SINGLE-MOLECULE STUDIES OF THE DYNAMICS OF SSDNA NEAR A DNA (P/T) JUNCTION AND THEIR ROLE IN NUCLEIC ACID PROTEIN INTERACTIONS OF THE T4 BACTERIOPHAGE

by

BRETT A. ISRAELS

A DISSERTATION

Presented to the Department of Chemistry and Biochemistry
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2020

DISSERTATION APPROVAL PAGE

Student: Brett A. Israels

Title: Single Molecule Studies of ssDNA Dynamics Near a DNA (p/t) Junction and Their Role in Protein Nucleic Acid Interactions of the T4 Bacteriophage.

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Biology by:

Dr. Julia Widom	Chairperson
Dr. Andrew H. Marcus	Advisor
Dr. Peter H. von Hippel	Co-Advisor
Dr. Marina Guenza	Core Member
Dr. Tristen Ursell	Institutional Representative

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2020

DISSERTATION ABSTRACT

Brett A. Israels

Doctor of Philosophy

Department of Chemistry and Biochemistry

May 2020

Title: Single Molecule Studies of ssDNA Dynamics Near a DNA (p/t) Junction and Their Role in Protein Nucleic Acid Interactions of the T4 Bacteriophage.

Thermally induced conformational fluctuations of deoxyribonucleic acid (DNA) play an important role in the regulation of DNA replication, recombination and repair, which depends on the ability of protein machinery to recognize and bind to selected conformations of DNA lattices. Obtaining information about the nature of these functionally relevant DNA conformations, in addition to the time scales of their inter-conversion, is critical to understanding the detailed molecular mechanisms of protein-DNA interactions. An important component to DNA replication in all organisms is the single-stranded (ss) DNA binding protein (ssb), which binds to ssDNA, protecting and priming it for interaction with other proteins. We used the T4 Bacteriophage as a model organism to study the process of DNA replication, with a focus on the T4 ssb, gene-product32.

I investigated fundamental questions: 1. How does the molecular structure of ssDNA affect its conformational fluctuations? 2. How do gp32 proteins assemble onto ssDNA to form functional microscopic machines? And, 3. How is the gp32 assembly mechanism affected by ssDNA polarity and length? To answer these questions, I used a combination of spectroscopic techniques including sub-millisecond single-molecule Förster Resonance Energy Transfer (FRET) measurements. We analyzed our data by performing numerical optimizations of a transport master equation to simulate multi-order time correlation functions (TCFs) and the equilibrium distribution of conformational macrostates.

We discovered that the pathway for gp32 dimer assembly onto short oligonucleotides near ss—double-stranded (ds) DNA junctions proceeds through a transiently bound monomer, which does not slide along ssDNA. I developed methods to analyze the single-molecule signal at sub-millisecond resolution, which led to new insights into the nature of ssDNA fluctuations as well as the gp32 dimer assembly mechanism. I found that the ssDNA backbone fluctuates between various conformational states on a sub-millisecond timescale, only some of which are available for binding by gp32. By examining different lengths and polarities of ssDNA p/t constructs, I show that gp32 dimer sliding occurs on the timescale of a millisecond and that it is sensitive to the polarity of the ssDNA to which it is bound.

This dissertation includes both previously published and unpublished co-authored materials.

CURRICULUM VITAE

NAME OF AUTHOR: Brett A. Israels

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Florida State University, Tallahassee

DEGREES AWARDED:

Doctor of Philosophy, Chemistry, 2020, University of Oregon
Master of Chemistry 2014, University of Oregon
Bachelor of Science, Chemistry & Biochemistry, 2012, Florida State University
Bachelor of Science, Physical Science, 2012, Florida State University

AREAS OF SPECIAL INTEREST:

Spectroscopy with an emphasis on single-molecule microscopy
Protein-nucleic acid interactions
Network Theory and Markov analysis

PROFESSIONAL EXPERIENCE:

Graduate Student Researcher, Andrew Marcus laboratory, Department of Chemistry and Biochemistry, University of Oregon, 2013-2020.

Graduate Teaching Fellow, Department of Chemistry and Biochemistry, University of Oregon, 2012-2013.

Research Assistant, Per Arne Rikvold laboratory, Department of Physics, Florida State University, 2009-2012.

GRANTS, AWARDS, AND HONORS:

Predoctoral Molecular Biology and Biophysics Training Grant, University of Oregon, National Institutes of Health, 2013-2015.

Travel Award, Biophysical Society, 2019.

PUBLICATIONS:

Phelps, C., **B. Israels**, D. Jose, M.C. Marsh, P.H. Von Hippel, and A.H. Marcus. 2017. “Using Microsecond Single-Molecule FRET to Determine the Assembly Pathways of T4 SsDNA Binding Protein onto Model DNA Replication Forks.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (18). https://doi.org/10.1073/pnas.1619819114.

Phelps, C., **B. Israels,** M.C. Marsh, P.H. Von Hippel, and A.H. Marcus. 2016. “Using Multiorder Time-Correlation Functions (TCFs) to Elucidate Biomolecular Reaction Pathways from Microsecond Single-Molecule Fluorescence Experiments.” *Journal of Physical Chemistry B* 120 (51). https://doi.org/10.1021/acs.jpcb.6b08449.

Lee, Wonbae, John P. Gillies, Davis Jose, **Brett A. Israels,** Peter H. von Hippel, and Andrew H. Marcus. 2016. “Single-Molecule FRET Studies of the Cooperative and Non-Cooperative Binding Kinetics of the Bacteriophage T4 Single-Stranded DNA Binding Protein (Gp32) to SsDNA Lattices at Replication Fork Junctions.” *Nucleic Acids Research* 44 (22): 10691–710. https://doi.org/10.1093/nar/gkw863.

ACKNOWLEDGMENTS

I’d like to especially thank Professor Andrew Marcus for being a great role model for an aspiring scientist. On that same note, I’m extremely grateful to Professor Peter von Hippel for his endless supply of enthusiasm. I’d like to express gratitude to Dr. Larry Scatena and my fellow lab mates Dr. Carey Phelps, Claire Albrecht and Jack Maurer for their intellectual contributions to this work. I really appreciate the rest of my committee as well, Professors Marina Guenza, Julia Widom and Tristan Ursell for valuable feedback. None of this work could be done without funding, so thank you to the University of Oregon, the NIH, and the NSF. Thanks so much to all my friends and family who provided invaluable support.

I dedicate this dissertation to my mother Brenda Israels for her never-ending supply of love and encouragement, and to all the teachers and friendships that got me here.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

CHAPTER I

You can know the name of a bird in all the languages of the world, but when you're finished you'll know absolutely nothing whatever about the bird. So, let's look at the bird and see what it's doing – that's what counts.

- Richard P. Feynman

Chapter I serves as the introduction to the rest of the thesis. It outlines several of the major topics covered in subsequent chapters and is designed to give a non-expert the sufficient background to understand the rest of the material. Chapter II was published in *the Journal of Physical Chemistry B* in December 2016 with co-authors Carey Phelps, Morgan C. Marsh, Peter H. von Hippel, and Andrew H. Marcus. In this chapter we first outline the theory of Markov chains applied to a time correlation analysis of single-molecule data. Chapter III was published in *Proceedings of National Academy of Sciences* in April 2017 with co-authors Carey Phelps, Davis Jose, Morgan C. Marsh, Peter H. von Hippel, and Andrew H. Marcus. It describes the investigation into the assembly mechanism of a dimer of the T4 bacteriophage single-stranded DNA binding protein gp32 onto a 3′-p(dT)$_{15}$ ssDNA segment near a DNA p/t junction. Chapter IV is an unpublished collaboration with Claire Albrecht, Anson Dang, Megan Barney, Peter H. von Hippel, and Andrew H. Marcus. In this chapter we investigate the sub-millisecond kinetics of various DNA constructs that differ in length and polarity. Chapter V is an unpublished collaboration with Claire Albrecht, Anson Dang, Megan Barney, Peter H. von Hippel, and Andrew H. Marcus. It examines the polarity dependent binding mechanisms of gp32 using new insights from modeling ssDNA near a p/t junction, discussed in the previous chapter. Chapter VI is a concluding summary that puts all the work into context and ponders the next logical steps in the project for future investigators.

Molecular Biology of DNA

Deoxyribose nucleic acid (DNA) is a polymer like no other. The specific sequence and number of its monomers determine not only the dynamic conformational behavior of the polymer, but also the macroscopic physical makeup of its vehicle, roughly 6-8 orders of magnitude larger than itself. The full complement of DNA in an organism is known as the genome and is responsible for heredity, and the diversity of life on Earth. To further differentiate DNA from other polymers, we note that a single-strand (ss) of DNA in general doesn't exist as a lone polymer, but rather it is paired with another ssDNA to form the full double-strand (ds) DNA double-helix that is the heart of every living cell.

The monomer of ssDNA is a nucleotide, which consists of a sugar bound to a phosphate and nucleobase: adenine (A), guanine (G), thymine (T), or cytosine (C).[1] The nucleobases on one ssDNA of the double helix are bound to nucleobases on the other strand in a specific manner: the purine A binds to pyrimidine T with two hydrogen bonds, and the purine G binds to pyrimidine C with three hydrogen bonds. Pyrimidines are smaller than purines and participate in weaker $\pi - \pi$ stacking interactions between adjacent bases on the same ssDNA strand. All of the studies presented in this dissertation feature a ssDNA polymer composed of thymine nucleotides, abbreviated as $p(dT)_N$, where N is the number of thymine monomers.

DNA is a dynamic molecule, with motions on a vast range of length- and time-scales. One type of conformational fluctuation, 'dsDNA breathing', involves the complementary ssDNA strands transiently separated from one another.[2] DNA doesn't work alone: proteins take advantage of these transient fluctuations in order to interact with specific regions of DNA and carry out the myriad of biological functions it is essential to. The evolutionary history of an organism determines the array of proteins that are available to interact with the genetic code.

Single-Stranded DNA Binding Proteins

One class of protein that interacts with DNA is the single-stranded DNA binding protein (ssb). As the name suggest, ssbs bind to ssDNA, generally to the phosphate backbone. Long segments of single-stranded DNA are exposed during DNA replication, and shorter segments during DNA recombination, and repair. Ssbs bind to ssDNA and protect it from forming unfavorable secondary structures and from being attacked by nucleases. The ssb proteins often bind cooperatively to DNA, such that long segments of ssDNA can be coated.

The molecular version of Darwin's theory of natural selection goes as follows: if the sequence of the genetic code influences the success of the organism, and the success of the organism affects the proliferation of the genetic code, then it comes to reason that sequences of nucleotides which benefit the organism's survival will increase in frequency in the population relative to sequences that do not. If this is true macroscopically, it is certainly true within the cell too. Single-stranded DNA binding proteins protect the DNA that codes for it, skipping the middle man of the organism entirely. It is no wonder that such an interesting class of protein has evolved across virtually all domains of life.

The T4 bacteriophage is an excellent model organism in the field of molecular biology.[3] It's a particularly useful system to investigate DNA replication, owing in part to its well understood genome and similarities to eukaryotic DNA replication.[4] Only a handful of the 289 protein coding genes are necessary to form the multiprotein DNA replication machine known as the T4 *replisome.*[5] The replisome has the same basic replication components as humans including a helicase, a polymerase, a primase, and a ssb protein, gene-product 32 (gp32).

The T4 ssb gp32 will be the focus of much of this dissertation. Through studying the assembly mechanism of gp32, we not only learn specifics about an important component of the DNA replication system of a model organism, but also general principles about protein-nucleic acid interactions, how cooperativity between ligands affects higher-order assembly, and the role that conformational fluctuations play in complex biomolecular mechanisms.

