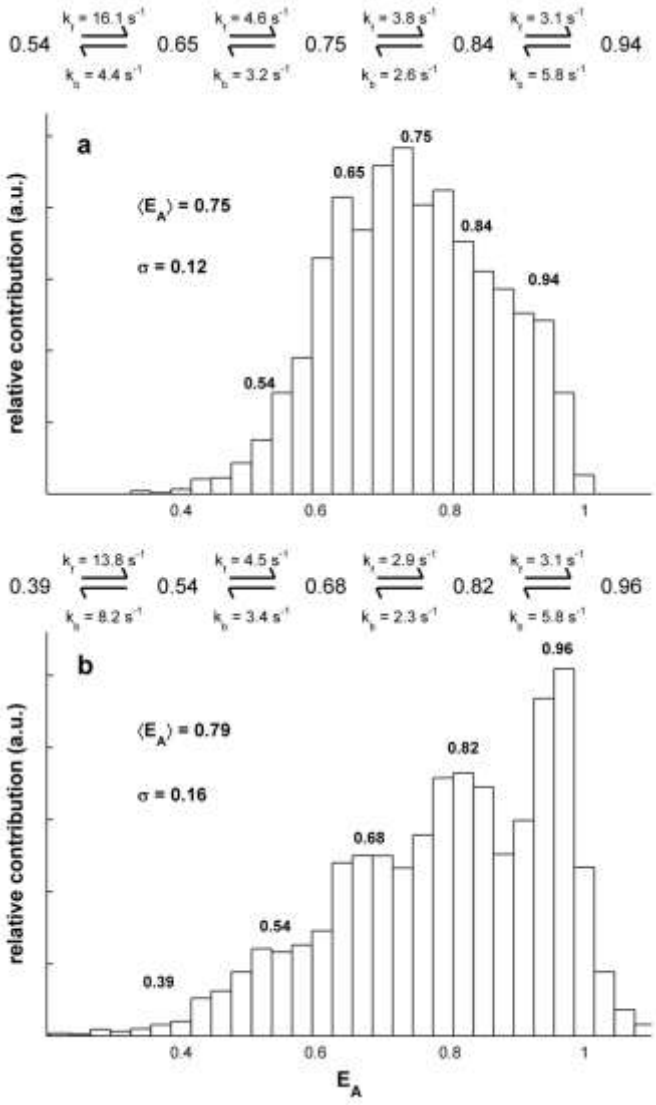


Figure 7.11. A) Denoised apo GluR2-ABD ensemble FRET histogram with the five preferred conformational states and the rate constants for the sequential equilibrium. B) Denoised glutamate bound form of GluR2-T686S-ABD ensemble FRET histogram. The histogram shows the five preferred states and the rate constants of the transitions.

wider range of conformations as predicted by MD simulations [2]. Lastly, the optimal number of states in this distribution, as identified by the HMM, was found to be five states as opposed to the four in the agonist-bound ABD. The efficiencies of the states that were identified are 0.54 ± 7.3%, 0.65 ± 6.1%, 0.75 ± 5.2%, 0.84 ± 4.7%, and 0.94± 4.2%, as shown with the extracted transition rate constants in Figure 7.11.

We propose two possible explanations for this result. In the first, we consider the 1-dimensional free energy plot for the apo form presented by Lau and Roux [2]. This curve contains four inflection points,

indicating the existence of four states, but there are other points along the curve in which the curvature flattens, but does not inflect. These indicate local free energy minima that may be stabilized by slight variation of ionic strength or temperature, creating new inflection points and thus new states. Furthermore, the theoretical simulation uses order parameters that were chosen for, among other qualities, their efficacy in extracting equilibrium properties [2]. The use of different order parameters may result in additional local minima resulting from stabilized conformational intermediates. Additionally, the HMM does not accommodate broadened distributions, as the widths of all states are assumed equivalent [20]. The energy landscape for the apo form is broader than that of glutamate bound protein [2]. This is especially true in the most open regions explored by the protein, so the HMM may approximate one broad state as two states with narrow distributions.

We also performed dwell time analysis on this data. Rates of transition among the five states are illustrated in Fig. 7.11. As in the glutamate-bound case, we find that 95% of the observed transitions occur between adjacent states, again suggesting a sequential equilibrium.

The primary difference, other than the number of states in the distribution, lies in the differences in rate constants for the 0.65 $\leftrightarrow$ 0.75 equilibrium. In the apo form, the protein prefers the more open state located near an efficiency of 0.65, but in the agonist-bound form, all transitions favor the more closed state. This suggests that the most probable state of the apo form has an efficiency that is lower than that of the

glutamate-bound form, which is expected. The distances resulting from these efficiencies correspond nicely with those obtained from x-ray structures [151].

7.3.4 *T686S Mutant*

Our final study involves the T686S mutant form of the glutamate-bound GluR2-ABD. The stabilization of the glutamate-bound form induced by hydrogen bonding is eliminated in this mutant [159]. As shown in Fig. 7.11 by the average smFRET efficiency of 0.79 and the broad distribution, the mutant form indeed occupies a more open and flexible average state. This distribution, as in the case of the apo form, is best fit to a 5 state model. These states are located at 0.39 ± 12%, 0.54 ± 8.7%, 0.68 ± 6.9%, 0.82 ± 5.7%, and 0.96 ± 4.9%. Dwell time analysis again yields rate constants (shown in Fig. 7.11) for the transitions among these states, and these transitions again infer a sequential equilibrium (as predicted in [2]), with 99% of all transitions occurring between adjacent states. While the smFRET data of the mutant form bears closed cleft with much resemblance to that of the agonist-bound wild-type form, consistent with crystal structures [159], the mutant form explores a broader range of conformations than does the wild-type. Considering that the T686S mutant has shown to have an increased rate of recovery from desensitization [159], this suggests that the range of conformations explored by the mutant plays an important role in the desensitization/recovery process. It is then prudent to infer that there is a functional advantage to a more flexible binding cleft, resulting in a trade-off between sensitivity and recovery. Additionally, simulations suggest [2] that the agonist-bound mutant, which is destabilized by the absence of hydrogen bonding, can occupy states with more

closed binding clefts than the agonist-bound wild-type protein. This results in a more flexible structure exploring a wider range of conformations, which is in agreement with the current work.

7.3.5 *GluR2-ABD Conformational Dynamics*

We may quantify this final point concerning overall structural rigidity via autocorrelation of the smFRET trajectories acquired in the experiments to extract the characteristic fluctuation time (τ) of each of the three protein forms studied in this work. We generate an average autocorrelation curve for each situation as described above and fit these curves to exponential decays [175], yielding a fluctuation time for each set of trajectories that represents the time scale of the decay of self- similarity within the sample. These data are shown in Fig 7.12, where we have compared the curves to the stable structure of the HIV1 TAR DNA hairpin [63] to illustrate that a longer decay time and a higher amplitude correspond to a more flexible system. Examining these data, we see immediately that all the GluR2-ABD results are more flexible than the stable structure of the hairpin.

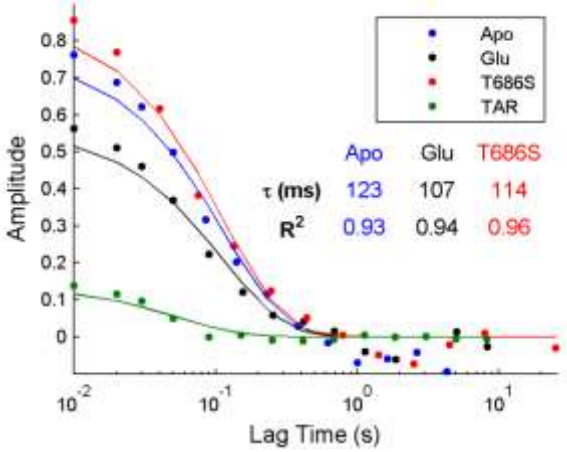


Figure 7.12. The average smFRET efficiency autocorrelation as a function of lag time is compared for the three GluR2-ABD proteins measured in the current work, and compared to a similar analysis of a model rigid biomolecule, TAR DNA [63].

Furthermore, we see that the glutamate-bound form is the most rigid and that the apo and T686S mutants exhibit similar behavior, considering their larger amplitudes and longer decay times that are with error of one another and the calculated values.

7.4 Conclusions

In conclusion, we have shown via smFRET that the glutamate-bound form of the GluR2-ABD is not in a locked conformation, but instead explores a range of conformations in a sequential equilibrium. We accomplish this with the use of an advanced signal processing method, wavelet denoising. We also extract rate constants for interconversions among the identified states. We compare these results to similar results of the apo form and the T686S mutant, and such comparison allows for the observation that these forms are more flexible than their counterpart, thus allowing for the hypothesis that the flexibility of the apo and mutant forms plays a role in the activity and desensitization of the neurological ion channel. Generally, our studies suggest that ion channel activation depends not just on a closed binding cleft, but on the probability that the ABD occupies a closed cleft conformation. This probability is determined not only by the native conformation, but by the range of conformations that are accessible to the GluR2-ABD.

Chapter 8

A Coarse-Grained Model for Molecular Ionic Transport of Rhodamine 6G within a Polystyrene Sulfonate Polymer Brush

ABSTRACT

The experiments of single-molecule spectroscopy eliminate the averaging of ensemble techniques and provide results that often reveal dynamic disorder in the pathways by which processes occur. Theoretical simulations are often used to provide detail concerning the physical system but are often limited by approximations and/or computational facility. Coarse-grained (CG) molecular dynamics simulations are emerging as a useful alternative to atomistic molecular dynamics simulations because they provide a relatively fine level of molecular detail at a reduced computational cost in comparison to atomistic simulations. Among the more common of CG approaches is that of the MARTINI force field. Here we describe a MARTINI-based CG model for Rhodamine 6G diffusing within a polystyrene sulfonate polymer brush. We construct CG models of the molecular ionic fluorophore and the polymer brush, and then perform temperature-dependent molecular dynamics simulations for temporal lengths of 20 ns. We compare our results to empirical data, and find that steric interactions are reliably

reproduced but that electrostatic interaction in the physical system is not adequately reproduced by our simulations.

8.1 Introduction

Single-molecule spectroscopic experiments are providing experimental results with more detail than ever before. These experiments have come in a range of applications, both in the nature of the experimental system, e.g., biological [12, 14, 62], electrochemical [9-11], photophysical [28], as well as in the nature of the spectroscopic method. Single-molecule measurements involving FRET have been applied to a variety of experimental systems [55], performing functions by providing a measure of the distance between two fluorophores that are attached to the molecule(s) under examination in some advantageous fashion. Other methods include FCS, which is often used to measure transport properties of a probe within a variety of experimental systems [7-11].

Attempts to characterize single-molecule experiments often involve theoretical simulations. Various forms of simulation have been used, including Monte Carlo methods [1, 4, 5], Rouse models [176], and atomistic molecular dynamics (MD) simulations [177]. Some of these methods, such as Monte Carlo methods and Rouse models offer the advantage simple implementation and achieving the experimental time resolution, but also offer disadvantages in terms of physical precision and realism. On the other hand, MD simulations offer the advantage of explicit atomistic detail, but computational facilities largely do not allow for the experimental time resolution to be achieved. Considering that experiments such as single-molecule FRET tend to capture

slower, larger scale behavior of the system, the fast, small-scale details provided by atomistic MD simulations are often of little practical value assisting in the interpretation of experimental results. In situations such as these, coarse-grained (CG) molecular dynamics offers an interesting alternative to atomistic MD simulation.

Coarse-grained molecular dynamics simulations have proven useful in describing dynamics of biomolecules [178, 179], polymers [180, 181], as well as some organic dye molecules [182]. CG approaches vary in their scale, e.g. some map CG beads in a 2-4 atom-to-bead ratio, such as those describing polymers and aromatic ring structures [180], and others take a coarser approach, such as the 3SPN approach (3-sites-per-nucleotide) [181], where an entire nucleotide is represented by 3 coarse-grained atoms. One represents the base, the next represents the sugar, and last represents the phosphate backbone.