Single Molecule FRET Spectroscopy

The word *single* in single-molecule spectroscopy means that the technique focuses on one *individual* molecule at a time. Because the signal from one molecule is not mixed with others as it is in traditional spectroscopic techniques, single-molecule spectroscopy is optimally sensitive to the moment-to-moment state of a molecule instead of an ensemble average value. The trade off one gets for more information content is a reduction in signal strength.

The FRET efficiency, E_{FRET}, is proportional to the distance, r, between two molecules called the donor (D) and acceptor (A) chromophore as described by Eq. (1.1).

$$E_{FRET} = \left[1 + \left(\frac{r}{R_0}\right)^6\right]^{-1}, \tag{1.1}$$

where R_0 is the donor-acceptor specific distance that the $E_{FRET} = 0.5$.[6] The relative orientation of the chromophores also affects the FRET efficiency, so the relationship between E_{FRET} and r is not exact. In smFRET, the donor molecule is excited by an external electric field (a narrowband laser tuned to a specific wavelength that corresponds to the donors S_0-S_1 transition). The excited-state Donor can either fluoresce or non-radiatively excite a nearby acceptor molecule, which can fluoresce in its stead. The closer the molecules are to one another, the more often the acceptor is excited. In fact, the relative fluorescence from the donor and acceptor molecules can be used to calculate E_{FRET} using Eq. (1.2),

$$E_{FRET} = \frac{I_A}{I_A+I_D}, \tag{1.2}$$

where I_A and I_D are the intensity per unit time of the acceptor and donor, respectively.

The experiments and analysis reported in this dissertation are about uncovering the details of how nature unwittingly coordinates complex molecular interactions. The specific sequence of biomolecular events that transpire from one initial state to a final state is called a mechanism. To understand biology, is to understand mechanisms.

Bridge to Chapter II

In the next chapter we introduce a generalized approach of using time-correlation functions (TCFs) to obtain kinetic information from single-molecule fluorescence measurements. The approach is illustrated in the context of two possible reaction mechanisms that each describe the assembly of gp32 monomers into cooperatively bound gp32 dimers on a ssDNA region near a DNA replication fork. DNA replication occurs at the rate of about a ms per nucleotide, and involves the coordination of many polymers (nucleic- and amino-acids) each with unique secondary structures and conformational dynamics. The TCF approach can yield information on timescales faster than concurrent hidden Markov modeling (HMM) methods, so is especially useful in microsecond-resolved spectroscopy to uncover information about biological mechanisms.

CHAPTER II

USING MULTI-ORDER TIME CORRELATION FUNCTIONS (TCFS) TO ELUCIDATE BIOMOLECULAR REACTION PATHWAYS FROM MICROSECOND SINGLE-MOLECULE FLUORESCENCE EXPERIMENTS

This work was published in *The Journal of Physical Chemistry B* (JPCB) in December 2016. I performed calculations featured in the manuscript and made some of the figures. I was involved in the drafting and editing of the paper along with coauthors Carey Phelps, Morgan Marsh, Andrew H. Marcus and Peter H. von Hippel. This work was supported by grants from the National Science Foundation (Chemistry of Life Processes Program grant CHE-1608915 to A.H.M.) and the National Institutes of Health (NIGMS Grant GM-15792 to P. H.v.H. and A.H.M.). Andrew H. Marcus was the principal investigator for this work.

Introduction

During the past several years, significant advances have been made in the use of single-molecule fluorescence methods to monitor conformational changes in the structure and dynamics of fluorescently labeled macromolecular systems. Such studies can provide detailed information about the assembly and function of protein-DNA complexes[1-9]. The recent development of sub-millisecond (tens-of-microseconds) single-molecule Förster resonance energy transfer (smFRET) experiments has opened the possibility to study relatively fast macromolecular processes, such as DNA ‘breathing’ and its role in the regulation of biochemical reactions,[2,10-11] which cannot be resolved on the time scales of most current single-molecule methods (~100 milliseconds). DNA breathing involves the thermal activation of segments of duplex DNA to form short-lived local ‘bubble-like’ states. Such locally disordered regions of DNA are thought to function as transient, secondary-structural motifs that can be bound by regulatory proteins as intermediate steps in the assembly and function of DNA-protein complexes. Microsecond-resolved smFRET experiments have the potential to reveal the mechanisms by which DNA-associated proteins can ‘harvest’ such specific thermally populated states in the course of carrying out reactions involved in the processes of genome expression.

Fast detection techniques, such as phase-synchronous single-photon-counting methods, can provide time-resolved data with tens-of-microsecond resolution.[2] Such experiments rapidly detect individual fluorescence photons from a single molecule, and store information about the intervening time intervals and optical phase conditions associated with each detection event. Even under optimal conditions, microsecond-resolved single-molecule fluorescence experiments produce 'sparse' data sets, because the average interval between successively detected signal photons can greatly exceed the experimental time resolution. In order to extract sub-millisecond kinetic information from sparse data sets, certain experimental challenges must be overcome. For example, transient intermediates may be difficult to detect due to the limited signal integration period. Under such low-signal conditions, the signal-to-noise (S/N) ratio is often too small to construct single-molecule trajectories in which transitions between distinct 'states' can be unambiguously identified and state-to-state transition 'pathways' can be visualized. Thus, the analysis of sparse trajectories must be carried out in non-standard ways.

In this paper, we show how mechanistic information can be obtained from microsecond single-molecule fluorescence experiments by applying generalized concepts of time correlation functions (TCFs).[12-21] TCFs provide a statistically meaningful way to characterize the time scales of stochastically fluctuating biochemical systems. Moreover, the time resolution of single molecule experiments can be maximized using TCFs, as demonstrated by Scherer and co-workers.[22] By correlating the fluctuations of individual molecules as a function of time, one can learn about the pathways connecting the conformational states that are accessible to the system at equilibrium. A commonly used approach to analyze single-molecule trajectories is to directly visualize the transition steps within a finite data set by fitting to a so-called hidden Markov model (HMM).[23] When utilized to their full advantage, TCFs constructed as a function of multiple time intervals can, in principle, provide more accurate and detailed information than HMM analyses.

In optimal situations, one can obtain several pieces of information from the analysis of single-molecule trajectories: (*i*) the number of conformational states reported by an experimental observation (such as a FRET measurement); (*ii*) the values of the observables associated with each state; and (*iii*) kinetic parameters associated with the inter-conversion between the states. When the experimental signal is especially noisy, as is the case for microsecond-resolved smFRET experiments, the application of HMM methods is inadequate to determine the above information. In contrast, TCFs provide an excellent approach to analyze the microsecond kinetics of macromolecular conformational transitions.

The situation can be described using the theory of Markov chains.[24] We assume that the instantaneous state of the system is mapped onto an experimentally accessible stochastic variable $A(t)$ that can be measured at discrete times. The distribution of A is characterized by its moments, and the time-dependent moments are the TCFs. In general, the nth-order TCF, $C^{(n)}(\tau_1, \tau_2, \dots, \tau_{n-1})$, can be written as the average product of n successive observations $\langle A(t_1)A(t_2) \dots A(t_n)\rangle$, which depends on the $n - 1$ time intervals $\tau_1 = t_2 - t_1$, $\tau_2 = t_3 - t_2$, …, $\tau_{n-1} = t_n - t_{n-1}$. The complexity of information that is potentially available from a TCF depends on its order. For example, the 2nd-order (two-point) TCF, $C^{(2)}(\tau) = \langle A(t_1)A(t_2)\rangle$, is the average product of two successive observations written as a function of the time interval $\tau = t_2 - t_1$. The 2nd-order TCF thus describes the average loss (or gain) in correlation of A over time, which can be used to obtain the average time scales of the fluctuations of the system. Nevertheless, 2nd-order TCFs do not provide information about 'transition pathways' – that is, whether a particular state-to-state transition must follow or precede another, or whether two such transitions occur independently. Such information is available through a higher-order TCF analysis. In the analysis that follows, we skip over 3rd-order TCFs and focus on the 4th-order (four-point) TCFs $C^{(4)}(\tau_1, \tau_2, \tau_3)$, because the latter contain more information and can handle reaction pathways that include a larger number of elementary steps. In principle, even higher-order TCFs (e.g., 5th-, 6th-order, etc.) could be employed, although this would require increasingly complex analyses that become more difficult due to the S/N limitations of finite data sets. We show, by performing a global analysis that includes 4th-order TCFs, that it is possible to characterize fundamental time scales of the system, including intervening (exchange) times that might be associated with short-lived chemical intermediates.

In addition, 4th-order TCFs are widely applied in molecular spectroscopy, such as two-dimensional (2D) NMR, 2D infra-red and 2D electronic spectroscopy.[25-26] For example, the nonlinear optical response of a molecule can be formulated in terms of the 4th-order TCFs of the appropriately defined transition dipole moment operator.[25] 4th-order TCFs have also been applied to study the stochastic microscopic fluctuations of complex chemical systems,[20-21] including protein reaction dynamics,[14] protein diffusion in solution,[15-16] liquid polymer diffusion,[17, 27] and protein conformation fluctuations in Molecular Dynamics (MD) simulations.[18]

In spite of their advantages, higher-order TCFs have not been previously used to study conformational transition pathways of biological macromolecules. This may be because the underlying concepts of TCFs are relatively abstract, and there are few sources on this topic that are accessible to a general scientific audience. Here we seek to demonstrate the utility of TCFs to extract mechanistic information from single-molecule fluorescence experiments. We show that by using TCFs of sufficiently high order, it is possible to distinguish between macromolecular binding pathways of varying levels of complexity.

In a recent study from our laboratory,[8] smFRET experiments were employed to analyze the cooperative binding of the single-stranded (ss) DNA binding protein of the T4 bacteriophage DNA replication complex (gp32) to single-stranded segments of primer-template (p/t) DNA constructs of varying lengths and polarities. These constructs can serve as models of DNA replication forks. Throughout this paper, we use a particular model experiment based on this study as an explicit molecular illustration of the principles and approaches developed in our analysis. As background, we note that gp32 protein molecules bind cooperatively and preferentially to ssDNA, with a binding site size of 7 nucleotide residues (nts, or DNA lattice positions) per gp32 molecule.[28] We have shown[8] that p/t DNA substrates with a ssDNA 'tail' region of 15 nts in length, which can cooperatively bind up to two gp32 proteins, can undergo stochastic fluctuations between 0-, 1- and 2-bound states (see Fig. 2.1A). In these experiments the ssDNA tail region was labeled on opposite ends with a FRET donor-acceptor chromophore pair that moves to longer inter-dye distances as gp32 molecules bind between them and thus increase the rigidity of the intervening ssDNA sequence. As a consequence, the sequential binding of gp32 molecules to the ssDNA tail can be monitored by tracking the changes in the FRET signal, as discussed further below.

While such experiments could detect the presence of distinct conformational sub-states of the ssDNA involved in association / dissociation events, the time-resolution of the experiments described in Lee *et al.*[8] (~100 ms) was not sufficient to determine either the lifetimes of the short-lived singly-bound intermediates, or to directly observe their conversions to longer-lived end-states. Nevertheless, this model system can serve as a concrete illustration of the potential uses of the theoretical approaches developed here. We are currently applying these TCF methods to analyze new microsecond-resolved single-molecule experiments on this gp32 binding system.

Conformational Transition Pathways and the Role of Intermediates

We consider an equilibrium system composed of N discrete microscopic states. At any instant, the system can undergo a transition from state-i to state-j where $i, j \in \{0, 1, \ldots, N-1\}$. We assume that there exists an experimentally accessible stochastic variable $A(t)$ that is coupled to the conformation of the system. For example, A might be a fluorescence signal from a single fluorophore or a collective signal from a FRET donor-acceptor pair that site-specifically labels a biological macromolecular complex and is sensitive to its local conformation or to a similar reaction coordinate. When the system occupies state-i, the variable A assumes a corresponding value A_i.

As indicated above, we illustrate our approach using the macromolecular system studied by Lee *et al.*,[8] in which a ssDNA template interacts with the T4 bacteriophage gp32 binding protein (see Fig. 2.1A, in Appendix A with all tables). The $N = 3$ reaction scheme (shown in Fig. 2.1B) is the simplest possible to describe the p(dT)$_{15}$-(gp32)$_n$ system (with $n = 0$, 1, or 2), which involves 0-, 1- and 2-bound gp32 molecule states. Since the gp32 protein occludes 7 nts on the ssDNA template, there are nine possible binding conformations available to the 1-bound state (e.g., at positions 1 – 7, 2 – 8, …, and 9 – 15). This simplest model treats all 1-bound states as experimentally indistinguishable species that may lie on the accessible pathway connecting the reactant 0-bound state to the product 2-bound state. In this reaction scheme, we do not indicate direct transitions between 0- and 2-bound states, since it is known that gp32 does not directly bind to ssDNA as a dimer.[29]

Despite the appealing simplicity of the $N = 3$ scheme (Fig. 1B) for the p(dT)$_{15}$-(gp32)$_n$ system, further consideration suggests that this mechanism cannot provide an adequate description of this gp32-binding model system because all of the 1-bound states on the 15 nts ssDNA 'tail' lattice cannot be treated as identical. Rather there are a number of ways in which a gp32 monomer might initially bind to the ssDNA template that would partially occlude the second binding site of 7 contiguous unoccupied nts, which is required to allow a second gp32 monomer to bind to the ssDNA tail of the p/t construct.[30] Such 1-bound states that 'overlap' the potential second binding site represent 'unproductive' intermediates, and thus inhibit transitions between the 0-bound and 2-bound states. Clearly, the first gp32 protein can bind productively only at the four possible positions (1 – 7, 2 – 8, 8 – 14 or 9 – 15) to allow the ssDNA 'tail' sequence to retain a contiguous (7 nts) binding site that can accommodate a second gp32 monomer.[31] These latter 1-bound states would function as 'productive' intermediates through which the 0-bound state can undergo transitions to the 2-bound states. The kinetics of a model of this type can be diagramed using the $N = 4$ scheme shown in Fig. 2.1C, in which we have labeled the 'unproductive' and 'productive' intermediates as state-1 and state-1', respectively.

As pointed out above, the binding states of the ssDNA-(gp32)$_n$ system and their inter-conversion pathways can be studied using smFRET techniques.[8] In the experiments by Lee *et al.*,[8] which were performed using 100-ms time resolution, only two states – a 0-bound state and a 2-bound state – could be unambiguously observed, although indirect evidence for the existence of short-lived 1-bound states was also obtained. These results suggested that 1-bound states are present, but are too short-lived to be resolved in experiments conducted at 100-ms resolution. Because gp32 binding to ssDNA is known to be highly cooperative, 1-bound states are expected to be unstable in comparison to 2-bound states. A reasonable model for the assembly mechanism of the system might involve an initial singly bound gp32 molecule that either rapidly recruits a second gp32 protein to the ssDNA lattice to form a high affinity (cooperatively bound) dimer of gp32 molecules, or that rapidly dissociates from the ssDNA lattice. The relative probabilities of these competing scenarios should depend in part on the location of the initially bound gp32 protein, as described by the four-state scheme of Fig. 2.1C. Indeed, a common situation for many single-molecule experiments is that intermediates can be very short-lived, and their observed signals might be degenerate. An idealized stochastic smFRET trajectory for the $N = 3$ scheme is shown in Fig. 2.1D, in which case A_0, A_1 and A_2 are the values of the observable $A(t)$ when the system is in states 0, 1 or 2, respectively.

To fully appreciate the kinetics of the ssDNA-(gp32)*n* system, one must properly account for the short-lived 1-bound intermediates, which may well give rise to indistinguishable signals. Experimentally, this requires making measurements at a higher time resolution than that used in the Lee *et al.* study.[8] As the time resolution of a single-molecule fluorescence measurement approaches a few milliseconds, the signal will necessarily become too noisy to extract the state of the system through direct visualization of single-molecule trajectory data (e.g., by HMM analysis). Rather, we show below how equivalent information may be obtained through the application of the generalized concepts of TCFs.

Definitions of 2nd- and 4th-Order Time Correlation Functions

The 2nd-order TCF of A is the average product of two successive measurements, made at times t_1 and t_2, which are separated by the interval $\tau = t_2 - t_1$

$$C^{(2)}(\tau) = \langle A(0)A(\tau)\rangle. \tag{2.1}$$

In Eq. (2.1), the angle brackets denote that the average has been performed over all possible starting times, according to $C^{(2)}(\tau) = \int_{-\infty}^{\infty} A(t)A(t+\tau)dt$. If the longest relaxation time of the system exceeds the duration of an individual data set, then the average two-point product is additionally integrated over a large number of single-molecule data sets. For a stochastic chemical system, $C^{(2)}(\tau)$ decays from its maximum value $\langle A^2\rangle$ at $\tau = 0$ to its asymptotic minimum $\langle A\rangle^2$ in the limit $\tau \rightarrow \infty$. For this reason, we define the fluctuation $\delta A(t) = A(t) - \langle A\rangle$, and its TCF:

$$\bar{C}^{(2)}(\tau) = \langle \delta A(0)\delta A(\tau)\rangle = \langle A(0)A(\tau)\rangle - \langle A\rangle^2 \tag{2.2}$$

The TCF $\bar{C}^{(2)}(\tau)$ defined by Eq. (2.2) decays from its maximum $\langle \delta A^2\rangle$ to zero over the characteristic time scales of the system.

One can predict the form of $\bar{C}^{(2)}(\tau)$ for a given model using the theory of Markov chains, which assumes that the time interval between successive observations is long in comparison to 'internal relaxation times,' and that the probability that the system undergoes a transition from state-i to state-j depends only on its occupancy of state-i.[24] This assumption ignores the possibility of memory effects, which become important if internal barriers associated with state-i influence the transition probability. The Markov chain expression for the 2nd-order TCF is:

$$\bar{C}^{(2)}(\tau) = \sum_{i,j=0}^{N-1} \delta A_j p_{ji}(\tau) \delta A_i p_i^{eq} \tag{2.3}$$

In Eq. (2.3), p_i^{eq} is the equilibrium (time-independent) probability to observe the system in state-i, δA_i is the value of the fluctuation observable associated with that state, and $p_{ji}(\tau)$ is the conditional probability that the system will be in state-j at a time τ after it was initially observed to be in state-i. Equation (3) shows that the 2nd-order TCF is the second moment of the time-dependent stochastic variable $\delta A(t)$, which is the weighted average of all possible two-point products $\delta A_j \delta A_i$ occurring within the time interval τ. It is instructive to note that when τ is short in comparison to the shortest transition time of the system, the two-point product is dominated by terms $\delta A_i \delta A_i$, such that $\bar{C}^{(2)}(\tau \to 0) = \sum_{i=0}^{N-1} \delta A_i^2 p_i^{eq} = \langle \delta A^2 \rangle$. In contrast, for τ longer than the longest transition time, the two-point product is dominated by uncorrelated successive observations, such that $\bar{C}^{(2)}(\tau \to \infty) = \left[\sum_{j=0}^{N-1} \delta A_j p_j^{eq}\right] \times \left[\sum_{i=0}^{N-1} \delta A_i p_i^{eq}\right] = 0$. When the time interval τ is comparable to the time scale of a particular transition from state-i to state-j, the two-point product is dominated by terms $\delta A_j \delta A_i$, which reflect the weighted contributions of these particular transitions.

The information provided by the 2nd-order TCF alone cannot be used to determine whether the states visited during a single-molecule trajectory occur independently, or are connected through a 'pathway' of correlated sequential events. One can imagine that a particular fluctuation must occur first in order for a subsequent fluctuation to follow. For example, the $N = 3$ and $N = 4$ schemes depicted in Fig. 2.1B and Fig. 2.1C, respectively, illustrates the ssDNA-(gp32)$_n$ assembly pathways as a system of coupled elementary chemical steps in which the 0-bound and 2-bound states are inter-connected through 'productive' (and sometimes 'unproductive') intermediates. The 2nd-order TCF does not contain information, for example, about how a transition between any particular 1-bound and 2-bound state might be correlated to a preceding transition between the 0-bound and a 1-bound state. As we shall see, information about the preferred sequences of transitions that occur at equilibrium is contained in 'higher-order' TCFs.

To distinguish between different mechanisms of coupled chemical transformations, we consider the information contained within 4th-order TCFs. The 4th-order TCF of δA is the average product of four sequential observations, separated by the three time intervals $\tau_1 = t_2 - t_1$, $\tau_2 = t_3 - t_2$, and $\tau_3 = t_4 - t_3$ (see Fig. 2.1D)

$$C^{(4)}(\tau_1, \tau_2, \tau_3) = \langle \delta A(0) \delta A(\tau_1) \delta A(\tau_2) \delta A(\tau_3) \rangle \tag{2.4}$$

In Eq. (2.4), the angle brackets have the same meaning as those in Eqs. (2.1) and (2.2). The 4th-order TCF $C^{(4)}(\tau_1, \tau_2, \tau_3)$ depends on the probability of sampling each possible time-ordered sequence of δA. For the $N = 4$ scheme of Fig. 2.1C, for example, we might observe the sequence $\delta A_0 \delta A_0 \delta A_{1'} \delta A_2$ at the four times sampled. If, for a particular set of time intervals, we were to observe this sequence with greater frequency than sequences that contain sequential occurrences of δA_0 followed by δA_2, then we might conclude that direct transitions between state-0 and state-2 are unlikely, and must proceed through an intermediate state-1'. Because the timescales of transitions between the various states have definite values, certain sequences will be more prevalent for short time intervals, while others will occur with greater frequency for long time intervals. Thus, the information encoded in $C^{(4)}(\tau_1, \tau_2, \tau_3)$ provides direct insight into the kinetic scheme that defines the time-ordered fluctuations of a single-molecule trajectory.

It is helpful to visualize $C^{(4)}(\tau_1, \tau_2, \tau_3)$ as a series of two-dimensional (2D) contour plots, with horizontal and vertical axes given by the intervals τ_1 and τ_3. We present model calculations of $C^{(4)}(\tau_1, \tau_2, \tau_3)$ in the next section. Such plots are presented as a parametric function of the interval τ_2, which is referred to as the waiting time. As mentioned above, the 4th-order TCF contains information about the presence of 'higher-order temporal correlations' between successive transitions, with the first transition occurring during τ_1 and the second during τ_3. By examining a series of 4th-order TCFs as a function of τ_2, we can determine the average timescales over which successive transitions are correlated. In the absence of higher-order correlations, upstream and downstream transitions occur independently. In the limit that the waiting time τ_2 becomes very long, or that higher-order correlations are short-lived, we see from Eq. (2.4) that $\lim_{\tau_2\to\infty} C^{(4)}(\tau_1,\tau_2,\tau_3) = \langle\delta A(0)\delta A(\tau_1)\rangle\langle\delta A(0)\delta A(\tau_3)\rangle = \bar{C}^{(2)}(\tau_1)\bar{C}^{(2)}(\tau_3)$. In this limit, the 4th-order TCF is equal to the square product of the 2nd-order TCF defined in Eq. (2.2). To isolate the effects of higher order correlations from those due to 2nd-order 'background' correlations, it is useful to define the 4th-order *difference* TCF

$$\bar{C}^{(4)}(\tau_1,\tau_2,\tau_3) = \langle\delta A(0)\delta A(\tau_1)\delta A(\tau_2)\delta A(\tau_3)\rangle - \bar{C}^{(2)}(\tau_1)\bar{C}^{(2)}(\tau_3) \qquad (2.5)$$

The 4th-order difference TCF $\bar{C}^{(4)}(\tau_1,\tau_2,\tau_3)$ defined by Eq. (2.5) decays as a function of τ_2 from its maximum value $\langle\delta A(0)[\delta A(\tau_1)]^2\delta A(\tau_3)\rangle$ to zero over the characteristic time scales for which higher-order correlations exist.