Among the more common of CG approaches is that of the MARTINI force field [183]. MARTINI was developed initially for use in biomolecular lipid simulations [184], but has since expanded to include CG descriptions of other systems such as proteins [179], polymers [180, 181], and carbohydrates [185]. MARTINI beads typically represent 4 heavy atoms, but some structures, such as phenyl rings [180], have been mapped to 2 heavy atoms per CG bead. Representing multiple atoms as a unified 'bead' has the obvious advantage of reducing the computational cost of integrating over each atom in the molecule and allowing for longer amounts of simulated time. Taking a more fine-grained CG approach also allows for some of the important physical properties of the

molecule, such as mass, size, shape, and charge to be represented with some degree of accuracy. This, of course, comes with the sacrifice of computational cost.

In this chapter we develop a CG model for the system described in Reznik, *et al* [9]. These experiments describe the 3-dimensional orientation of an electrostatically charged organic fluorophore, rhodamine 6G (R6G), as it moves within a polystyrene sulfonate (PSS) polymer brush. We first develop a CG representation molecular ionic fluorophore whose objective is to accurately represent the size, shape, and charge of R6G. Next, we construct an atomistic representation of R6G for use in the characterization and validation of the CG model. In particular, we use the rates of diffusion for the atomistic model and the CG model to calculate an effective time dilation for the CG model, and we use selected atomistic angle distributions to calibrate the angles of the CG model. We then use the CG model developed for polystyrene [180] as a template in the development of a CG representation for polystyrene sulfonate. We characterize the structural properties of the CG model, and then construct a polymer brush representation of the CG model of PSS. We place the polymer brush representation in a simulation volume containing multipolar, CG water [186], Na^+ counterions, and R6G, and perform temperature dependent molecular dynamics simulations for a temporal length of 20 ns. We then characterize the ionic behavior within the system, the structural characteristics of the PSS oligomers, and the dynamical behavior of R6G within the simulated polymer brush construct.

8.2 A Coarse-Grained Model for R6G

In this section we describe two models of R6G. The first is an atomistic model constructed in the GROMOS 43a1 force field that is used for the validation and calibration of a coarse-grained model. The second is a coarse-grained model that was constructed as a means to accurately reproduce the mass, size, shape, and charge of the molecular ionic fluorophore. All simulations are performed in GROMACS version 4.5.5 [39-42].

8.2.1 *The Atomistic R6G model*

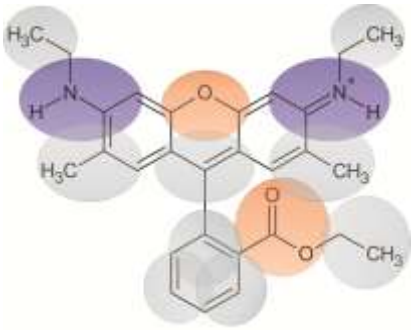


Figure 8.1. The molecular structure of Rhodamine 6G pictured beneath its MARTINI-based coarse-grained mapping. Purple CG beads have a charge of *+0.5q*, orange beads represent polar groups, and gray beads are nonpolar.

The structure of the all-atom R6G molecule is pictured in Figure 8.1. The model is a complete atomistic description of the molecular R6G ion, including explicit hydrogen atoms. The model was constructed in the GROMOS 43a1 force field [187], and the charge on each of the nitrogen atoms was adjusted by *+0.5q* to accommodate the positive Coulombic charge of the molecular ion. The atomistic model was placed in two simulation volumes, each of which is 5x5x5 *nm*. The first simulation volume contained no solvent molecules, and the second contained explicit water as a solvent. Brownian dynamics simulations were performed on the solvent-free simulation volume, with molecular diffusion

characteristics and selected angle distributions being extracted from the data for comparison to the other models. Molecular dynamics simulations were performed on the simulation volume containing explicit water, from which we extract the same properties as in the solvent-free case.

8.2.2 *The R6G-CG Model*

In this section we describe the coarse-grained model of R6G containing 13 CG beads, and term this model to be R6G-CG. The mapping of R6G-CG is shown in Figure 8.1, where blue represents the charged beads containing the nitrogen atoms of the molecule and the orange beads represent the polar groups containing oxygen atoms. We use the CG atoms provided in the MARTINI force field to define the beads in the R6G-CG model, and adjust their masses to appropriately represent the mass of the constituent atoms within each bead. Bonded parameters and constraint distances were defined manually, but nonbonded interactions remain unchanged. This particular CG mapping was constructed to accurately represent the size and shape of the R6G molecule, considering that a more spherical representation may not accurately represent steric interactions within some ordered environment such as the polystyrene sulfonate polymer brush described in [9]. For the CG model, we also perform solvent-free Brownian dynamics simulations and solvated molecular dynamics simulations. Instead of the atomistic water model, we use the multipolar CG water model described in [186].

8.2.3 *Results of the R6G simulations*

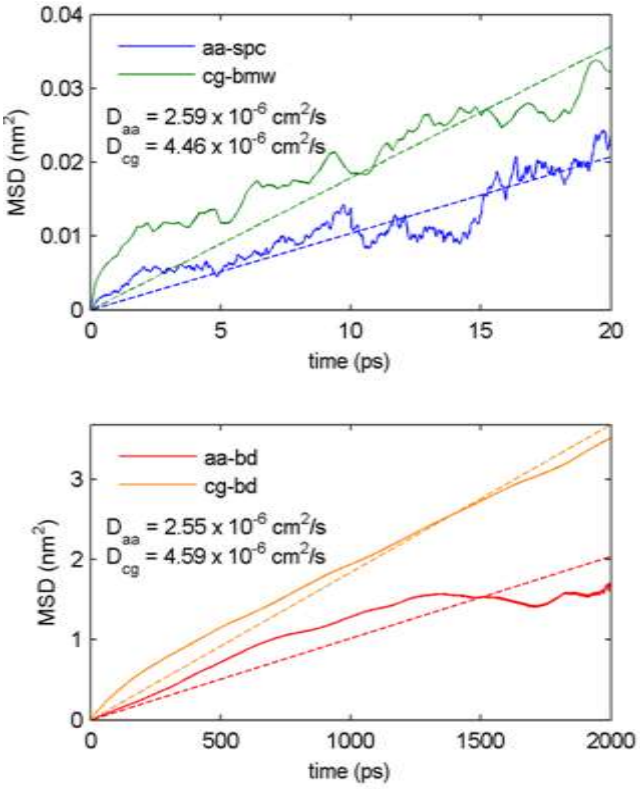


Figure 8.2. Mean-squared displacement curves and fits for the solvated (above) and Brownian dynamics (below) simulations.

Firstly, to measure the time dilation of the CG model that arises from lack of the effective friction of small atomistic motions [180] we measure the molecular diffusion constants of the atomistic and CG simulations. The mean-squared displacements of each simulation are plotted versus time in Figure 8.2 along with their linear fits and extracted diffusion constants. As shown by Figure 8.2, the diffusion constants extracted from the atomistic simulations are comparable to one another, as are those extracted from the CG simulations. Additionally, we compare the atomistic diffusion constant of approximately 2.6 x10^{-6} cm^2/s to the experimental one of 2.84 x10^{-6} cm^2/s [10], and find good agreement. Furthermore, we see that the CG-based diffusion constants are larger than their atomistic counterparts, indicating the time dilation occurs as expected. Specifically, the ratio the CG diffusion constant to the atomistic diffusion constant indicates a time dilation on the order of 1.7.

Next, we characterize the structural similarity of the CG model with respect to the atomistic model. To accomplish this, we select 4 structural angles within the atomistic model (Figure 8.3) and compare their distributions to the corresponding angles within the atomistic model. The CG angle distributions extracted from the solvated simulations (Figure 8.4, upper row) show good agreement overall with their atomistic counterparts, indicating that the R6G-CG model retains good structural similarity with respect to the structure of the atomistic R6G. The angle distributions extracted from the Brownian dynamics simulations (Figure 8.4, lower row) also show good agreement with their atomistic counterparts; however these distributions display less similarity than the solvated configuration. This is expected however, and is also a product of the reduced effective friction in the CG model as discussed above. In particular, this friction that is reduced in the CG model is effectively increased in the atomistic model, leading to sharper angle distributions. In combination with the temporal results described above, these structural results indicate

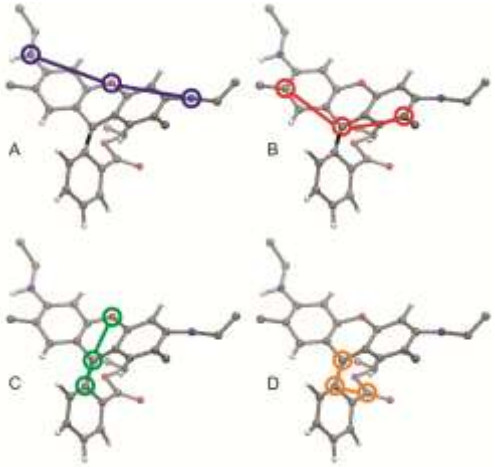


Figure 8.3. Structural angles selected for CG calibration. The angles α (A) and β (B) span the upper ring structure. The angle γ (C) spans the lower ring attachment, and the angle δ spans the lower ring's attachment to the ester tail.

that R6G-CG is indeed a representative coarse-grained model of the atomistic R6G model.

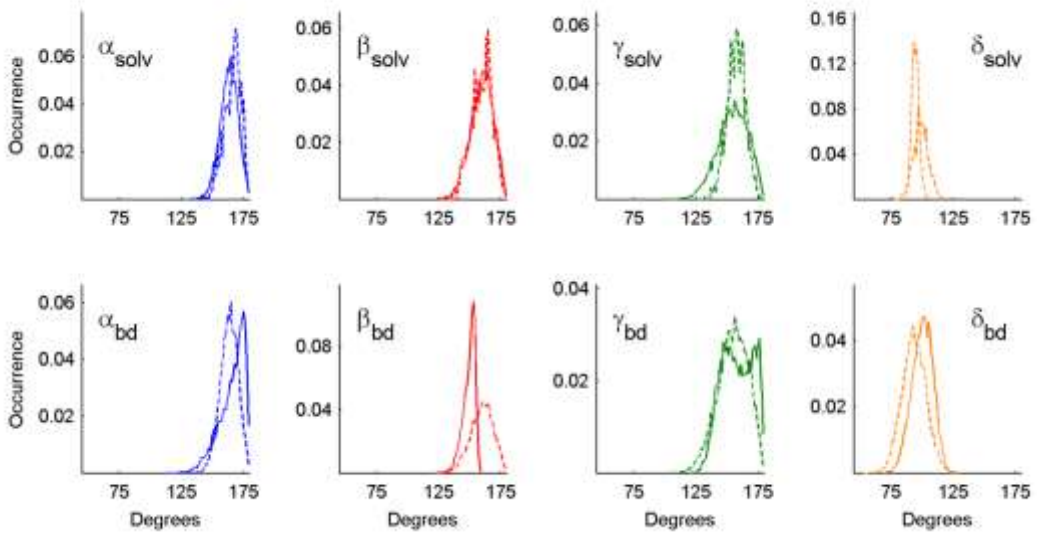


Figure 8.4. The angle distributions of the solvated simulations are shown (above) along with those extracted from the Brownian dynamics simulations (below). Solid lines are the atomistic distributions and dashed lines are the CG distributions.

8.3 A Coarse-Grained Model for a Polystyrene Sulfonate Polymer Brush

8.3.1 Description of the PSS CG Model

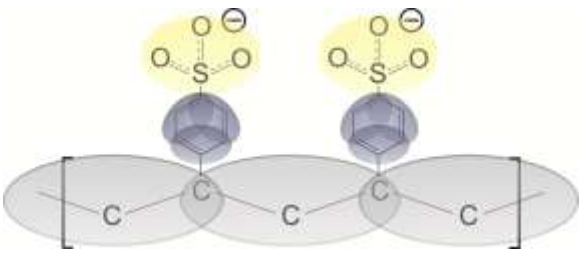


Figure 8.5. Coarse-grained mapping of polystyrene sulfonate.