The Markov chain expression for the 4th-order TCF can be written

$$\langle\delta A(0)\delta A(\tau_1)\delta A(\tau_2)\delta A(\tau_3)\rangle \qquad (2.6)$$
$$= \sum_{i,j,k,l=0}^{N-1} \delta A_l p_{lk}(\tau_3)\delta A_k p_{kj}(\tau_2)\delta A_j p_{ji}(\tau_1)\delta A_i p_i^{eq}$$

where the conditional probability $p_{ji}(\tau)$ is defined similarly as in Eq. (2.3). Since the system may only occupy discrete states, the 4th-order TCF is the weighted sum of a finite number of four-point products $\delta A_l\delta A_k\delta A_j\delta A_i$. For the $N = 3$ example of Fig. 2.1B, each observation can take only one of three possible values: δA_0, δA_1 or δA_2. Thus for $N = 3$, the four-point product can acquire (3)(3)(3)(3) = 81 possible outcomes (or pathways). In general, the number of possible outcomes for an N-state system is N^4, and the 4th-order TCF is composed of the weighted average of these outcomes as described by Eq. (2.6). In order to apply Eqs. (2.3) and (2.6) to a specific N-state system, one must solve for the conditional probabilities $p_{ji}(\tau)$. In the following sections, we show how the conditional probabilities may be obtained as the formal solution to a master equation for a system of N coupled differential equations that characterize the reaction pathway.

Calculation of TCFs using Markov Chains

We apply the theory of Markov chains to relate the 2nd- and 4th-order TCFs defined in the previous sections to specific N-state models.[24,32] Such analyses are generally useful for the interpretation of single-molecule trajectories in which stochastic transitions occur between a few discrete states. We write the memory-less master equation for an *N*-state system

$$\dot{\boldsymbol{p}}(t) = \boldsymbol{K}\boldsymbol{p}(t) \equiv \begin{bmatrix} \dot{p}_0 \\ \dot{p}_1 \\ \vdots \\ \dot{p}_{N-1} \end{bmatrix} \tag{2.7}$$

$$= \begin{bmatrix} -\sum_{i=0}^{N-1} k_{0,i} & k_{1,0} & \cdots & k_{N-1,0} \\ k_{0,1} & -\sum_{i=0}^{N-1} k_{1,i} & \ddots & \vdots \\ \vdots & \ddots & \ddots & k_{N-1,N-2} \\ k_{0,N-1} & \cdots & k_{N-2,N-1} & -\sum_{i=0}^{N-1} k_{N-1,i} \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ \vdots \\ p_{N-1} \end{bmatrix}$$

In Eq. (2.7), $\boldsymbol{p}(t)$ is an *N*-dimensional vector containing the probabilities to find the system in each of its *N* states at time *t*, and $\boldsymbol{K}$ is the *N*×*N* rate matrix, with elements k_{ij} associated with the transitions from state-i to state-j. We constrain the diagonal elements of the rate matrix $k_{ii} = -\sum_{j\neq i}^{N-1} k_{ij}$ to enforce the mass action law, and we set the sum of the instantaneous state probabilities $\sum_{i=0}^{N-1} p_i\ (t) = 1$.

When constructing the rate matrix $\boldsymbol{K}$, the elements k_{ij} must be chosen to satisfy the detailed balance condition, $p_i^{eq} k_{ij} = p_j^{eq} k_{ji}$ where $p_i^{eq} = lim_{t\to\infty} p_i(t)$ is the stationary (equilibrium) occupancy of state-i. The detailed balance condition requires that in the long-time limit, the flow of probability from state-i to state-j is equal to the flow of probability from state-j to state-i. For coupled reactions that involve cyclical pathways, the requirements of the detailed balance condition lead to additional inter-dependencies of the rate constant matrix elements. In Fig. 2.2, we depict three reaction schemes as examples to illustrate this point. For a system that contains a single cyclical pathway (Fig. 2.2A), the product of rate constants moving along the clockwise path must equal the product of rate constants moving along the counter-clockwise path; i.e. $k_{03}k_{32}k_{21}k_{10} = k_{30}k_{01}k_{12}k_{23}$. Thus, a system that contains a single cyclical pathway leads to the constraint that one rate constant must depend on all others. This relationship ensures that the flow of probability in the clockwise direction is precisely balanced by the flow of occupancies in the counter-clockwise direction, as must be the case for an equilibrium system. In the absence of a cyclical pathway, the detailed balance condition can be satisfied locally for each successive step of the coupled chemical reaction (see Fig. 2.2B), so that the rate constants may be chosen independently of each other. When the system contains multiple cyclical pathways, such as the situation depicted in Fig. 2.2C, more complicated interrelationships between rate constants exist. The relationship between cyclical pathways in this instance leads to the requirement that two rate constants must be dependent on all others. An excellent description of enforcing detailed balance can be found in reference 33.

Provided that a rate matrix $\boldsymbol{K}$ can be found to satisfy the detailed balance condition, a general solution of Eq. (2.7) can be obtained using the spectral decomposition method[24]

$$\boldsymbol{p}(t) \equiv \sum_{i=0}^{N-1} c_i \boldsymbol{v}_i e^{-\lambda_i t} = c_0 \boldsymbol{v}_0 e^{-\lambda_0 t} + c_1 \boldsymbol{v}_1 e^{-\lambda_1 t} + \cdots + c_{N-1} \boldsymbol{v}_{N-1} e^{-\lambda_{N-1} t} \tag{2.8}$$

In Eq. (2.8), λ_i and $\boldsymbol{v}_i$ are, respectively, the eigenvalues and the corresponding eigenvectors of the rate matrix of Eq. (2.7). We set the first eigenvalue $\lambda_0 = 0$ to allow the time-dependent populations to decay to the constant equilibrium distribution $c_0 \boldsymbol{v}_0 = \boldsymbol{p}^{eq}$. We may thus rewrite Eq. (2.8) explicitly in terms of the equilibrium distribution

$$\boldsymbol{p}(t) = \boldsymbol{p}^{eq} + c_1 \boldsymbol{v}_1 e^{-\lambda_1 t} + \cdots + c_{N-1} \boldsymbol{v}_{N-1} e^{-\lambda_{N-1} t} \tag{2.9}$$

The conditional probabilities $p_{ji}(\tau)$ needed for the evaluation of 2nd-order and 4th-order TCFs described by Eqs. (2.3) and (2.6) respectively, can be obtained using Eq. (2.9), with proper enforcement of the boundary conditions. For example, $p_{21}(\tau)$ is the conditional probability that the system resides in state-2 at time τ, given that it was in state-1 at time zero. In this case, the initial condition is $p_1(0) = 1$ and $p_{i\neq 1}(0) = 0$. We may thus solve Eq. (2.9) for the set of expansion coefficients $\{c_1, c_2, \cdots, c_{N-1}\}$, and for the conditional probability $p_{21}(\tau)$. We carry out a similar procedure for each conditional probability $p_{ji}(\tau)$ with $i, j \in \{0,1, \cdots, N-1\}$.

Analytical expressions for the 2nd- and 4th-order TCFs for N = 2 and N = 3.

We next consider analytical expressions for the 2nd and 4th-order TCFs that follow from Eq. (2.9) for common situations with $N = 2$ and $N = 3$. Although the expressions for $N = 2$ systems are trivial, we include them for completeness before examining the more complex situations with $N = 3$.

Two-state system. For an $N = 2$ scheme, the 2nd-order TCF described by Eq. (2.2) is a weighted average of 4 possible two-point product pathways, as shown schematically in Fig. 2.3A.

The master equation solution [Eq. (2.9)] specified for $N = 2$ yields the time-dependent conditional probabilities

$$p_{ji}(\tau) = p_j^{eq} + \left[p_j(0) - p_j^{eq}\right]e^{-\lambda_1\tau}, \quad N = 2 \tag{2.10}$$

where $\lambda_1 = k_{12} + k_{21}$ is the only non-zero eigenvalue. An analytical expression for the 2nd-order TCF follows from substitution of Eq. (2.10) into Eq. (2.3).

$$\bar{C}^{(2)}(\tau) = \langle \delta A^2 \rangle e^{-\lambda_1\tau}, \quad N = 2 \tag{2.11}$$

Equation (11) shows that the 2nd-order TCF for a two-state system decays exponentially with rate constant $\lambda_1 = k_{12} + k_{21}$.

For the $N = 2$ scheme, the 4th-order TCF described by Eq. (2.4) is a weighted average of 16 possible four-point product pathways, as shown schematically in Fig. 2.3B. Upon substitution of Eq. (2.10) into Eq. (2.6), it is straightforward to show that the 4th-order TCF for a two-state system has the form

$$\langle \delta A(0)\delta A(\tau_1)\delta A(\tau_2)\delta A(\tau_3)\rangle = \mathcal{A}_{11}e^{-\lambda_1(\tau_1+\tau_3)}, \quad N = 2 \tag{2.12}$$

where the constant $\mathcal{A}_{11} = \langle \delta A^2 \rangle^2$. Equation (12) shows that the 4th-order TCF for an $N = 2$ system is simply the product of the 2nd-order TCFs $\bar{C}^{(2)}(\tau_1)\bar{C}^{(2)}(\tau_3)$ for all values of τ_2. This follows since there are no intermediates in an $N = 2$ scheme, and therefore no 'higher-order' transition pathways can exist. In this case, the 4th-order difference TCF $\bar{C}^{(4)}(\tau_1, \tau_2, \tau_3)$, defined by Eq. (2.5), is equal to zero for all values of τ_2.

Three-state system. We next consider the three-state scheme ($N = 3$) introduced in Fig. 2.1B, and redrawn for the following discussion in Fig. 2.4. In the redrawn scheme, we have allowed for the hypothetical transition between the 0-bound (reactant) state and the

2-bound (product) state, so that these might (or might not) be bridged by a 1-bound (intermediate) state. The 0 ⇄ 2 reaction pathway would require the binding of an appropriately pre-formed gp32 dimer directly from solution. This does not happen in the real system, but we include the possibility here to provide generality. Such schemes are the simplest that may exhibit higher-order temporal correlations, as reflected by the behavior of the 4th-order TCF. The derivations of the corresponding analytical expressions are straightforward, yet somewhat involved. We present the derivation here to illustrate how higher-order correlations emerge.

The master equation for an $N = 3$ system is specified, using Eq. (2.7), according to:

$$\begin{bmatrix} \dot{p}_0 \\ \dot{p}_1 \\ \dot{p}_2 \end{bmatrix} = \begin{bmatrix} -k_{01} - k_{02} & k_{10} & k_{20} \\ k_{01} & -k_{10} - k_{12} & k_{21} \\ k_{02} & k_{12} & -k_{20} - k_{21} \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ p_2 \end{bmatrix} \qquad (2.13)$$

The general solution to Eq. (2.13) is

$$\boldsymbol{p}(t) = \boldsymbol{p}^{eq} + c_1 \boldsymbol{v}_1 e^{-\lambda_1 t} + c_2 \boldsymbol{v}_2 e^{-\lambda_2 t}, \qquad N = 3 \qquad (2.14)$$

where the eigenvalues λ_1 and λ_2 and the eigenvectors $\boldsymbol{v}_1 = [v_1^0, v_1^1, v_1^2]$ and $\boldsymbol{v}_2 = [v_2^0, v_2^1, v_2^2]$ are functions of the rate constants (derivation given in Appendix C, Chapter III). To satisfy detailed balance, one rate constant must depend on the others, such that $k_{20} = k_{02}k_{21}k_{10}/k_{12}k_{01}$. The equilibrium populations $\boldsymbol{p}^{eq} = \left[p_0^{eq}, p_1^{eq}, p_2^{eq}\right]$ are found by solving Eq. (2.13) with the boundary condition $\dot{\boldsymbol{p}}(t) = 0$. These solutions must also satisfy completeness: $\sum_{i=0}^{2} p_i\,(t) = 1$. The above conditions lead to explicit forms for the component equilibrium populations p_0^{eq}, p_1^{eq} and p_2^{eq}, which are explicit functions of the rate constants (see Appendix C, Chapter III).

To determine the nine conditional probabilities $p_{ji}(\tau)$ with $i, j \in \{0,1,2\}$, we solve Eq. (2.14) for the expansion coefficients c_1 and c_2, while assuming the appropriate boundary conditions. We label each expansion coefficient with a superscript to indicate the boundary condition. For example, the expansion coefficient c_1^0 corresponds to the case when all population resides in state-0 at time zero, i.e. $p_0(0) = 1$ and $p_1(0) = p_2(0) = 0$. This leads to closed form expressions for the six expansion coefficients: c_2^0, c_2^1, c_2^2, c_1^0, c_1^1 and c_1^2 (see Appendix C, Chapter III). Upon substitution of these into Eq. (2.14), we obtain the conditional probabilities

$$p_{ji}(\tau) = p_j^{eq} + c_1^i v_1^j e^{-\lambda_1 \tau} + c_2^i v_2^j e^{-\lambda_2 \tau}, \quad N = 3 \tag{2.15}$$

Substitution of Eq. (2.15) into Eqs. (2.3) and (2.6) provides analytical expressions for the 2nd- and 4th-order TCFs, respectively. Although these expressions are unwieldy to write in extended form, their solutions are readily obtained using a desktop computer. The 2nd-order TCF can be written succinctly

$$\bar{C}^{(2)}(\tau) = \mathcal{A}_1 e^{-\lambda_1 \tau} + \mathcal{A}_2 e^{-\lambda_2 \tau}, \quad N = 3 \tag{2.16}$$

Equation (16) is composed of two exponentially decaying terms, with decay rates λ_1 and λ_2 and amplitudes $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. The constants λ_1, λ_2, $\mathcal{A}_1$ and $\mathcal{A}_2$ are polynomial functions of the six rate constants k_{ij}, with $i, j \in \{0,1,2\}$ and $i \neq j$.

It is straightforward to show that the difference 4th-order TCF, which is given by Eq. (2.5), has the succinct form

$$\bar{C}^{(4)}(\tau_1, \tau_3)]_{\tau_2\, fixed} = \mathcal{A}_{11}(\tau_2) e^{-\lambda_1(\tau_1+\tau_3)} + \mathcal{A}_{12}(\tau_2) e^{-\lambda_1\tau_1 - \lambda_2\tau_3} \tag{2.17}$$

$$+\mathcal{A}_{21}(\tau_2) e^{-\lambda_2\tau_1 - \lambda_1\tau_3} + \mathcal{A}_{22}(\tau_2) e^{-\lambda_2(\tau_1+\tau_3)}, \qquad N = 3$$

Equation (17) is composed of four terms, each with an amplitude $\mathcal{A}_{mn}$ $[n, m \in \{1,2\}]$ that depends on the waiting time τ_2. Similar to the 2nd-order TCF, the 4th-order TCF decays exponentially. For a fixed waiting time τ_2, the decay of the 4th-order TCF occurs in two dimensions, corresponding to the time intervals τ_1 and τ_3. The characteristic decay rates of the 4th-order TCF are the same as those of the 2nd-order TCF. In Eq. (2.17), the two terms with amplitudes $\mathcal{A}_{11}$ and $\mathcal{A}_{22}$ designate global relaxation self-terms (i.e. terms that each depend on a single eigenvalue, λ_1 *or* λ_2, respectively), while the terms with amplitudes $\mathcal{A}_{12}$ and $\mathcal{A}_{21}$ designate inter-dependent cross-terms, which each depend on both decay constants, λ_1 *and* λ_2. For an equilibrium system, the detailed balance condition requires that $\mathcal{A}_{12} = \mathcal{A}_{21}$.[19] As we discuss further below, the self-term amplitudes, $\mathcal{A}_{11}$ and $\mathcal{A}_{22}$, indicate the relative weights of the global relaxation processes, while the sign and magnitude of the cross-term amplitudes, $\mathcal{A}_{12}$ and $\mathcal{A}_{21}$, indicate positive or negative 4th-order correlations that effectively couple these processes.

We now return to the example of the ssDNA-(gp32)$_2$ assembly reaction, as depicted in Fig. 2.4. To illustrate how the local connectivity between states can affect the collective dynamics characterized by the 4th-order TCF, we present in Fig. 2.5A – 2.5D calculations for a specific case in which the rate constants k_{12} and k_{21} are varied while the remaining parameters are held fixed. For the purpose of this discussion, we have set the waiting time interval τ_2 = 1 ms, and we have chosen plausible values for the rate constants $k_{01} = 10$ s^{-1}, k_{10} = 20 s^{-1}, k_{02} = 2 s^{-1}, and k_{20} = 4 s^{-1} with signal observables A_0 = 0.9, A_1 = 0.3, and A_2 = 0.1. This particular choice of parameters assumes that the time scales of exchange between reactant state-0 and intermediate state-1 are much faster than those between reactant and product state-2. It is worth noting that for time intervals in which four-point pathways are dominated by recurring observations of the end state-0 or state-2 (e.g., $\delta A_0 \delta A_0 \delta A_0 \delta A_0$), the 4th-order TCF will tend to be high-valued. Alternatively, for intervals in which the majority of four-point pathways include observations of the intermediate state-1 (e.g., $\delta A_0 \delta A_1 \delta A_1 \delta A_2$), the 4th-order TCF will tend to be low-valued. For this particular example with the given rates under the detailed balance condition, the symmetry of the system dictates that for all values of the rate constants k_{12} = k_{21}, the equilibrium distribution of populations are given by p_0^{eq} = 0.5, p_1^{eq} = 0.25, and p_2^{eq} = 0.25.

We initially consider the case in which transitions between state-1 and state-2 are prohibitively slow (i.e., k_{12} = 0). The time scales of the local elementary chemical reaction steps 0 ⇄ 1 and 0 ⇄ 2 can be estimated by assuming that these transitions occur independently of one another. We thus estimate the time scale of 'fast' transitions between state-0 and state-1 as $(k_{01} + k_{10})^{-1}$ = 33 ms, and that of 'slow' transitions between state-0 and state-2 as $(k_{02} + k_{20})^{-1}$ = 167 ms. By solving the master equation for the coupled system [Eq. (2.13)], we determine the time scales of the global relaxations (eigenvalues) λ_1 = 31 s^{-1} and λ_2 = 5.2 s^{-1}, which correspond to the times λ_1^{-1} = 32 ms and λ_2^{-1} = 193 ms, respectively. Because in this example there is a clear separation between fast and slow elementary chemical steps (i.e. 0 ⇄ 1 and 0 ⇄ 2), these time scales closely approximate those of the eigenvalues of the coupled system (1 ⇄ 0 ⇄ 2). In Fig. 2.5A, we plot the 4th-order TCF corresponding to these conditions. We note that this function slowly rises to a peak value close to the point τ_1 = τ_3 ~ 33 ms and then gradually decays to zero with increasing values of τ_1 and τ_3. This behavior reflects the fact that multi-step transitions occur only rarely on time scales shorter than the fastest exchange process of the system. For values of τ_1 and τ_3 that match the time scale of the fast 0 ⇄ 1 exchange process, the 4th-order TCF is heavily weighted by terms that involve successive observations of the reactant and intermediate states (e.g., $\delta A_0 \delta A_1 \delta A_1 \delta A_0$). For values of τ_1 and τ_3 in which one or the other of these intervals approaches time scales comparable to the slow 0 ⇄ 2 exchange process, the 4th-order TCF is composed mostly of terms that include successive observations of all three states involved in both fast and slow local reactions (e.g., $\delta A_1 \delta A_0 \delta A_0 \delta A_2$), which in turn cause the function to decay. The self- and cross-term amplitudes corresponding to these conditions are $\mathcal{A}_{11}$ = 1.08, $\mathcal{A}_{22}$ = 6.73, and $\mathcal{A}_{12}$ = $\mathcal{A}_{21}$ = -2.68, which indicates that the slow eigen-mode is dominant. We note that the negative sign of the cross-term amplitudes are responsible for the concave downward shape of the three-dimensional surface, and for its convex contours for values of τ_1, τ_3 > 32 ms. From the above analysis, we conclude that for this model, the 32 ms time scale serves as an experimental demarcation point. For short time intervals (τ_1, $\tau_3 \approx$ 32 ms), the system primarily undergoes 'fast' exchange of population between state-0 and state-1, and for longer time intervals (τ_1, τ_3 > 32 ms), the system undergoes a combination of 'fast' and 'slow' processes that exchanges population between all three states.

We next examine the possibility that state-1 shown in Fig. 2.4 can function as an intermediate, so that the exchange reactions 1 ⇄ 2 (shown in red) can bridge the 0 ⇄ 1 and the 0 ⇄ 2 reactions (shown in black). We first outline our expectations based on qualitative arguments before examining the theoretical results of the model. Suppose, for example, that when a gp32 monomer binds to the ssDNA template to form state-1, that it might rapidly slide to a 'productive' site allowing for a second gp32 monomer to bind cooperatively, and thus to form a stable dimer. Were this the prevalent mechanism, it would be reflected by the occurrence of four-point pathways at short time intervals that lead to the assembly of the ssDNA-(gp32)$_2$ product (e.g., $\delta A_0 \delta A_1 \delta A_1 \delta A_2$). The resulting 4th-order TCF would then decay rapidly with increasing values of τ_1, τ_3, and exhibit a pattern of positive correlation between successive elementary steps 0 ⇄ 1 and 1 ⇄ 2, which collectively lead to the formation of product. In contrast, if the gp32 monomer state-1 were unstable (due to its presumably slow exchange with state-2), its rapid dissociation would block its ability to act as a 'gateway' intermediate along the assembly pathway. In this latter situation, the intermediate state-1 behaves as a competitive inhibitor to the direct formation of state-2, so that the 4th-order TCF would decay slowly and exhibit a pattern of negative correlation between the successive elementary steps 0 ⇄ 1 and 1 ⇄ 2 (or 0 ⇄ 2). Therefore, depending on whether the 1 ⇄ 2 exchange time scale is fast, slow or intermediate in comparison to the fastest local relaxation time of the system (in the current example, ~ 32 ms), the global rate of population exchange can either be sped up, slowed down, or left unaffected by the presence of the intermediate state-1. These three scenarios correspond to positive, negative, and zero 4th-order correlation, respectively, between successive elementary chemical steps. The signs and magnitudes of the cross-term amplitudes, $\mathcal{A}_{12}$ and $\mathcal{A}_{21}$ serve to characterize whether 4th-order correlation is positive, negative or zero.