The coarse-grained model for PSS was adapted from the MARTINI-based CG model for polystyrene [180]. Specifically, the A-mapping of [180] was modified with an additional CG bead to describe the sulfonate group attached to

the phenyl ring. This mapping then is characterized by 5 CG beads per monomer, 1 representing the backbone and 3 representing the phenyl ring, as in [180], and 1 representing the sulfonate group. An additional CG counterion, the MARTINI representation of Na^+, accompanies each monomer. The LJ parameters of the sulfonate bead are defined via the bond lengths described for the sulfonate group in the crystal structures of p-toluene sulfonic acid [188]. Considering that the phenyl-C bound to the sulfonate group is shared between two beads in the CG representation, bonds were constructed between two of the phenyl CG beads and the sulfonate bead. Bond lengths and angles were also inferred from the crystal structures in [188]. The mapping of a 2 monomer segment is illustrated in Figure 8.5.

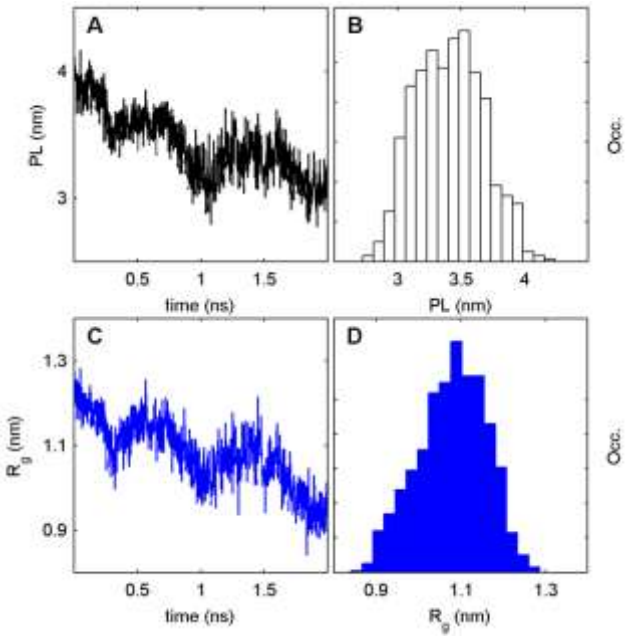


Figure 8.6. Structural characterization of a single PSS20 oligomer. A) Time trajectory for the average radius of gyration for all backbone beads at each time step. B) The distribution associated with the data in A). C) Time trajectory for the persistence length of PSS20. D) The distribution associated with the data shown in B)

8.3.2 *Simulation of a Single PSS Oligomer*

A single PSS oligomer comprised of 20 monomers and their counterions was placed in a 5x5x10 *nm* simulation volume, and molecular dynamics simulations were performed the after the oligomer had been solvated in the multipolar CG water model described in [186] for a temporal length of 2 *ns*. To simulate the effect of the oligomer being bound to a surface as would be the case in a polymer brush construct, we place a 3-dimensional position restraint on the bottom-most backbone bead so that any movement of the restrained bead is counteracted by a restoring force. From each of these simulations, we extract end-to-end distances (persistence lengths, PLs) of the oligomer as well as radii of gyration associated with the polymer backbone. These data are illustrated in Figure 8.6. Considering that the oligomer's initial configuration was an extended chain, we observe a collapsing of the oligomer that appears to occur in discrete steps. This is also indicated by the multimodal nature of the distribution. Furthermore, each discrete step in the persistence length trajectory is accompanied by a corresponding step in the radius of gyration trajectory, indicating that as the PSS chain becomes more and more collapsed, the movement of the polymer backbone becomes more and more constrained.

8.4 Simulation of R6G Diffusing within a PSS Polymer Brush Construct

8.4.1 *Description of the R6G-CG/PSS Simulation Volume*

In order to simulate the experimental environment of R6G diffusing within a PSS polymer brush as described in [9], we construct simulation volumes containing R6G-CG and a polymer brush representation of PSS. Specifically, the oligomer constructed in section 8.3 was replicated to a 2-dimensional grid to simulate the polymer brush

construct on a small spatial scale. We align the oligomers in a 5x5 array such that the system consists of 25 oligomers each having 20 monomers, and we place a position restraint on the bottom-most bead in each oligomer to simulate being bound to a surface. The distance between neighboring oligomers is 2.5 *nm.* The total simulation volume is 12.5x12.5x10 *nm*, thereby allowing for a volume consisting solely of water molecules to surround the polymer brush construct. We then solvate the polymer brush construct with the BMW model [186], add a single R6G-CG to the simulation volume, and perform molecular dynamics simulations with a time step of 20 fs for a total temporal length of 20 ns. Considering that the masses of the MARTINI-based CG beads were adapted to accurately represent the masses of the groups of atoms from which they are comprised, we conduct a temperature dependent study to characterize the behavior of the CG model at temperatures of 300 K, 350 K, 400 K, and 450 K. From these simulations we extract radii of gyration for each backbone bead in each oligomer, persistence lengths for a selected oligomer, diffusion constants for R6G-CG and Na^+ ions, a radial distribution function for Na^+/SO_3^- distances, and a rotational correlation function for R6G-CG.

8.4.2 *Ionic Behavior within the R6G-CG/PSS System*

To ensure that the ionic behavior within our system is appropriate, i.e., that Na^+ ion mobility plays an appropriate role in the dynamics of the polymer brush construct, we extract temperature-dependent diffusion constants for Na^+ from the simulated trajectories. These data are shown in Figure 8.7A. Specifically, the diffusion constants (in units of cm^2/s) are 2.3 x 10^{-6} (+/- 2.4 x 10^{-7}), 6.8 X 10^{-6} (+/- 1.37 x 10^{-7}), 1.5 x 10^{-5} (+/-

8.8 X 10^{-7}), and 2.4 X 10^{-5} (+/- 4.4 X 10^{-7}), at temperatures of 300 K, 350 K, 400 K, and 450 K, respectively. From these data we can clearly see that the Na^+ diffusion constant increases with temperature as expected. Furthermore, when we examine the radial distribution of Na^+ ions with respect to the locations of the SO_3^- groups of each monomer that are reported in Figure 8.7B, we see that the distributions indicate the presence of at least 2 preferred ionic arrangements, which is indicative of complicated charge-shielding within the polymer brush construct. Moreover, we observe a decrease in the relative occupation of these arrangements as temperature increases that is also indicative of increased Na+ ion mobility at higher temperature. In the comparison of these distributions to those of the atomistic PSS simulations reported in [189], we find excellent agreement between the atomistic and coarse-grained models.

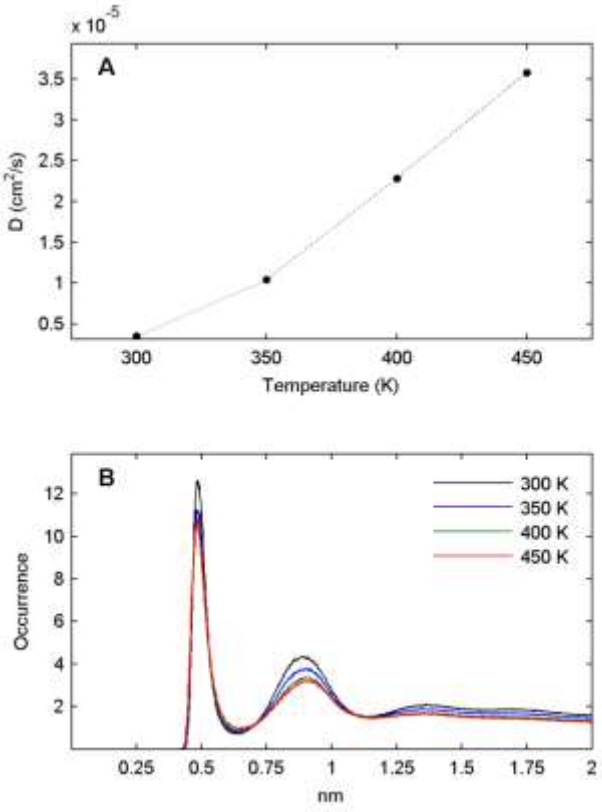


Figure 8.7. A) Collective diffusion constants (cm^2/s) extracted for Na+ ions. B) Radial distribution functions for Na^+/SO_3^- distances.

8.4.3 *Structural Characteristics of the PSS oligomers*

In this section we discuss the structural characteristics of the simulated PSS oligomers as a function of temperature. In particular, we extract radii of gyration for the backbone beads in each oligomer to measure the effects of temperature on the conformational dynamics of the polymer backbone. Next, we select a single oligomer that resides in a central location within the polymer brush matrix and observe its persistence length, defined as the radial distance from the position-restrained bottom-most backbone bead to the top backbone bead in the single PSS oligomer.

Time trajectories of the average radius of gyration (R_g) for all backbone beads at each time step in the simulation are shown in the left-hand column of Figure 8.8, and the associated distributions of these average values are shown in the right-hand column of Figure 8.8. The most obvious feature in the time trajectories is that, as expected, the average radius of gyration and its associated distribution width, does indeed increase with

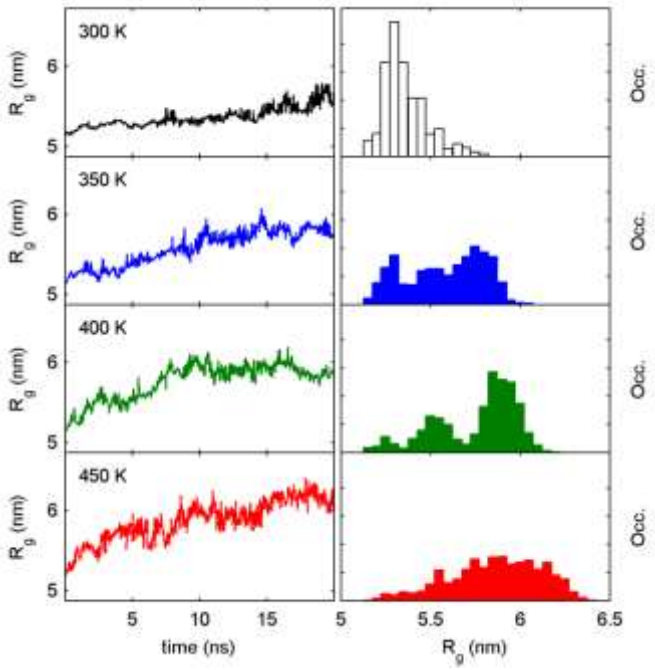


Figure 8.8. Time trajectories for radius of gyration of polymer backbone beads (left) are shown with their corresponding distributions (right). Temperature increases from 300 K, to 350 K, to 400 K, to 450 K (downwards).

temperature. Considering the multi-modal nature of the distributions shown in Figure 8.8 may lead one to conclude that interactions among the oligomers causes the motion of the backbone beads to be constrained in some systematic fashion. However, we observe a positive slope in all time trajectories, indicating the process of equilibration is not yet complete, even after 1 million time steps and 20 ns of simulation time. This effect is most noticeable at the highest temperature of 450 K but appears at all temperatures. The zero slope in the latter 10 ns of the 400 K trajectory indicates that this system is equilibrated, but the positive slope remains at lower temperatures and does not allow for this conclusion to be drawn from these data alone.