We now consider the case in which the exchange rate constants between state-1 and state-2 are assigned to an intermediate value k_{12} = 17 s^{-1} (k_{12}^{-1} = 60 ms) in comparison to the 'fast' and 'slow' local exchange processes (30 s^{-1} and 6 s^{-1}, respectively) described for the case of k_{12} = 0. These conditions are expected to mimic the scenario of competitive inhibition described above. In Fig. 2.5B, we plot the 4th-order TCF using these parameters, which decays for all non-zero values of τ_1 and τ_3 with collective relaxation rates λ_1 = 50 s^{-1} and λ_2 = 19 s^{-1} (λ_1^{-1} = 20 ms and λ_2^{-1} = 53 ms). The introduction of the 1 ⇄ 2 step permits a new pathway for population exchange to occur between all three states, which leads to a dramatic speedup of the slow collective relaxation (i.e. the second eigenvalue λ_2: 5.2 → 19 s^{-1}). Under these conditions, the self-term amplitudes are determined to be $\mathcal{A}_{11}$ = 0.472, $\mathcal{A}_{22}$ = 4.94, and the cross-term amplitudes $\mathcal{A}_{12}$ = $\mathcal{A}_{21}$ = -1.52. As in the previous case, the convex contour lines exhibited by the 4th-order TCF are due to the negative cross-term amplitudes, which indicate the presence of kinetic 'bottleneck' states within the four-point pathways that lead to the exchange of population between all three states. Under these conditions, the reactant state-0 is much more likely to form the intermediate state-1 than to directly form the product state-2. However, once formed, the intermediate is much more likely to undergo the reverse dissociation reaction than to proceed to form product. Thus, for short intervals τ_1 and τ_3 (< 32 ms), the 4th-order TCF is most heavily weighted by the 'fast' exchange between state-0 and state-1. Only at longer time intervals does the 4th-order TCF decay due to the contributions of slower processes such as the coupling step from state-1 to state-2.

In Fig. 2.5C, we plot the 4th-order TCF for the case k_{12} = 33 s^{-1} (k_{12}^{-1} = 30 ms). Under these conditions the rate constants for the 1 ⇄ 2 exchange reactions closely match those of the 0 ⇄ 1 process discussed above for the k_{12} = 0 ms^{-1} case. The 4th-order TCF decays for all values of τ_1 and τ_3 with collective relaxation rates λ_1 = 81 s^{-1} and λ_2 = 22 s^{-1} (λ_1^{-1} = 12 ms and λ_2^{-1} = 45 ms), and with self- and cross-term amplitudes $\mathcal{A}_{11}$ = 0.0001, $\mathcal{A}_{22}$ = 2.26, and $\mathcal{A}_{12}$ = $\mathcal{A}_{21}$ = 0.017, respectively. Under these conditions, only the slower of the two collective relaxation processes carries significant amplitude, and the curvature of the 4th-order TCF is neither convex nor concave. From Eq. (2.17), we see that in the absence of cross-term amplitude (i.e., for $\mathcal{A}_{12}$ = $\mathcal{A}_{21}$ $\approx$ 0), a cross-section of the 4th-order TCF along a vertical slice (with respect to τ_3 and, for example, setting τ_1 = 0) decays at precisely half the rate as does the decay along the diagonal line (with respect to $\tau_1 + \tau_3$, and setting $\tau_1 = \tau_3$), so that the contours of the 2D surfaces are straight anti-diagonal lines. The absence of 4th-order correlation can be understood as a consequence of the close matching of time scales between the 1 ⇄ 2 and 0 ⇄ 1 exchange processes. Because population can readily exchange between all three states via the intermediate state-1, successive elementary reaction steps may occur in an uncorrelated manner.

By further increasing the 1 ⇄ 2 exchange rate constants to the value k_{12} = 67 s^{-1} (k_{12}^{-1} = 15 ms), we model the situation of enhanced kinetic exchange between the intermediate and product states, as described above. In Fig. 2.5D, we plot the 4th-order TCF for these conditions, which decays for all non-zero values of τ_1 and τ_3 with characteristic relaxation rates λ_1 = 146 s^{-1} and λ_2 = 23 s^{-1} (λ_1^{-1} = 6.8 ms and λ_2^{-1} = 43 ms), and with self- and cross-term amplitudes $\mathcal{A}_{11}$ = 0.155, $\mathcal{A}_{22}$ = 1.16, and $\mathcal{A}_{12}$ = $\mathcal{A}_{21}$ = 0.419, respectively. In this case, the 1 ⇄ 2 exchange rate constants are much faster than those of the 32 ms 0 ⇄ 1 process. This leads the 4th-order TCF to decay much more rapidly than in any of the previous situations, and to exhibit concave surface contours as a consequence of the positive-valued cross-term amplitudes. The concave surface curvature indicates that under these conditions, the intermediate state-1 functions as a 'gateway' species whose presence enhances the formation of the product state.

It is often useful to represent Eq. (2.17) as a two-dimensional (2D) rate domain spectrum through the inverse Laplace transform (ILT) – i.e., $\bar{C}^{(4)}(\tau_1, \tau_3)]_{\tau_{2\,fixed}} \xrightarrow{\text{ILT},\tau_1,\tau_3}$

$$\hat{C}^{(4)}(k_1, k_3)]_{\tau_2\,fixed} =$$
$$\mathcal{A}_{11}(\tau_2)\delta(k_1 - \lambda_1)\delta(k_3 - \lambda_1) + \mathcal{A}_{12}(\tau_2)\delta(k_1 - \lambda_1)\delta(k_3 - \lambda_2) \tag{2.18}$$
$$+\mathcal{A}_{21}(\tau_2)\delta(k_1 - \lambda_1)\delta(k_3 - \lambda_2) + \mathcal{A}_{22}(\tau_2)\delta(k_1 - \lambda_2)\delta(k_3 - \lambda_2).$$

The 2D rate spectrum is a sum of four delta functions, which are defined in the k_1, k_3-plane. Comparison between Eq. (2.17) and Eq. (2.18) shows that exponentially decaying terms in the 4th-order TCF are represented as delta functions centered at values corresponding to the collective relaxation rates, λ_1 and λ_2 (see Figs. 2.5E – 2.5H). The two terms positioned along the 'diagonal' line ($k_1 = k_3$), which occur at the positions $(k_1, k_3) = (\lambda_1, \lambda_1)$ and (λ_2, λ_2), respectively, correspond to the self-terms with amplitudes $\mathcal{A}_{11}$ and $\mathcal{A}_{22}$. The cross-terms with amplitudes $\mathcal{A}_{12}$ and $\mathcal{A}_{21}$ occur above and below the diagonal, at the positions $(k_1, k_3) = (\lambda_1, \lambda_2)$ and (λ_2, λ_1), respectively. These self- and cross-term features of the 2D rate spectrum represent the same amplitudes discussed above for the 4th-order TCF, and thus serve as an equivalent representation of the collective dynamics of the coupled cyclical $N = 3$ system.

Such 2D rate spectra are made in analogy to the often-used frequency domain spectra of 2D Fourier transform spectroscopy.[17, 25-27] The diagonal and off-diagonal terms generally decay as a function of the waiting time τ_2. Cross-term amplitudes indicate the 'exchange' of populations between states involved in collective relaxation processes, and these terms decay on time scales that match the exchange dynamics. For situations in which the cross-term amplitudes are zero, the collective relaxation processes (defined by the eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$) are independent as depicted in Fig. 2.5G. Negative or positive cross-term amplitudes (as depicted in Figs. 2.5F and 2.5H, respectively) indicate that such processes are negatively or positively correlated, which is possible for pathways with $N \geq 3$. As discussed in the context of our model calculations, the $N = 3$ scheme shown in Fig. 2.4, in which the intermediate state-1 functions as a rate-limiting 'bottleneck' (i.e., with $k_{01}, k_{10} \gg k_{12} \approx k_{21} \gg k_{02}, k_{20}$), exhibits negative 4th-order correlation. In contrast, the same scheme in which the intermediate functions as a 'gateway' species (i.e., with $k_{01}, k_{10} \ll k_{12} \approx k_{21} \gg k_{02}, k_{20}$) exhibits positive 4th-order correlation. For display purposes, we have artificially broadened the diagonal and off-diagonal features in our 2D rate spectra in Figs. 2.5E – 2.5H using a Gaussian convolution.

As previously mentioned, both TCF and HMM analyses can, in principle, provide similar information about the states and kinetics of a stochastically fluctuating chemical system. To illustrate this point, we plot in Fig. 2.6 the so-called 'transition density plot' (TDP) alongside the corresponding 4th-order TCFs and 2D rate spectra. A TDP is a useful way to present the information about transition pathways that is potentially available from an HMM analysis.[23] The TDP describes the time-integrated joint distribution $p_{ji}\left(A_j, \tau; A_i, 0\right)$ of molecules that are initially in state-i with observable value A_i, and which at a later time τ undergo a transition directly to state-j with observable value A_j. The weights of the TDP are given by the expression

$$p_{ij}\left(A_j, \tau; A_i, 0\right) = k_{ij}\, p_i^{eq}(1 - e^{-k_{ji}\tau}) \tag{2.19}$$

(see Appendix C, Chapter III for derivation). Thus, a time-dependent TDP contains information about the direct state-to-state transitions that occur within the time interval τ, and such information could be useful, in principle, to infer assignments to the various states involved within a transition pathway. We note that in the long-time limit, the joint distribution must be a symmetric function, i.e., $p_{ji}^{eq}\left(A_j, A_i\right) = k_{ij}\, p_j^{eq} \cdot k_{ji}\, p_i^{eq}$ with $p_j^{eq} = p_i^{eq}$, which is necessary to satisfy detailed balance. Nevertheless, this symmetry need not be valid at short or intermediate times, since the various state-to-state transitions may occur on entirely different time scales. Only in the limit of very long time intervals (i.e., longer than the slowest relaxation of the system) are the forward and backward flow of state occupancies along all inter-connected transition paths expected to be equal.

In Fig. 2.6, we present model calculations for the *linear* N = 3 scheme (shown in Fig. 2.1B) of the 4th-order TCFs, the 2D rate spectra, and the TDPs as a function of the waiting period τ_2. For these calculations, we have chosen the rate constants $k_{01} = k_{21} =$ 10 s^{-1} and $k_{10} = k_{12}$ = 20 s^{-1}, with signal observables A_0 = 0.9, A_1 = 0.3, and A_2 = 0.1 (see Fig. 2.1B). The collective relaxation rates of the system are λ_1 = 50 s^{-1} and λ_2 = 10 s^{-1} (λ_1^{-1} = 20 ms and λ_2^{-1} = 100 ms), and the equilibrium distribution of populations is given by p_0^{eq} = 0.4, p_1^{eq} = 0.2, and p_2^{eq} = 0.4. This system has the interesting property that it crosses over from a regime of negative 4th-order correlation at short waiting intervals

($\tau_2 < 27$ ms) to one of positive 4th-order correlation at long waiting intervals ($\tau_2 > 27$ ms). The time-dependent crossover is evident from the shapes of the contour lines of the 4th-order TCFs (Fig. 2.6A) and the signs of the cross-term amplitudes of the 2D rate spectra (Fig. 2.6B). This is due to the fact that for waiting periods less than 27 ms, the four-point pathways are heavily weighted by transitions leading away from the intermediate state-1, either in the backward direction toward the reactant state-0, or in the forward direction toward the product state-2. An initial step in either direction will tend to inhibit the successive step in the opposite direction, thereby inhibiting the global exchange of population between all three states. An example four-point pathway for a short waiting time is $\delta A_1 \delta A_0 \cdot \tau_{2,\text{short}} \cdot \delta A_0 \delta A_1$. In contrast, for waiting time intervals greater than 27 ms, the four-point pathways will tend to be dominated by sequences in which an initial fast step in the direction away from the intermediate state-1 (towards state-0 or state-2) will, after an intervening waiting time that exceeds the fast process, be positively correlated to a subsequent fast step in the opposite direction. An example four-point pathway such a waiting time is $\delta A_1 \delta A_0 \cdots \tau_{2,\text{long}} \cdots \delta A_1 \delta A_2$. This example illustrates that the time-dependences of the 4th-order TCFs, 2D rate spectra, and the TDPs can provide information about the connectivity of a chemical network, its rate constants, and the observable values A_0, A_1 and A_2.

Optimization of N-State Kinetic Models to Sub-Millisecond Single-Molecule Fluorescence Data

In the preceding discussion, we have shown that analytical expressions for the TCFs of discrete stochastic systems with $N = 2$ or 3 can be readily obtained. For systems of higher complexity ($N \geq 4$), it is often practical to solve Eq. (2.8) numerically. These solutions can be used to rapidly generate: (*i*) the 2nd-order TCF $\bar{C}^{(2)}(\tau)$; (*ii*) the 4th-order TCF $\bar{C}^{(4)}(\tau_1, \tau_2, \tau_3)$ and its corresponding 2D rate spectrum; (*iii*) the equilibrium distribution of states $p_i^{eq}(A_i)$; and (*iv*) the time-dependent joint distribution of states (i.e. the time-dependent transition density plot, TDP) $p_{ji}\big(A_j, \tau; A_i, 0\big)$. By applying the algorithms discussed in Section IV to calculate quantities (*i*) – (*iv*), we may implement a multi-parameter optimization strategy to obtain the simplest kinetic scheme that can accurately represent the experimental behavior of single-molecule fluorescence data.

As indicated above, conventional single-molecule fluorescence experiments performed on discrete-state systems often employ 100-ms time resolution. Such measurements can provide useful kinetic information on this time scale through direct visual inspection, or by using hidden Markov model (HMM) analyses to obtain idealized single-molecule trajectories.[23] Single-molecule experiments with sub-millisecond time resolution provide only sparse trajectory data[2] that are not strictly amenable to direct visual inspection or HMM analyses. This is mostly due to the influence of stochastic noise – i.e., when a fixed number (n) of data points is measured over a short period of time, the signal-to-noise ratio (S/N) during this interval has a lower bound of $\sqrt{n}$. We therefore turn to the analysis described in this work, which is based on the use of TCFs and state distribution functions to extract detailed and useful kinetic information about multi-step transition pathways.

Here we prescribe a step-by-step protocol to analyze sparse single-molecule trajectory data. This approach is based on multi-parameter optimization algorithms that have been widely applied in numerous experimental contexts.[34-36] We must first consider the 2nd-order TCF, which is constructed from individual single-molecule trajectories as described by Eq. (2.2). Each TCF may vary from trajectory-to-trajectory, depending on system heterogeneity, the experimental S/N, and on the number of data-points included in the calculation. The characteristic relaxation times are reflected by the decay of the 2nd-order TCF. These are limited in range by the time-resolution of the measurement and by the maximum duration of a single-molecule trajectory. To reduce the effects of stochastic noise, the TCFs constructed from many individual trajectories should be averaged together.[17] By fitting this decay to a model multi-exponential function, one determines the minimum number ($N - 1$) of relaxation components necessary to represent the system. The value of N so determined represents the minimum number of states, since the presence of additional relaxation components might be difficult to detect due to relatively small contributing amplitudes, or to the presence of eigen-mode degeneracy – i.e. the possibility that multiple relaxation components share the same (or nearly the same) relaxation time (eigenvalue).

Information about forward and backward rate constants associated with individual steps along the reaction pathway is contained within the 4th-order TCF. The 4th-order TCF is constructed from experimental trajectories using Eq. (2.5). The time intervals τ_1 and τ_3 must be varied over a range that spans the individual decay components present in the 2nd-order TCF, while the waiting time interval τ_2 must be varied over a range that spans slow exchange time dynamics of the system.

Simulated expressions for the 2nd- and 4th-order TCFs, and the equilibrium distribution of states, are calculated using the *N*-state master equation that is described by Eq. (2.7). An optimized solution can be determined by minimizing the difference between the experimentally derived functions, and the simulated functions while varying the input parameters specified by the rate constants k_{ij} and the observable values A_i. We thus achieve a globally optimized solution to the kinetic problem of the *N*-state system.

Conclusions

We have shown how the analysis of 2nd- and 4th-order TCFs of single-molecule trajectories can be used to learn about the roles of short-lived intermediates in biochemical reactions. In principle, 6th-order and higher TCFs could be used to study the details of even more complex biochemical reactions than the relatively simple $N = 3$ and $N = 4$ schemes examined here. The implementation of higher dimensionality TCFs is, of course, limited by S/N and data availability. Nevertheless, with the steady improvements that are currently underway to single-molecule methodology and detector technologies, such applications of generalized TCFs to elucidate complex biochemical pathways are now feasible.

The implementation of a generalized TCF analysis to microsecond-resolved single-molecule fluorescence measurements can be a powerful way to extract detailed information when the signal is too noisy to warrant analysis by direct visualization methods (e.g., HMM). However, unlike HMM, generalized TCFs are rarely utilized for such experiments. This is likely because the theory surrounding this analysis is relatively abstract and not easily approached by a general biophysical audience. In this manuscript, we have outlined the theoretical foundations to apply a generalized TCF approach to analyze single-molecule data, and we illustrated these ideas in the context of the ssDNA-(gp32)$_n$ binding system shown in Fig. 2.1A.

While the generalized concepts of TCF have not yet been widely applied to the analysis of single-molecule fluorescence measurements, they hold great promise for future microsecond kinetic studies, and for experiments carried out under low signal conditions. Since many important bio-molecular interactions occur on sub-millisecond timescales, we anticipate that the application of TCF methodology can help to provide new insights to understand these dynamics, which have thus far proven difficult to access experimentally.

Bridge to Chapter III

This work first establishes the theory behind using Markov models to simulate time correlation functions and FRET probability distributions. The methodology we presented is general, though this chapter used examples corresponding to single-molecule experiments. The next chapter takes the analytical methods established here and applies them to microsecond-resolved experimental single-molecule FRET measurements to elucidate possible models for gp32 dimer assembly.

CHAPTER III

USING MICROSECOND SINGLE-MOLECULE FRET TO DETERMINE THE ASSEMBLY PATHWAYS OF T4 SSDNA BINDING PROTEIN ONTO MODEL DNA REPLICATION FORKS

This manuscript was published in the *Proceedings of the National Academy of Sciences of the united states of America* (PNAS) in April 2017. I prepared DNA samples for the experiment, as well as constructed the microscopy chambers along with Carey Phelps. I took ensemble and single-molecule data for the experiment along with Carey Phelps and Morgan Marsh. I was involved in the drafting and editing of the paper along with coauthors Carey Phelps, Davis Jose, Morgan Marsh, Andrew H. Marcus and Peter H. von Hippel. This work was supported by grants from the National Science Foundation (Chemistry of Life Processes Program grant CHE-1608915 to A.H.M.) and the National Institutes of Health (NIGMS Grant GM-15792 to P.H.v. H. and A.H.M.). Andrew H. Marcus was the principal investigator for this work.

Introduction

The DNA replication complex of the T4 bacteriophage is an excellent model to understand the mechanistic details of DNA synthesis, since it employs the same three protein sub-assemblies as found in higher organisms, albeit without the many additional layers of regulatory complexity[1-5]. These are: (i) the helicase / primase (primosome) complex that unwinds the double-stranded (ds) DNA genome and synthesizes pentameric RNA primer strands while exposing the leading and lagging single-stranded (ss) DNA templates; (ii) the DNA polymerases that use the exposed templates to synthesize complementary DNA daughter strands; and (iii) the replication clamp-clamp loader complexes that load and unload the sliding clamps from the functioning polymerases and thereby control the processivity of DNA synthesis.

An integral component of DNA replication is the ssDNA binding protein (ssb)[6-7]. The ssbs bind to the exposed ssDNA templates during the critical period after the helicase has unwound these sequences, and before the DNA polymerases have incorporated complementary paired nucleotides into the newly formed daughter strands. The ssbs are thought to protect ssDNA from nuclease activity, and to remove unfavorable secondary structures that would otherwise hinder the efficiency of the replication process. The T4 ssb is referred to as the gene product 32-protein (or gp32), and it is known to form association complexes with ssDNA through a cooperative binding mechanism. The cooperative binding mode of gp32 is thought to be important to achieve complete coverage of the exposed ssDNA templates at the precisely regulated concentration of protein that needs to be maintained in the infected *E. coli* cell (8, 9). The gp32 protein has an N-terminal domain, a C-terminal domain, and a core domain.

The N-terminal domain is necessary for the cooperative binding of the gp32 protein through its interactions with the core domain of an adjacent gp32 protein. To bind to ssDNA, the C-terminal domain of the gp32 protein must undergo a conformational change that exposes the positively charged region of its core domain, which in turn interacts with the negatively charged ssDNA backbone. The binding site size of the gp32 protein is 7 nucleotide residues (nts)[10]. While many of the ssbs of higher organisms have larger binding footprints, the functional role of binding cooperativity in these systems is less straightforward. Thus, oligomers of *E. coli* ssb can bind to ssDNA in more than one binding mode [11] while mammalian RPAs may bind and function as monomers[12]. Both E. coli ssb and human RPA have been shown to diffuse on ssDNA, and this may be important in discharging their functions [13-14]. Although sliding of the gp32 protein along the template strand is expected to be important to its biological function, there is currently little direct information available about the role of sliding in the process of gp32 cluster assembly on ssDNA template sequences during replication.

In this work we use microsecond-resolved single-molecule Förster resonance energy transfer (smFRET) experiments to study the kinetics and mechanism of gp32 dimer assembly on a short 15-nt ssDNA template. Such ssDNA-(gp32)$_2$ complexes can serve as simple model systems to examine the basic biochemical steps involved in ssb filament assembly and sliding. Our experiments employed fluorescently labeled primer-template (p/t) DNA constructs in which a Cy3 donor chromophore was attached to the 3'-end of the template strand, and a Cy5 acceptor chromophore was attached to the 5'-end of the primer strand near the p/t junction (see Fig. 3.1A, see Appendix A for all figures). The template strand contains a 15-nt poly(deoxythymidine) [p(dT)$_{15}$] sequence that can form an association complex with up to two cooperatively bound gp32 protein monomers at one time. Recently, Lee *et al.* showed that smFRET signals observed from this same p/t DNA construct in the presence of gp32 undergoes intermittent protein-induced fluctuations, which reflect changes in the end-to-end distance of the ssDNA template due to binding and dissociation of gp32 proteins[15].