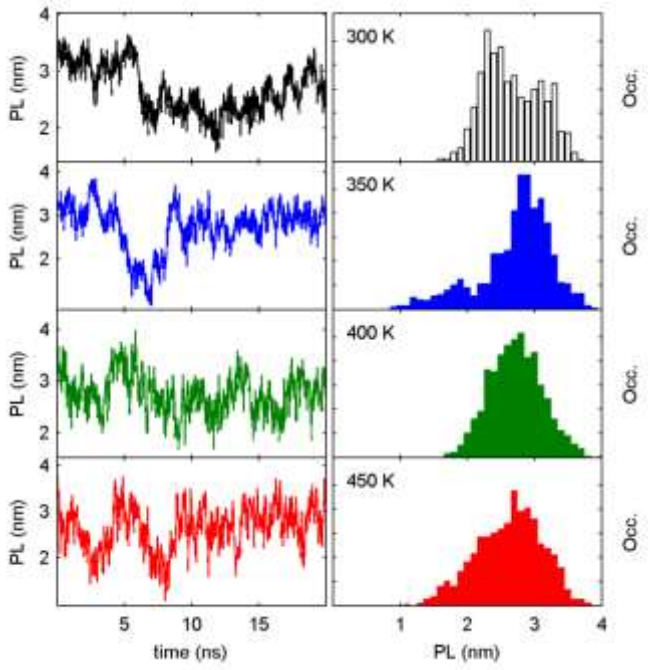


Figure 8.9. Time trajectories for persistence length between the polymer's bottom-most and topmost backbone beads (left) are shown with their corresponding distributions (right). Temperature increases from 300 K, to 350 K, to 400 K, to 450 K (downwards).

Similarly, the persistence lengths (PLs) associated with a selected oligomer are reported in Figure 8.9. The same oligomer was selected for each study reported in Figure 8.9. The left-hand column contains time trajectories observed for this particular oligomer and the right-

hand column contains the distribution associated with the time trajectory. From these data, we immediately observe that the persistence length of the PSS oligomer exhibits more rapid fluctuation at higher temperatures in agreement with the data reported for the radius of gyration. Upon examination of the distributions however, we observe that this faster fluctuation does not lead to the occupation of a broader range of PLs. This leads us to draw the conclusion that the PSS oligomers do not sample from a broader configurational space at higher temperature, but rather sample this configurational space more rapidly. Furthermore, examination of the time trajectories reveals the reoccurrence of particular values for the PLs, leading to the conclusion that there are indeed preferred conformations of the oligomer, i.e., there are local minima within the configurational sampling space.

8.4.4 *Dynamical Characteristics of R6G-CG within the PSS Polymer Brush Construct*

To quantify the dynamical behavior of R6G-CG within the PSS brush, we extract diffusion constants and rotational decay times from the simulated trajectories. These data are shown in Figure 8.10. Specifically, Figure 8.10A displays the first 5 ns of the mean-squared displacement (MSD) of R6G-CG at each temperature, and Figure 8.10B shows the associated rotational correlation functions. Figures 8.10C and 8.10D display the extracted values as a function of temperature.

From the data shown in Figure 8.10A, we see that the MSD of R6G-CG remains reasonably linear for the first 5 ns of each simulation, and we extract diffusion constants from these data. We thereby conclude from Figure 8.10C that the diffusion constant does indeed increase with temperature as expected. Specifically, the extracted diffusion

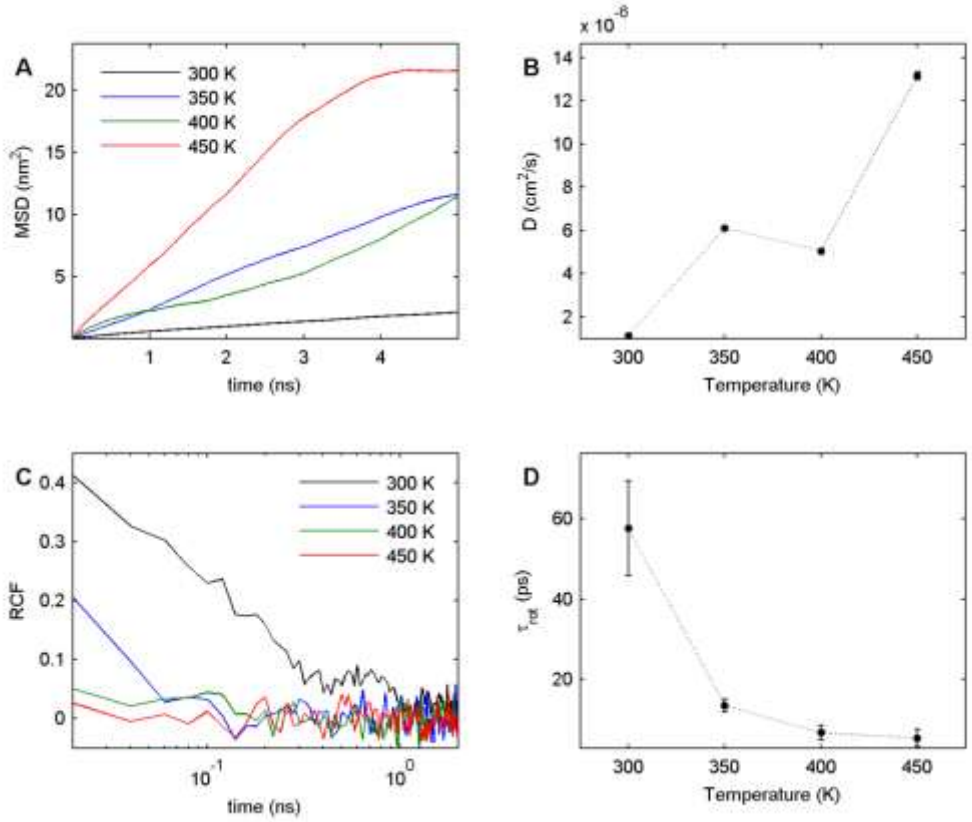


Figure 8.10. Dynamical analysis of R6G-CG within the polymer brush construct. A) Mean-squared displacement trajectories of R6G-CG. B) Rotational correlation functions for R6G-CG. C) Diffusion constants for R6G-CG as a function of temperature. D) Rotational decay times as a function of temperature for R6G-CG.

constants are (in cm^2/s) 3.0 x 10^{-7} (+/- 2.0 x 10^{-7}), 6.0 x 10^{-7} (+/- 2.8 x 10^{-7}), 5.4 x 10^{-6} (+/- 4.2 x 10^{-7}), and 2.5 x 10^{-5} (+/- 2.5 x 10^{-6}) at 300 K, 350 K, 400 K, and 450 K, respectively.

In comparison to the R6G-CG in BMW simulations performed above (at 298K), we observe a decrease in the rate of translational diffusion by an order of magnitude, suggesting that interaction of R6G-CG with the polymer brush matrix does indeed occur, effectively raising the local viscosity and hindering translational diffusion. We also note

here that, although the diffusion constant extracted at 400 K appears at first to be anomalous, examination of the spatial trajectory reveals increased residence time within the polymer brush matrix as compared to the other temperatures. This increased residence time is the result of polymer brush dynamics in the local environment surrounding R6G-CG. In short, polymer dynamics lead to the formation of a 'pocket' surrounding the polymer, thus hindering translational diffusion. This hindrance appears to be largely steric in nature, as will be revealed by the upcoming rotational studies.

The rotational correlation functions shown in Figure 8.10B indicate that temperature is the main factor influencing the rotational diffusion of the molecular ionic fluorophore. This is indicated by the observation of substantial correlation amplitude at the lowest temperature of 300K. Each rotational correlation function was fit to a single exponential decay to extract the rotational decay times that are shown in Figure 8.10D. These values are (in *ps*) 57.5 +/- 11.7, 13.6 +/- 1.6, 6.8 +/- 1.9, and 5.5 +/- 2.1 at 300 K, 350 K, 400 K, and 450 K, respectively. Firstly, these data reveal that the rotational decay times decrease with temperature as expected. Furthermore, considering that the trajectories used to construct the rotational correlation functions are sampled at a time resolution of 20 ps, the only temperature at which R6G-CG exhibits significant rotational correlation is 300 K. Upon further examination, one recognizes that the hindered translational diffusion that was observed at 400 K does not result in a corresponding hindrance to rotational diffusion. This result suggests, as mentioned above, that increased residence of R6G-CG within the polymer brush is not induced by a preferred orientation, but rather is largely governed by steric interactions at this temperature.

Lastly, we observe rapid rotational diffusion on the order of picoseconds at all temperatures, leading to the conclusion that, despite the electrostatic charge of +1q that resides on the molecular ion, there is not a long-lasting electrostatic interaction between R6G-CG and the polymer brush. Considering the combination of these results we must conclude that the interaction that is observed via a reduction in the rate of translational diffusion is primarily steric in nature.

8.5 Conclusions

In conclusion, we have constructed atomistic and coarse-grained representations of the molecular ionic fluorophore, Rhodamine 6G. The coarse-grained representation is based in the MARTINI coarse-grained force field. Simulations of the atomistic representation were performed via water-solvated molecular dynamics as well as via solvent-free Brownian dynamics, and the extracted diffusion constants agree well with empirical data (Carmen ref). Structural angle distributions were extracted from the atomistic data and were used to calibrate the coarse-grained representation. We use the multipolar, coarse-grained water model described in [186] to solvate R6G-CG, and performed both molecular dynamics and Brownian dynamics simulations as in the atomistic case. We use the extracted diffusion constants for the atomistic and coarse-grained molecular models to calculate an effective time dilation of 1.7 for the coarse-grained model.

We also constructed a coarse-grained model for polystyrene sulfonate, also MARTINI-based, that used the similarly MARTINI-based polystyrene construct [180] as a basic template. We tested a single, BMW-solvated polymer, comprised of 20 monomers,

against an atomistic representation [189], and found good agreement between the structural and dynamic characteristics of the isolated polymer. We then extended our CG representation to a small polymer brush construct to model the system described in [9], solvated the system with BMW water, and placed a single R6G-CG within the simulation volume. Molecular dynamics simulations were then performed for a temporal length of 20 ns at temperatures of 300 K, 350 K, 400 K, and 450 K. From these simulations we extracted dynamical characteristics of the ions, of polystyrene sulfonate, and of R6G-CG, including diffusion constants for Na+ and R6G-CG, persistence lengths and radii of gyration for of PSS, radial distribution functions for Na^+/SO_3^- distances, and rotational decay times for R6G-CG.

These characteristics reveal several properties of the system under examination. Firstly, diffusion constants extracted for Na^+ ions reveal temperature dependent ion mobility, and the radial distribution functions for Na^+/SO_3^- distances reveal at least 2 preferred ionic arrangements for charge-shielding within the polyelectrolytic matrix that are in good agreement with previous atomistic simulations [189]. As expected, the average R_g for the polymer backbone is larger and the distribution width is broader at higher temperature, and the positive slope in the time trajectories suggests that the systems are not thermally equilibrated after 20 ns of simulated time. The persistence lengths of polymers in the brush construct suggest that configurational space is sampled more rapidly at higher temperature but the widths of the PL distributions do not change appreciably at higher temperature, suggesting the polymer samples the same configurational space at a rate proportional to temperature. Furthermore, the

reoccurrence of particular PLs within time trajectories at lower temperatures indicates that the polymer occupies preferred conformations to some degree. Concerning R6G-CG, the rates of translational diffusion of R6G-CG within the polymer brush constructs show that they are temperature dependent. Furthermore, the interaction of R6G-CG with the polymer brush is observed via a reduction in the translational diffusion constant by an order of magnitude. We also observe that reduced translational diffusion is not associated with a decrease in the rate of rotational diffusion, suggesting that the hindrance of translational diffusion is steric in nature. Moreover, this conclusion is further supported by the rotational decay times extracted from the simulations. While the rotational decay time of R6G-CG decreases with increasing temperature as expected, rotational diffusion is on the order of 1-100 picoseconds, thereby indicating fast rotation within the polymer brush at all temperatures. We therefore conclude that, despite the electrostatic charge on the molecular ion, that hindrance of translational diffusion is primarily steric in nature.

The reduction of the rate of translational diffusion by an order of magnitude is in agreement with the empirical data for a neutral fluorophore [10, 11], but is not in agreement with the empirical data concerning the ionic R6G probe [9], in which the authors observed a reduction in the rate of translational diffusion by approximately four orders of magnitude. Furthermore, our simulations indicate rotational diffusion on the order of 1-100 picoseconds while the same empirical study found hindered rotational diffusion on the order of milliseconds. The authors of the empirical study attribute the decreased rates of rotational and translational diffusion to electrostatic interaction,

which is suggested by their results concerning a neutral fluorophore. This suggests that, while the steric aspects of our simulation show good agreement with empirical data, the electrostatic aspects remain questionable. While this may simply be the result of the PSS oligomers constructed in this simulation being too short, or equivalently, too sparsely spaced in the polymer brush construct, thereby allowing free rotation and faster translational diffusion of R6G within the water-filled gaps between the oligomers, other factors may contribute as well. Specifically, the electrostatic cutoff of 1.4 *nm* that is allowed by the force field parameterization is much smaller than the Debye length calculated for R6G (approximately 10 *nm* at 298 K), which may result in the simulated molecular ion feeling less electrostatic force than is the case in the empirical system, leading to faster rates of translational and rotational diffusion. Simulations testing the first hypothesis regarding chain length and chain spacing are underway, but the results presented here suggest that a more complicated electrostatic model may be necessary to model these types of dynamical electrostatic phenomena.

Chapter 9

Conclusions and Future Directions

9.1 Conclusions

In conclusion, we have performed smFRET experiments on a system comprised of VEGF and the aV aptamer as a means to quantify structural fluctuations within the aptamer and observed the aV/VEGF interaction. Specifically, we found that the aptamer occupies a range of conformations that displays Mg^{2+} dependence as well as VEGF dependence. In particular, we find that the aptamer favors the predicted lowest energy conformation under all conditions, especially at higher Mg^{2+} concentration, and that the conformational equilibrium shifts towards lower FRET efficiency states in the presence of VEGF, indicated the aV/VEGF interaction occurs as expected. However, experimental noise limits the amount of available information within the experimental results and does not allow for more specific conclusions to be drawn.

Next, a method to identify and remove a considerable source of uncertainty in smFRET time trajectories – photoblinks – was developed and tested on simulated smFRET data. It is particularly important to maintain an objective approach in the removal of such phenomena. To such an end, a method based in Bayesian probability was developed and tested. The development of an unbiased method of photoblink detection eliminates the need to manually preprocess the trajectories, and perhaps more importantly, removes bias introduced into the measurement by subjective selection of photoblink thresholds. Tests of the algorithm's efficacy on simulated smFRET data resulted in nearly complete elimination of photoblinks with little effect on the actual data. This improvement is observed both in the ensemble analysis of structural distributions, and in kinetic analysis of dwell times. Although there are caveats involved with the method of photoblink detection, we have also shown that the caveats can be avoided through establishment of a lower efficiency boundary.

Another source of uncertainty in smFRET measurements is that of electronic shot-noise. To address these maleficent effects, a two component interpretation of noise observed in smFRET signals was taken, allowing for the removal the component we can quantify, thereby enhancing the accuracy of these measurements. In particular, a method based in wavelet decomposition was developed and tested on simulated smFRET trajectories. It was found that the widths of the photon distributions acquired from smFRET experiments can be significantly reduced by the application of the wavelet-based method, leading to a significant increase in the information content of the smFRET data. The method was found to be effective under a range of simulated and

experimental circumstances, including an oscillating system, an indistinguishable system, a resolvable two state system, and a variety of experimental systems containing a one, two, and multiple states. It is of note that the method is applied with only one assumption, that photon collection is a Poisson process that may be approximated normally. This assumption is valid under all but the most extreme of experimental circumstances, and as such the method constitutes an objective approach, being based solely on known properties of the experimental signals, that is capable of removing a substantial portion of the shot-noise contribution from a smFRET trajectory.

This wavelet-based method was then modified to be specific to smFRET measurements. These modifications include the determination that the best wavelet basis for shot-noise removal is the simplest of all bases, the Haar basis. The method known as cycle-spinning was applied to remove the translation dependence that is inherent to the wavelet transformation, and resulted in the reduction of high frequency artifacts that are introduced by the denoising procedure. A variety of thresholding, or noise removal, methods were tested, and it was found that two methods, firm thresholding and hard thresholding perform best in the applications to simulated data. Lastly, a time-local noise estimation method was developed to accommodate the non-stationary Poisson processes that are inherent to smFRET measurements, and the performance of the denoising procedure with respect to non-stationary processes was greatly improved. Overall, through the application of these methods to two model systems, it was shown that, with respect to the procedure described previously, these modifications improve the accuracy of denoised estimates by more than 200 %. In

general, it was found that the application of the modified wavelet denoising procedure resulted in up to 75 % reduction of the shot-noise contribution to smFRET data.

This objective, wavelet-based procedure was then applied to smFRET trajectories obtained for the neurological ion channel protein known as GluR2. Specifically, the agonist binding domain of the GluR2 receptor was studied in various forms that included the glutamate-bound, apo, and T686S mutant proteins. Prior to denoising, broad and featureless smFRET efficiency distributions were observed, leaving little information to be extracted from the data. However, after wavelet-based denoising, it was found that, in agreement with theoretical prediction, the glutamate-bound form of the GluR2-ABD is not in a locked conformation, but instead explores a range of conformations in a sequential equilibrium. Similar analyses were performed on the apo and T686S mutant forms, and it was found that these forms of the protein are more flexible than their glutamate-bound counterpart, allowing for the hypothesis that protein flexibility plays a role in the activity and desensitization of the neurological ion channel. In a general sense, the results of these studies suggested that activation of the neurological ion channel is dependent not on binding cleft closure, but on the probability the closed cleft conformation is occupied considering the range of conformations explored by the protein.

Lastly, the focus of my research shifted to a system involving the transport of a molecular ionic fluorophore within an electrostatically charged polystyrene sulfonate polymer brush. In particular atomistic and coarse-grained models were developed for the fluorophore Rhodamine 6G. A time dilation of 1.7 was measured for the CG model

and atomistic structural angle distributions were used to calibrate the CG angles. A CG representation of PSS was then constructed, based on a previously constructed CG model of polystyrene, and a polymer brush representation was also constructed. The CG model was solvated in multipolar water and molecular dynamics simulations were performed for four differing temperatures and a temporal length of 20 ns. From these simulations we extracted dynamical characteristics of the ions, of polystyrene sulfonate, and of R6G-CG, including diffusion constants for Na^+ and R6G-CG, persistence lengths and radii of gyration for of PSS, radial distribution functions for Na^+/SO_3^- distances, and rotational decay times for R6G-CG. In general, good agreement between previous atomistic simulations and the CG simulation was found. Such agreement, however, was not found between the simulations and empirical evidence. Specifically, the rates of the fluorophore's translational and rotational diffusion were not accurately reproduced, leading to two hypotheses about the nature on the disagreement. The first hypothesis is simple in nature, and involves the spacing of the polyelectrolytic polymer brush. The second hypothesis was one involving the computation of the electrostatics of the system. Specifically, considering the Debye length of the fluorophore in this environment is calculated to be approximately 6.5 times larger than the electrostatic cutoff radius, the distance at which electrostatic interactions are calculated may not be large enough to accurately model the physical system. Simulations are underway in which the first hypothesis concerning structural characteristics of the polymer brush will be tested to determine the exact nature of the disagreement.

9.2 Future Directions

9.2.1 *Small-Scale Extension of the R6G-CG/PSS CG Model*

The coarse-grained molecular dynamics simulations described in chapter 8 reveal several things about both the theoretical simulations as well as the experimental system. Firstly, the simulations reveal that the system has not reached thermal equilibrium after 20 ns of simulated time. Secondly, they reveal that the empirical aspects of rotational and translational diffusion are not reproduced accurately. As discussed in chapter 8, this may be a limitation of the force field parameterization or it may be that the PSS chains are too sparsely spaced to allow for the appropriate polymer dynamics to occur. To remedy the first situation, longer simulations are currently underway that will span 100 – 500 ns. To test the latter hypothesis concerning the aspects of rotational and translational diffusion of R6G within a PSS polymer brush, simulations are underway in which the lengths of the PSS chains have been extended by a factor of three but the chain spacing remains. This will effectively move the PSS chains closer together and allow them to interact in a more ordered fashion.

9.2.2 *Large-Scale Extension of the R6G-CG/PSS CG Model and Synthetic Photon Trajectories*

As computational facilities continue to increase, we may continue to expand our simulated systems both in time and space. Large scale simulations of the R6G-CG/PSS system are planned, in which the construct will be expanded to a spatial scale that is comparable to the experimental observation volume and to a temporal scale that is comparable to the experimental time resolution. This will allow for the generation of synthetic photon trajectories that may be directly compared and analyzed alongside the

experimental data, thereby allowing for the quantification and distinction of photophysical phenomena that often plague experimental interpretations. In fact, a general infrastructure for generating polarized, FCS-style, 3-component, synthetic photon trajectories from molecular dynamics trajectories has already been developed. This method takes advantage of the statistics of photon kinetics and uses stochastic modeling to generate synthetic trajectories from spatial trajectories in a photon-by-photon fashion.

9.2.3 *Extension of the Stochastic Photon Model to Other Photophysical Phenomena*

The aforementioned photophysical phenomena that hinder experimental interpretation may also be stochastically modeled if their kinetics are known or can be measured directly. Furthermore, other experimental methods, such as smFRET for example, may be modeled using photon statistics and stochastic modeling as well. In fact, such phenomena are trivial extensions to the basics infrastructure that is already in place. Such an approach that uses stochastic modeling of photon kinetics in combination with molecular dynamics simulation to generate synthetic data would be of considerable practical value in the understanding and quantification of troublesome photophysical phenomena. For example, it is known that the instantaneous orientation of the molecular transition dipoles within the two fluorophores in smFRET experiments greatly affects the efficiency of energy transfer. As such, it is a large source of uncertainty in smFRET measurements. However, a microscopic description of such instantaneous orientation coupled with a synthesized manifestation within macroscopic data may aid in its identification on the macroscopic level. Furthermore, knowledge of

such microscopic detail may possibly lead to its use in an advantageous manner, thereby transforming what had been an experimental disadvantage into an experimental advantage.

References

1. Landes, C.F., et al., *Structural landscape of isolated agonist-binding domains from single AMPA receptors.* Nat Chem Biol, 2011. **7**(3): p. 168-173.
2. Lau, A.Y. and B. Roux, *The Free Energy Landscapes Governing Conformational Changes in a Glutamate Receptor Ligand-Binding Domain.* Structure, 2007. **15**(10): p. 1203-1214.
3. Lau, A.Y. and B. Roux, *Structural biology: One domain, multiple conformations.* Nat Chem Biol, 2011. **7**(3): p. 130-131.
4. Taylor, J.N. and C.F. Landes, *Improved Resolution of Complex Single-Molecule FRET Systems via Wavelet Shrinkage.* The Journal of Physical Chemistry B, 2011. **115**(5): p. 1105-1114.
5. Taylor, J.N., D.E. Makarov, and C.F. Landes, *Denoising Single-Molecule FRET Trajectories with Wavelets and Bayesian Inference.* Biophysical Journal, 2010. **98**(1): p. 164-173.
6. Huang, B., M. Bates, and X. Zhuang, Annu. Rev. Biochem., 2009. **78**: p. 993.
7. Daniels, C.R., et al., *Permeability of anti-fouling PEGylated surfaces probed by fluorescence correlation spectroscopy.* Colloids and Surfaces B: Biointerfaces, 2011. **88**(1): p. 31-38.
8. Daniels, C.R., C. Reznik, and C.F. Landes, *Dye Diffusion at Surfaces: Charge Matters.* Langmuir, 2010. **26**(7): p. 4807-4812.
9. Reznik, C., et al., *Transient Three-Dimensional Orientation of Molecular Ions in an Ordered Polyelectrolyte Membrane.* The Journal of Physical Chemistry Letters, 2011. **2**(6): p. 592-598.
10. Reznik, C., et al., *Single Ion Diffusive Transport within a Poly(styrene sulfonate) Polymer Brush Matrix Probed by Fluorescence Correlation Spectroscopy.* The Journal of Physical Chemistry B, 2008. **112**(35): p. 10890-10897.
11. Reznik, C., et al., *Single Molecule Spectroscopy Reveals Heterogeneous Transport Mechanisms for Molecular Ions in a Polyelectrolyte Polymer Brush.* The Journal of Physical Chemistry B, 2009. **113**(44): p. 14611-14618.
12. Darugar, Q., et al., *Human T-Cell Lymphotropic Virus Type 1 Nucleocapsid Protein-Induced Structural Changes in Transactivation Response DNA Hairpin Measured by Single-Molecule Fluorescence Resonance Energy Transfer.* The Journal of Virology, 2008. **82**(24): p. 12164-12171.
13. Selvin, P.R., *The renaissance of fluorescence resonance energy transfer.* Nat Struct Mol Biol, 2000. **7**(9): p. 730-734.
14. Taylor, J.N., et al., *Dynamics of an anti-VEGF DNA aptamer: A single-molecule study.* Biochem. Biophys. Res. Comm., 2008. **373**(2): p. 213-218.
15. Gopich, I., *Theory of photon statistics in single-molecule Forster resonance energy transfer.* J. Chem. Phys, 2005. **122**.
16. Gopich, I.V. and A. Szabo, *Decoding the Pattern of Photon Colors in Single-Molecule FRET.* The Journal of Physical Chemistry B, 2009. **113**(31): p. 10965-10973.
17. Watkins, L.P. and H. Yang, *Information Bounds and Optimal Analysis of Dynamic Single Molecule Measurements.* Biophys. J., 2004. **86**(6): p. 4015-4029.
18. Ober, R.J., S. Ram, and E.S. Ward, *Localization Accuracy in Single-Molecule Microscopy.* Biophys. J., 2004. **86**(2): p. 1185-1200.
19. Schenter, G.K., H.P. Lu, and X.S. Xie, *Statistical Analyses and Theoretical Models of Single-Molecule Enzymatic Dynamics.* J. Phys. Chem. A, 1999. **103**(49): p. 10477-10488.

20. McKinney, S.A., C. Joo, and T. Ha, *Analysis of single-molecule FRET trajectories using hidden Markov modeling.* Biophys. J., 2006. **91**(5): p. 1941-1951.
21. Andrec, M., R.M. Levy, and D.S. Talaga, *Direct Determination of Kinetic Rates from Single-Molecule Photon Arrival Trajectories Using Hidden Markov Models.* The Journal of Physical Chemistry A, 2003. **107**(38): p. 7454-7464.
22. Yang, H., *Detection and characterization of dynamical heterogeneity in an event series using wavelet correlation.* J. Chem. Phys., 2008. **129**(7): p. 074701-11.
23. Baba, A. and T. Komatsuzaki, *Construction of effective free energy landscape from single-molecule time series.* Proceedings of the National Academy of Sciences, 2007. **104**(49): p. 19297-19302.
24. Li, C.-B., H. Yang, and T. Komatsuzaki, *Multiscale complex network of protein conformational fluctuations in single-molecule time series.* Proceedings of the National Academy of Sciences, 2008. **105**(2): p. 536-541.
25. Rylance, G.J., et al., *Topographical complexity of multidimensional energy landscapes.* Proceedings of the National Academy of Sciences, 2006. **103**(49): p. 18551-18555.
26. Donoho, D.L., *De-noising by soft-thresholding.* Information Theory, IEEE Transactions on, 1995. **41**(3): p. 613-627.
27. Di Fiori, N. and A. Meller, *The Effect of Dye-Dye Interactions on the Spatial Resolution of Single-Molecule FRET Measurements in Nucleic Acids.* Biophysical Journal, 2010. **98**(10): p. 2265-2272.
28. Ha, T. and P. Tinnefeld, *Photophysics of Fluorescent Probes for Single-Molecule Biophysics and Super-Resolution Imaging.* Annual Review of Physical Chemistry, 2012. **63**(1): p. 595-617.
29. Ranjit, S., K. Gurunathan, and M. Levitus, *Photophysics of Backbone Fluorescent DNA Modifications: Reducing Uncertainties in FRET.* The Journal of Physical Chemistry B, 2009. **113**(22): p. 7861-7866.
30. Leung, D.W., et al., *Vascular endothelial growth factor is a secreted angiogenic mitogen.* Science, 1989. **246**(4935): p. 1306-1309.
31. Kim, K.J., et al., *Inhibition of vascular endothelial growth factor-induced angiogenesis suppresses tumour growth in vivo.* Nature, 1993. **362**(6423): p. 841-844.
32. Ferrara, N., et al., *Discovery and development of bevacizumab, an anti-VEGF antibody for treating cancer.* Nat Rev Drug Discov, 2004. **3**(5): p. 391-400.
33. Holash, J., et al., *VEGF-Trap: A VEGF blocker with potent antitumor effects.* 2002. **99**(17): p. 11393-11398.
34. Ellington, A.D. and J.W. Szostak, *In vitro selection of RNA molecules that bind specific ligands.* Nature, 1990. **346**(6287): p. 818-822.
35. Tuerk, C. and L. Gold, *Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase.* Science, 1990. **249**(4968): p. 505-510.
36. Gouaux, E., *Structure and function of AMPA receptors.* The Journal of Physiology, 2004. **554**(2): p. 249-253.
37. Viterbi, A., *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.* Information Theory, IEEE Transactions on, 1967. **13**(2): p. 260-269.
38. Metropolis, N., et al., *Equation of State Calculations by Fast Computing Machines.* The Journal of Chemical Physics, 1953. **21**(6): p. 1087-1092.
39. Berendsen, H.J.C., D. van der Spoel, and R. van Drunen, *GROMACS: A message-passing parallel molecular dynamics implementation.* Computer Physics Communications, 1995. **91**(1–3): p. 43-56.

40. Hess, B., et al., *GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation.* Journal of Chemical Theory and Computation, 2008. **4**(3): p. 435-447.
41. Lindahl, E., B. Hess, and D. van der Spoel, *GROMACS 3.0: a package for molecular simulation and trajectory analysis.* Journal of Molecular Modeling, 2001. **7**(8): p. 306-317.
42. Van Der Spoel, D., et al., *GROMACS: Fast, flexible, and free.* Journal of Computational Chemistry, 2005. **26**(16): p. 1701-1718.
43. Nimjee, S.M., C.P. Rusconi, and B.A. Sullenger, *APTAMERS: An Emerging Class of Therapeutics.* Ann. Rev. Med., 2005. **56**(1): p. 555-583.
44. Ruckman, J., *2'-Fluoropyrimidine RNA-based Aptamers to the 165-Amino Acid Form of Vascular Endothelial Growth Factor (VEGF165).* J. Biol. Chem, 1998. **273**(32).
45. Ng, E., *Pegaptanib, a targeted anti-VEGF aptamer for ocular vascular disease.* Nat. Rev. Drug Disc., 2006. **5**(2).
46. Mallikaratchy, P., et al., *Selection of DNA ligands for protein kinase C-[small delta].* 2006(30): p. 3229-3231.
47. Potty, A. 2007, University of Houston.
48. Brion, P. and E. Westhof, *Hierarchy and Dynamics of RNA Folding.* Ann. Rev. Biophys. Biomol. Struct., 1997. **26**(1): p. 113-137.
49. Misra, V.K. and D.E. Draper, *The linkage between magnesium binding and RNA folding.* Journal of Molecular Biology, 2002. **317**(4): p. 507-521.
50. Thirumalai, D., et al., *Early Events in RNA Folding.* Ann. Rev. Phys. Chem., 2001. **52**(1): p. 751-762.
51. Bokinsky, G., et al., *Single-Molecule Transition-state analysis of RNA folding.* Proc. Nat. Acad. Sci. USA, 2003. **100**(16): p. 9302-9307.
52. Draper, D.E., D. Grilley, and A.M. Soto, *Ions and RNA Folding.* Ann. Rev. Biophys. Biomol. Struct., 2005. **34**(1): p. 221-243.
53. Kim, H.D., et al., *Mg^{2+}-dependent conformational change of RNA studied by fluorescence correlation and FRET on immobilized single molecules.* Proc. Nat. Acad. Sci. USA, 2002. **99**(7).
54. Pyle, A., *Metal ions in the structure and function of RNA.* Journal of Biological Inorganic Chemistry, 2002. **7**(7): p. 679-690.
55. Ha, T., *Single-Molecule Fluorescence Resonance Energy Transfer.* Methods, 2001. **25**(1): p. 78-86.
56. Moerner, W.E. and M. Orrit, *Illuminating Single Molecules in Condensed Matter.* Science, 1999. **283**(5408): p. 1670.
57. Weiss, S., *Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy.* Nat. Struct. Mol. Biol., 2000. **7**(9): p. 724-729.
58. Zhuang, X., *Single-Molecule RNA Science.* Ann. Rev. Biophys. Biomol. Struct., 2005. **34**(1): p. 399-414.
59. Liu, S., et al., *Single-molecule Chemistry and Biology Special Feature: Dissecting the multistep reaction pathway of an RNA enzyme by single-molecule kinetic "fingerprinting".* Proc. Nat. Acad. Sci. USA, 2007. **104**(31): p. 12634-12639.
60. Nahas, M.K., et al., *Observation of internal cleavage and ligation reactions of a ribozyme.* Nat. Struct. Mol. Biol., 2004. **11**(11): p. 1107-1113.
61. Bokinsky, G., et al., *Two Distinct Binding Modes of a Protein Cofactor with its Target RNA.* J. Mol. Bio., 2006. **361**: p. 771-784.

62. Cosa, G., et al., *Evidence for Non-Two-State Kinetics in the Nucleocapsid Protein Chaperoned Opening of DNA Hairpins.* J. Phys. Chem. B, 2006. **110**(5): p. 2419-2426.
63. Landes, C.F., et al., *Single-Molecule Study of the Inhibition of HIV-1 Transactivation Response Region DNA/DNA Annealing by Argininamide.* J. Am. Chem. Soc., 2007. **129**(33): p. 10181-10188.
64. Nir, E., et al., *Shot-Noise Limited Single-Molecule FRET Histograms: Comparison between Theory and Experiments.* J. Phys. Chem. B, 2006. **110**(44): p. 22103-22124.
65. Sabanayagam, C.R., J.S. Eid, and A. Meller, *Long time scale blinking kinetics of cyanine fluorophores conjugated to DNA and its effect on FÃ¶rster resonance energy transfer.* Journal of Chemical Physics, 2005. **123**(22): p. N.PAG.
66. Rasnik, I., S.A. McKinney, and T. Ha, *Nonblinking and long-lasting single-molecule fluorescence imaging.* Nat Meth, 2006. **3**(11): p. 891-893.
67. Zhang, X., et al., *Engineered 5S Ribosomal RNAs Displaying Aptamers Recognizing Vascular Endothelial Growth Factor and Malachite Green.* J. Mol. Recognit., 2008: p. in press.
68. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nuc. Acids Res., 2003. **31**(13): p. 3406-3415.
69. Dickson, R.M., et al., *On/off blinking and switching behaviour of single molecules of green fluorescent protein.* Nature, 1997. **388**(6640): p. 355-358.
70. Huang, Z., et al., *Spectral Identification of Specific Photophysics of Cy5 by Means of Ensemble and Single Molecule Measurements.* J. Phys. Chem. A, 2005. **110**(1): p. 45-50.
71. Jia, K., et al., *Characterization of Photoinduced Isomerization and Intersystem Crossing of the Cyanine Dye Cy3.* J. Phys. Chem. A, 2007. **111**(9): p. 1593-1597.
72. Haran, G., *Single-molecule fluorescence spectroscopy of biomolecular folding.* J. Phys.: Condens. Matter, 2003. **15**(32): p. R1291-R1317.
73. Schuler, B., E.A. Lipman, and W.A. Eaton, *Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy.* Nature, 2002. **419**(6908): p. 743-747.
74. Winkler, R.L., *An Introduction to Bayesian Inference and Decision*. 1972, New York: Holt, Rinehart and Winston.
75. Liu, P., et al., *A Bayesian statistics approach to multiscale coarse graining.* J. Chem. Phys., 2008. **129**(21): p. 214114-11.
76. Sabanayagam, C.R., J.S. Eid, and A. Meller, *Using fluorescence resonance energy transfer to measure distances along individual DNA molecules: Corrections due to nonideal transfer.* J. Chem. Phys., 2005. **122**(6): p. 061103-5.
77. Haar, A., *Zur Theorie der orthogonalen Funktionensysteme.* Mathematische Annalen, 1910. **69**(3): p. 331-371.
78. Makarov, D.E. and H. Metiu, *A model for the kinetics of protein folding: Kinetic Monte Carlo simulations and analytical results.* J. Chem. Phys., 2002. **116**(12): p. 5205-5216.
79. Fichthorn, K.A. and W.H. Weinberg, *Theoretical foundations of dynamical Monte Carlo simulations.* J. Chem. Phys., 1991. **95**(2): p. 1090-1096.
80. Metiu, H., Y.-T. Lu, and Z. Zhang, *Epitaxial growth and the art of computer simulations. (Frontiers in Materials Science).* Science, 1992. **255**(5048): p. 1088-1093.
81. Voter, A.F., *Classically exact overlayer dynamics: Diffusion of rhodium clusters on Rh(100).* Phys. Rev. B, 1986. **34**(10): p. 6819.
82. Sabanayagam, C.R., J.S. Eid, and A. Meller, *Long time scale blinking kinetics of cyanine fluorophores conjugated to DNA and its effect on F[o-umlaut]rster resonance energy transfer.* J. Chem. Phys., 2005. **123**(22): p. 224708-7.

83. Darugar, Q., et al., *Human T-Cell Lymphotropic Virus Type 1 Nucleocapsid Protein-Induced Structural Changes in Transactivation Response DNA Hairpin Measured by Single-Molecule Fluorescence Resonance Energy Transfer.* J. Virology, 2008. **82**(24): p. 12164-12171.
84. Moerner, W.E. and D.P. Fromm, *Methods of single-molecule fluorescence spectroscopy and microscopy.* Rev. Sci. Inst., 2003. **74**(8): p. 3597-3619.
85. Barsegov, V. and S. Mukamel, *Probing single molecule kinetics by photon arrival trajectories.* J. Chem. Phys., 2002. **116**(22): p. 9802-9810.
86. Enderlein, J., Robbins, David L., Ambrose, W. Patrick, Goodwin, Peter M., Keller, Richard A. , *The statistics of single molecule detection: An overview.* Bioimaging, 1997. **5**(3): p. 88-98.
87. Gopich, I. and A. Szabo, *Theory of photon statistics in single-molecule Forster resonance energy transfer.* J. Chem. Phys., 2005. **122**(1): p. 014707-1-014707-18.
88. Gopich, I.V. and A. Szabo, *Statistics of transitions in single molecule kinetics.* J. Chem. Phys., 2003. **118**(1): p. 454-455.
89. Nettels, D., et al., *Ultrafast dynamics of protein collapse from single-molecule photon statistics.* Proc. Natl. Acad. Sci. U.S.A., 2007. **104**(8): p. 2655-2660.
90. Onuchic, J.N., J. Wang, and P.G. Wolynes, *Analyzing single molecule trajectories on complex energy landscapes using replica correlation functions.* Chem. Phys., 1999. **247**(1): p. 175-184.
91. Wang, J. and P. Wolynes, *Intermittency of Single Molecule Reaction Dynamics in Fluctuating Environments.* Phys. Rev. Lett., 1995. **74**(21): p. 4317-4320.
92. Yang, H. and X.S. Xie, *Statistical approaches for probing single-molecule dynamics photon-by-photon.* Chem. Phys., 2002. **284**(1-2): p. 423-437.
93. Yang, S. and J. Cao, *Two-Event Echos in Single-Molecule Kinetics: A Signature of Conformational Fluctuations.* J. Am. Chem. Soc., 2001. **105**(28): p. 6536-6549.
94. English, B.P.M., Wei; van Oijen, Antoine M; Lee, Kang Taek; Luo, Guobin; Sun, Hongye; Cherayil, Binny J.; Kou, S.C.; Xie, X. Sunney *Ever-fluctuating single enzyme molecules: Michaelis-Menten revisited.* Nat. Chem. Biol., 2006. **2**(2): p. 87-94.
95. Okumus, B., et al., *Vesicle Encapsulation Studies Reveal that Single Molecule Ribozyme Heterogeneities Are Intrinsic.* Biophys. J., 2004. **87**(4): p. 2798-2806.
96. Adak, S., *Time-Dependent Spectral Analysis of Nonstationary Time Series.* J. Amer. Statistical Assoc., 1998. **93**(444): p. 1488-1501.
97. Cai, C. and P.d.B. Harrington, *Different Discrete Wavelet Transforms Applied to Denoising Analytical Data.* J. Chem. Inf. Comput. Sci., 1998. **38**(6): p. 1161-1170.
98. Mallat, S., G. Papanicolaou, and Z. Zhang, *Adaptive Covariance Estimation of Locally Stationary Processes.* Ann. Stat., 1998. **26**(1): p. 1-47.
99. Ombao, H.C., et al., *Automatic Statistical Analysis of Bivariate Nonstationary Time Series. In Memory of Jonathan A. Raz.* J. Amer. Statistical Assoc., 2001. **96**(454): p. 543-560.
100. Mallat, S., *A Wavelet Tour of Signal Processing*. 1998, San Diego, London: Academic Press.
101. Mallat, S.G., *A theory for multiresolution signal decomposition: the wavelet representation.* IEEE Trans. Patt. An. Mach. Int., 1989. **11**(7): p. 674-693.
102. Daubechies, I., *Orthonormal bases of compactly supported wavelets.* Commun. Pure. Appl. Math., 1988. **41**(7): p. 909-996.
103. Donoho, D.L. and I.M. Johnstone, *Ideal Spatial Adaptation by Wavelet Shrinkage.* Biometrika, 1994. **81**(3): p. 425-455.

104. Meyer, Y., *Wavelets: Algorithms and Applications*. 1993, Cambridge: Cambridge University Press.
105. Nason, G.P., S. Rainer von, and G. Kroisandt, *Wavelet Processes and Adaptive Estimation of the Evolutionary Wavelet Spectrum.* J. Roy. Statistical Soc. Ser. B 2000. **62**(2): p. 271-292.
106. Nason, G.P., *Wavelet Methods in Statistics with R*. Use R!, ed. R.H. Gentleman, K.; Parmigiani, G. 2008, New York: Springer.
107. Donoho, D. and J. Jin, *Higher criticism thresholding: Optimal feature selection when useful features are rare and weak.* Proc. Natl. Acad. Sci. U.S.A., 2008. **105**(39): p. 14790-14795.
108. Berger, J., M. Goldberg, and R. Coifman, *Removing Noise from Music Using Local Trigonometric Bases and Wavelet Packets.* J. Audio Eng. Soc., 1994. **42**(10): p. 808-818.
109. Daubechies, I., *10 Lectures on Wavelets*. CBMS-NSF regional conference series in applied mathematics. 1992, Philadelphia, PA: Society for Industrial and Applied Mathematics. 357.
110. Donoho, D.L., *De-noising by soft-thresholding.* IEEE Trans. Inf. Theory, 1995. **41**(3): p. 613-627.
111. Murphy, M.C., et al., *Probing Single-Stranded DNA Conformational Flexibility Using Fluorescence Spectroscopy.* Biophys. J., 2004. **86**(4): p. 2530-2537.
112. Cannone, F., et al., *Environment effects on the oscillatory unfolding kinetics of GFP.* Eur. Biophys. J., 2007. **36**(7): p. 795-803.
113. Iqbal, A., et al., *Orientation dependence in fluorescent energy transfer between Cy3 and Cy5 terminally attached to double-stranded nucleic acids.* Proc. Natl. Acad. Sci. U.S.A., 2008. **105**(32): p. 11176-11181.
114. Ranjit, S., K. Gurunathan, and M. Levitus, *Photophysics of Backbone Fluorescent DNA Modifications: Reducing Uncertainties in FRET.* J. Phys. Chem. B, 2009. **113**(22): p. 7861-7866.
115. Rasnik, I., et al., *DNA-binding Orientation and Domain Conformation of the E. coli Rep Helicase Monomer Bound to a Partial Duplex Junction: Single-molecule Studies of Fluorescently Labeled Enzymes.* J. Mol. Biol., 2004. **336**(2): p. 395-408.
116. Brakmann, S., *Single-molecule analysis: A ribosome in action.* Nature, 2010. **464**(7291): p. 987-988.
117. Forget, A.L. and S.C. Kowalczykowski, *Single-molecule imaging brings Rad51 nucleoprotein filaments into focus.* Trends in Cell Biology, 2010. **20**(5): p. 269-276.
118. Goldsmith, R.H. and W.E. Moerner, *Watching conformational- and photodynamics of single fluorescent proteins in solution.* Nature Chemistry, 2010. **2**(3): p. 179-186.
119. Lord, S.J., H.-L.D. Lee, and W.E. Moerner, *Single-Molecule Spectroscopy and Imaging of Biomolecules in Living Cells.* Analytical Chemistry, 2010. **82**(6): p. 2192-2203.
120. Pugh, S.D., et al., *Single-Molecule Studies of the Im7 Folding Landscape.* Journal of Molecular Biology, 2010. **398**(1): p. 132-145.
121. Gurunathan, K. and M. Levitus, *FRET Fluctuation Spectroscopy of Diffusing Biopolymers: Contributions of Conformational Dynamics and Translational Diffusion.* The Journal of Physical Chemistry B, 2009. **114**(2): p. 980-986.
122. Torres, T. and M. Levitus, *Measuring Conformational Dynamics: A New FCS-FRET Approach.* The Journal of Physical Chemistry B, 2007. **111**(25): p. 7392-7400.
123. Wang, Y. and H.P. Lu, *Bunching Effect in Single-Molecule T4 Lysozyme Nonequilibrium Conformational Dynamics under Enzymatic Reactions.* The Journal of Physical Chemistry B, 2010. **114**(19): p. 6669-6674.

124. Chung, H.S., J.M. Louis, and W.A. Eaton, *Distinguishing between Protein Dynamics and Dye Photophysics in Single-Molecule FRET Experiments.* Biophysical Journal, 2010. **98**(4): p. 696-706.
125. Liu, Y., et al., *A Comparative Study of Multivariate and Univariate Hidden Markov Modelings in Time-Binned Single-Molecule FRET Data Analysis.* The Journal of Physical Chemistry B, 2010. **114**(16): p. 5386-5403.
126. Nir, E., et al., *Shot-Noise Limited Single-Molecule FRET Histograms: Comparison between Theory and Experiments.* The Journal of Physical Chemistry A, 2006. **110**(44): p. 22103-22124.
127. Harvey, B.J., C. Perez, and M. Levitus, *DNA Sequence-Dependent Enhancement of Cy3 Fluorescence.* Photochem. Photobiol. Sci., 2009. **8**: p. 1105-1110.
128. Talaga, D.S., *Information-Theoretical Analysis of Time-Correlated Single-Photon Counting Measurements of Single Molecules.* The Journal of Physical Chemistry A, 2009. **113**(17): p. 5251-5263.
129. Watkins, L.P. and H. Yang, *Information Bounds and Optimal Analysis of Dynamic Single Molecule Measurements.* Biophysical Journal, 2004. **86**(6): p. 4015-4029.
130. Ensign, D.L. and V.S. Pande, *Bayesian Detection of Intensity Changes in Single Molecule and Molecular Dynamics Trajectories.* The Journal of Physical Chemistry B, 2009. **114**(1): p. 280-292.
131. Chipman, H.A., E.D. Kolaczyk, and R.E. McCulloch, *Adaptive Bayesian Wavelet Shrinkage.* Journal of the American Statistical Association, 1997. **92**(440): p. 1413-1421.
132. Daubechies, I., *10 Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics. Vol. 61. 1992, Philadelphia, PA: Society for Industrial and Applied Mathematics.
133. Donoho, D.L. and I.M. Johnstone, *Adapting to Unknown Smoothness via Wavelet Shrinkage.* Journal of the American Statistical Association, 1995. **90**(432): p. 1200-1224.
134. Gao, H.-Y., *Wavelet Shrinkage Denoising Using the Non-Negative Garrote.* Journal of Computational and Graphical Statistics, 1998. **7**(4): p. 469-488.
135. Gao, H.-Y. and A.G. Bruce, *WaveShrink with Firm Shrinkage.* Statistica Sinica, 1997. **7**: p. 855-871.
136. Ogden, T. and E. Parzen, *Change-point approach to data analytic wavelet thresholding.* Statistics and Computing, 1996. **6**(2): p. 93-99.
137. Vidakovic, B., *Nonlinear Wavelet Shrinkage with Bayes Rules and Bayes Factors.* Journal of the American Statistical Association, 1998. **93**(441): p. 173-179.
138. Vidakovic, B., *Statistical Modeling by Wavelets*. Wiley Series in Probability and Statistics. 1999, New York, NY: John Wiler & Sons, Inc.
139. Müller, P. and B. Vidakovic, *Bayesian Inference with Wavelets: Density Estimation.* Journal of Computational and Graphical Statistics, 1998. **7**(4): p. 456-468.
140. Perrin, C., B. Walczak, and D.L. Massart, *The Use of Wavelets for Signal Denoising in Capillary Electrophoresis.* Analytical Chemistry, 2001. **73**(20): p. 4903-4917.
141. Coifman, R.R. and D.L. Donoho, *Translation-Invariant De-Noising*, in *Lect. Notes Statist.*, A. Antoniadis and G. Oppenheim, Editors. 1995, Springer-Verlag: New York, NY. p. 125-150.
142. Kovac, A. and B.W. Silverman, *Extending the Scope of Wavelet Regression Methods by Coefficient-Dependent Thresholding.* Journal of the American Statistical Association, 2000. **95**(449): p. 172-183.
143. Hogg, R.V. and E.A. Tanis, *Probability and Statistical Inference*. 7 ed. 2006, Upper Saddle River, NJ: Pearson.

144. McKinney, S.A., C. Joo, and T. Ha, *Analysis of Single-Molecule FRET Trajectories Using Hidden Markov Modeling.* Biophysical Journal, 2006. **91**(5): p. 1941-1951.
145. Keinanen, K., et al., *A family of AMPA-selective glutamate receptors.* Science (Washington, D. C., 1883-), 1990. **249**(4968): p. 556-60.
146. Nakanishi, S. and M. Masu, *Molecular diversity and functions of glutamate receptors.* Annu Rev Biophys Biomol Struct, 1994. **23**: p. 319-48.
147. Dingledine, R., et al., *The glutamate receptor ion channels.* Pharmacol. Rev., 1999. **51**(1): p. 7-61.
148. Fleming, J.J. and P.M. England, *AMPA receptors and synaptic plasticity: a chemist's perspective.* Nat. Chem. Biol., 2010. **6**(2): p. 89-97.
149. Rosenmund, C., Y. Stern-Bach, and C.F. Stevens, *The tetrameric structure of a glutamate receptor channel.* Science (Washington, D. C.), 1998. **280**(5369): p. 1596-1599.
150. Sobolevsky, A.I., M.P. Rosconi, and E. Gouaux, *X-ray structure, symmetry and mechanism of an AMPA-subtype glutamate receptor.* Nature (London, U. K.), 2009. **462**(7274): p. 745-756.
151. Armstrong, N. and E. Gouaux, *Mechanisms for activation and antagonism of an AMPA-sensitive glutamate receptor: crystal structures of the GluR2 ligand binding core.* Neuron, 2000. **28**(1): p. 165-181.
152. Armstrong, N., M. Mayer, and E. Gouaux, *Tuning activation of the AMPA-sensitive GluR2 ion channel by genetic adjustment of agonist-induced conformational changes.* Proc. Natl. Acad. Sci. U. S. A., 2003. **100**(10): p. 5736-5741.
153. Gouaux, E., *Structure and function of AMPA receptors.* J. Physiol. (Oxford, U. K.), 2004. **554**(2): p. 249-253.
154. Armstrong, N., et al., *Structure of a glutamate-receptor ligand-binding core in complex with kainate.* Nature (London), 1998. **395**(6705): p. 913-917.
155. Birdsey-Benson, A., et al., *Enhanced efficacy without further cleft closure: reevaluating twist as a source of agonist efficacy in AMPA receptors.* J. Neurosci., 2010. **30**(4): p. 1463-1470.
156. Ahmed, A.H., et al., *Structure of the S1S2 glutamate binding domain of GluR3.* Proteins Struct., Funct., Bioinf., 2009. **75**(3): p. 628-637.
157. Maltsev, A.S., et al., *Mechanism of Partial Agonism at the GluR2 AMPA Receptor: Measurements of Lobe Orientation in Solution.* Biochemistry, 2008. **47**(40): p. 10600-10610.
158. Ramanoudjame, G., et al., *Allosteric mechanism in AMPA receptors: a FRET-based investigation of conformational changes.* Proc. Natl. Acad. Sci. USA, 2006. **103**(27): p. 10473-10478.
159. Robert, A., et al., *AMPA receptor binding cleft mutations that alter affinity, efficacy, and recovery from desensitization.* J. Neurosci., 2005. **25**(15): p. 3752-3762.
160. Li, C.-B., H. Yang, and T. Komatsuzaki, *Multiscale complex network of protein conformational fluctuations in single-molecule time series.* Proc. Natl. Acad. Sci. USA, 2008. **105**(2): p. 536-541.
161. Schuler, B. and W.A. Eaton, *Protein folding studied by single-molecule FRET.* Curr. Opin. Struct. Biol., 2008. **18**(1): p. 16-26.
162. Flynn, E.M., et al., *Dynamic Active-Site Protection by the M. tuberculosis Protein Tyrosine Phosphatase PtpB Lid Domain.* J. Am. Chem. Soc., 2010. **132**(13): p. 4772-4780.
163. Chung Hoi, S., M. Louis John, and A. Eaton William, *Distinguishing between protein dynamics and dye photophysics in single-molecule FRET experiments.* Biophys J, 2010. **98**(4): p. 696-706.

164. Talaga, D.S., *Markov processes in single molecule fluorescence.* Curr. Opin. Colloid Interface Sci., 2007. **12**(6): p. 285-296.
165. Darugar, Q., et al., *Human T-Cell lympotropic virus type 1 nucleocapsid protein-induced structural changes in transactivation response DNA measured by single molecule fluorescence resonance energy transfer.* J. Virol., 2008. **92**(24): p. 12164-12171.
166. Taylor, J.N., et al., *Dynamics of an anti-VEGF aptamer: A single molecule study.* Biochem. Biophys. Res. Comm., 2008. **373**: p. 213-218.
167. Lau, A.Y. and B. Roux, *The Free Energy Landscapes Governing Conformational Changes in a Glutamate Receptor Ligand-Binding Domain.* Structure (Cambridge, MA, U. S.), 2007. **15**(10): p. 1203-1214.
168. Mamonova, T., M.J. Yonkunas, and M.G. Kurnikova, *Energetics of the Cleft Closing Transition and the Role of Electrostatic Interactions in Conformational Rearrangements of the Glutamate Receptor Ligand Binding Domain.* Biochemistry, 2008. **47**(42): p. 11077-11085.
169. McFeeters, R.L. and R.E. Oswald, *Structural mobility of the extracellular ligand-binding core of an ionotropic glutamate receptor. Analysis of NMR relaxation dynamics.* Biochemistry, 2002. **41**(33): p. 10472-10481.
170. Ahmed, A.H., et al., *Dynamics of the S1S2 glutamate binding domain of GluR2 measured using 19F NMR spectroscopy.* J. Biol. Chem., 2007. **282**(17): p. 12773-12784.
171. Fenwick, M.K. and R.E. Oswald, *On the Mechanisms of α-Amino-3-hydroxy-5-methylisoxazole-4-propionic Acid (AMPA) Receptor Binding to Glutamate and Kainate.* J. Biol. Chem., 2010. **285**(16): p. 12334-12343.
172. Cheng, Q., et al., *Evolution of glutamate interactions during binding to a glutamate receptor.* Nat. Chem. Biol., 2005. **1**(6): p. 329-332.
173. Robert, A. and J.R. Howe, *How AMPA receptor desensitization depends on receptor occupancy.* J. Neurosci., 2003. **23**(3): p. 847-858.
174. Zhang, W., et al., *Structural and single-channel results indicate that the rates of ligand binding domain closing and opening directly impact AMPA receptor gating.* J. Neurosci., 2008. **28**(4): p. 932-943.
175. Ha, T.J., et al., *Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism.* Proc. Natl. Acad. Sci. USA, 1999. **96**(3): p. 893-898.
176. Makarov, D.E., *Spatiotemporal correlations in denatured proteins: The dependence of fluorescence resonance energy transfer (FRET)-derived protein reconfiguration times on the location of the FRET probes.* The Journal of Chemical Physics, 2010. **132**(3): p. 035104-7.
177. Brunger, A.T., et al., *Three-dimensional molecular modeling with single molecule FRET.* Journal of Structural Biology, 2011. **173**(3): p. 497-505.
178. Freeman, G.S., D.M. Hinckley, and J.J. de Pablo, *A coarse-grain three-site-per-nucleotide model for DNA with explicit ions.* The Journal of Chemical Physics, 2011. **135**(16): p. 165104-13.
179. Monticelli, L., et al., *The MARTINI Coarse-Grained Force Field: Extension to Proteins.* Journal of Chemical Theory and Computation, 2008. **4**(5): p. 819-834.
180. Rossi, G., et al., *Coarse-graining polymers with the MARTINI force-field: polystyrene as a benchmark case.* Soft Matter, 2011. **7**(2): p. 698-708.
181. Tian, W.-d. and Y.-q. Ma, *Coarse-grained molecular simulation of interacting dendrimers.* Soft Matter, 2011. **7**(2): p. 500-505.

182. Hinner, M.J., S.-J. Marrink, and A.H. de Vries, *Location, Tilt, and Binding: A Molecular Dynamics Study of Voltage-Sensitive Dyes in Biomembranes.* The Journal of Physical Chemistry B, 2009. **113**(48): p. 15807-15819.
183. Marrink, S.J., et al., *The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations.* The Journal of Physical Chemistry B, 2007. **111**(27): p. 7812-7824.
184. Marrink, S.J., A.H. de Vries, and A.E. Mark, *Coarse Grained Model for Semiquantitative Lipid Simulations.* The Journal of Physical Chemistry B, 2003. **108**(2): p. 750-760.
185. López, C.A., et al., *Martini Coarse-Grained Force Field: Extension to Carbohydrates.* Journal of Chemical Theory and Computation, 2009. **5**(12): p. 3195-3210.
186. Wu, Z., Q. Cui, and A. Yethiraj, *A New Coarse-Grained Model for Water: The Importance of Electrostatic Interactions.* The Journal of Physical Chemistry B, 2010. **114**(32): p. 10524-10529.
187. Oostenbrink, C., et al., *A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6.* Journal of Computational Chemistry, 2004. **25**(13): p. 1656-1676.
188. Arora, S.K. and M. Sundaralingam, *The crystal and molecular structure of 4-methyl sulfonic acid (p-toluenesulfonic acid) monohydrate, C7H8SO3-.H3O+, an oxonium salt.* Acta Crystallographica Section B, 1971. **27**(7): p. 1293-1298.
189. Park, S., X. Zhu, and A. Yethiraj, *Atomistic Simulations of Dilute Polyelectrolyte Solutions.* The Journal of Physical Chemistry B, 2012. **116**(14): p. 4319-4327.