The experiments performed by Lee *et al*. were sensitive to changes in ssDNA template conformation, which could be observed using the tens-of-milliseconds time resolution of standard smFRET experiments. While these studies could distinguish between unbound template conformations and those with two cooperatively bound gp32 proteins, they could not clearly resolve short-lived singly-bound intermediate states. In previous work we used microsecond resolved smFRET and single-molecule linear dichroism (smLD) to study DNA breathing fluctuations at model replication forks, and the influence of these fluctuations on helicase binding[16]. These measurements detected individual fluorescence photo-counts from a single molecule, and stored the ensuing information about the intervening time intervals and optical phase conditions associated with each detection event. The method is technically similar to those implemented by others to study fast activated-barrier-crossing in protein folding[17], and photon anti-bunching in quantum dot nanocrystals[18]. Here we show how microsecond-resolved smFRET experiments can be used to probe short-lived, singly-bound intermediate states and their influence on the assembly pathways of gp32-$(ssDNA)_2$ complexes. These studies thus permit us to examine the detailed mechanisms of non-cooperative and cooperative protein binding to the $p(dT)_{15}$ lattice, and the potential involvement of 'sliding' of the various bound species along the ssDNA strand.

A challenge to the interpretation of microsecond smFRET experiments lies in the difficulties involved in resolving short-lived intermediates. This problem is further complicated by the necessity to partition the finite signal intensity into very short (microsecond) sampling intervals, which leads to smFRET trajectories that appear 'sparse' and dominated by stochastic noise when viewed with high temporal resolution. Such trajectories are not amenable to analysis by 'direct visualization methods,' which attempt to assign discrete smFRET efficiency values to corresponding conformational states and thus to observe and resolve sequences of state-to-state transitions that make up a biochemical reaction pathway. For example, hidden Markov model (HMM) analysis is a particularly useful method to evaluate smFRET trajectories that use tens-of-milliseconds resolution[19], yet this approach becomes less accurate for microsecond resolved experiments.

In this work, we apply generalized concepts of time-correlation functions (TCFs) to study the ssDNA-(gp32)$_2$ assembly pathways[20-30]. We assume that our single-molecule experiments probe the instantaneous conformational state of the system at equilibrium, and that the stochastically fluctuating smFRET efficiency, E_{FRET}, can be directly mapped onto this state. A TCF is a time-dependent moment of the variable $E_{FRET}(t)$ that, when utilized to its full potential, can provide a statistically meaningful way to characterize the dynamics of the interconverting species lying along the reaction pathway. In general, the n^{th}-order TCF, $C^{(n)}(\tau_1, \tau_2, \ldots, \tau_{n-1})$, can be written as the average product of n successive observations $\langle E_{FRET}(t_1), E_{FRET}(t_2), \ldots, E_{FRET}(t_n) \rangle$, which depends on the $n-1$ time intervals $\tau_1 = t_2 - t_1$, $\tau_2 = t_3 - t_2$, …, $\tau_{n-1} = t_n - t_{n-1}$. The complexity of information available from a TCF depends on its order. For example, the 2nd-order TCF, $C^{(2)}(\tau)$, which is often referred to as the 'time-autocorrelation function,' contains information about the average time scales of the fluctuations of the system. However, the 2nd-order TCF cannot be used to determine information about the role of intermediates that may lie along a reaction pathway. Such information is available through the 4th-order TCF, $C^{(4)}(\tau_1, \tau_2, \tau_3)$, which is sensitive to whether the system, on average, undergoes a particular state-to-state transition that typically follows or precedes another, or whether two such transitions occur independently.

In the analysis that follows, we implemented 2nd- and 4th-order TCFs to study the role of intermediates along the gp32-ssDNA assembly pathway. An overview of the theoretical method that we implemented for this purpose is given in[20]. We note that 4th-order TCFs are important in the analysis of nonlinear spectroscopies, including two-dimensional (2D) NMR, 2D infra-red, and 2D electronic spectroscopy[31]. High-order TCFs have also been used to characterize the dynamics of a number of stochastic chemical systems[24,25] including protein reaction dynamics[23], molecular and protein diffusion in solution[26, 27, 32], polymer diffusion in the liquid phase[28], and protein conformation fluctuations in molecular dynamics (MD) simulations[29].

Materials and Methods

Instrumentation for Microsecond-Resolved smFRET.

We performed smFRET experiments at both 30-msec- and 1-μsec-resolution using custom instrumentation and procedures previously developed in our laboratory. The instrument applies a 1-MHz modulation to the optical phase of the 532 nm continuous wave excitation, and records each single photon detection event with an associated 64-bin phase and microsecond-resolved time 'stamp.' Instrumental details and theoretical considerations to perform these measurements are provided elsewhere[16].

Sample Preparation and Model DNA Replication Fork Constructs.

Model p/t DNA constructs that were labeled with the Cy3/Cy5 donor-acceptor FRET pairs were purchased from Integrated DNA Technologies. Strand sequences of these substrates are shown in Table S3.1 of in Appendix B Chapter III Supporting Information. Microfluidic sample chambers were constructed from microscope slides and coverslips. Details of the sample chamber construction, cleaning procedures, and sample preparation are the same as those reported by Lee *et al.*[15].

Results

gp32 binding to 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA induces stochastic transitions between conformational states of the p(dT)$_{15}$ template strand.

We initially performed 30 msec resolved smFRET experiments on the donor (Cy3) – acceptor (Cy5) labeled p/t DNA construct, as shown schematically in Fig. 3.1A. The nucleic acid base sequences used for the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct, as well as our nomenclature conventions for these p/t DNA constructs, is shown in Table S3.1 and are the same as those used in related studies[15]. As mentioned previously, the p(dT)$_{15}$ segment of the template strand can support binding of up to two gp32 proteins, each of which possess a binding site size of 7 nucleotide residues[10]. Binding of the gp32 protein to ssDNA is known to stiffen the p(dT)$_{15}$ segment of the p/t DNA substrate [15,33]. This interaction increases the average separation between the Cy3 and Cy5 labels, which leads to a decreased smFRET efficiency, E_{FRET} (defined below).

Because the ssDNA segment of the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA can bind up to a maximum of two gp32 monomer proteins, we anticipate the presence of numerous ssDNA-(gp32)$_n$ assembly complexes, corresponding to n = 0-, 1-, or 2-bound gp32 proteins. The interfacial free energy between two cooperatively bound gp32 monomers on a ssDNA template leads to a thousand-fold increase in binding affinity per monomer[34]. Thus, the cooperative binding mode of gp32 to ssDNA likely plays an important role in the transformation between 1-bound intermediates and 2-bound end-states. Because each gp32 protein occludes 7 nts, there are nine (= 15 – 7 + 1) possible 1-bound gp32-p(dT)$_{15}$ conformations (e.g., bound at positions 1 – 7, 2 – 8, …, 9 – 15), and there are two cooperative 2-bound conformations (at positions 1 – 14, and 2 – 15). In Fig. 3.1A, we depict a simplified three-state reaction scheme in which the various sub-states of 0-, 1- and 2-bound conformations are assumed to be experimentally indistinguishable, and the ssDNA-(gp32)$_2$ assembly pathway occurs via a single 1-bound intermediate. We have neglected in this scheme the possibility that the 0- and 2-bound states can interconvert directly, since it is known that gp32 dimers formed in solution cannot bind to ssDNA, but rather must dissociate and bind initially as monomers prior to forming cooperatively-bound dimer clusters[35].

The three-state scheme would be an adequate description of the assembly process if all nine possible singly-bound conformations (and the two possible doubly-bound conformations) could to rapidly interconvert on time scales faster than the ~ 10 µs instrument resolution. For example, a 1-bound intermediate might rapidly 'slide' along the ssDNA template from one to another of the nine possible template positions, some of which are better (or worse) suited to accommodate the binding of a second gp32 monomer for the creation of a 2-bound conformation. On the other hand, if the dissociation of gp32 monomers from the ssDNA template was significantly faster than monomer sliding, the latter process would not occur and it might be possible to distinguish between 'unproductive' and 'productive' intermediates that lie along the reaction pathway. We depict such a four-state scheme in Fig. 3.1B, where we have identified the 'unproductive' intermediate as state-1 and the 'productive' intermediate as state-1'. There are five possible 'unproductive' 1-bound conformations for which the initially bound protein (at positions 3 – 9, 4 – 10, …, 7 – 13) occludes the 7 nt binding site of the second protein, and there are four ‘productive’ 1’-bound conformations (at positions 1 – 7, 2 – 8, 8 – 14, and 10 – 15). Of course, even more complex schemes would need to be considered if the interconversion time scales of various sub-states could be resolved by the experiment.

In Fig. 3.1C, we present the results of bulk FRET experiments, in which the p/t DNA construct was titrated against gp32 protein in buffer consisting of 10 mM Tris at pH 8, 100 mM NaCl, and 6 mM $MgCl_2$. In the absence of gp32, the emission intensity corresponding to Cy5 (peaked at 663 nm) was much larger than that corresponding to Cy3 (peaked at 561 nm). This is due to the relatively short average distance between the 3’ end-labeled Cy3 chromophore and the p/t junction-labeled Cy5 chromophore in this rapidly fluctuating random coil ssDNA sequence. Comparison of the relative Cy3/Cy5 peak intensities (I_{Cy3}/I_{Cy5}) lead us to estimate the average bulk FRET efficiency of the p/t DNA substrate in the absence of gp32 to be $E_{FRET} = I_{Cy5}/\left(I_{Cy3} + I_{Cy5}\right) \approx 0.81$. Upon titration of gp32 into the solution, the ratio I_{Cy3}/I_{Cy5} increased sensitively with increasing gp32 concentration. These observations are consistent with those of previous studies[15], which show that binding of gp32 to the ssDNA region of the p/t DNA construct leads to chain stiffening and thus a decrease in the E_{FRET} value.

We next carried out smFRET experiments on the p/t DNA construct using a split-screen electron multiplied (em) CCD camera, which was capable of 30-msec time resolution. In Fig. 3.2A, we show representative single-molecule trajectories, which were taken from the 3’-Cy3/Cy5-p(dT)$_{15}$-p/t DNA samples incubated in the presence of 0 μM gp32, 0.1 μM gp32 and 1.0 μM gp32, respectively. Here, the values of E_{FRET} were calculated from Cy3 and Cy5 intensities according to the same formula as for our bulk experiments. In the absence of protein (Fig. 3.2A, top panel), the E_{FRET} values were typically constant over the course of an ~ 30 sec trajectory, although the precise value of E_{FRET} varied somewhat from molecule to molecule over a range of $\Delta E_{FRET} = \pm 0.5$. We characterized the distribution of state occupancies by constructing histograms of E_{FRET} values, each of which were made up of several thousand data point entries (see Fig. 3.2B). In the absence of the gp32 protein, there is a single feature centered at E_{FRET} ~0.81

(Fig. 3.2B, top panel). We note this peak value of E_{FRET} is very similar to that we obtained from our bulk measurements (see Fig. 3.1C).

For the experiments with p/t DNA constructs and gp32 concentrations of 0.1 μM and 1.0 μM, we observed smFRET trajectories that typically exhibited continuous stochastic transitions between relatively high and low E_{FRET} values (see middle and bottom panels of Fig. 3.2A, respectively, for representative trajectories). For samples incubated with 0.1 μM gp32, the system occupied primarily a state with a relatively high E_{FRET} value, and for brief periods of time a state with relatively low E_{FRET} values. Histograms constructed from smFRET trajectories recorded under these conditions exhibited a dominant feature centered at E_{FRET} ~0.81, and a minor feature centered at E_{FRET} ~0.56 (see Fig. 3.2B, middle panel). We assigned the state with E_{FRET} values of ~0.81 to the 0-bound conformation, and the state with E_{FRET} values near ~0.56 to the 2-bound conformation, consistent with previous studies[15]. For samples incubated with 1.0 μM gp32, we found that the equilibrium occupancies of the 0- and 2-bound conformations were reversed relative to those seen in the 0.1 μM gp32 experiments (see Fig 3.2B, bottom panel). At the elevated protein concentration, the system occupied primarily a 2-bound state, and for relatively brief periods also a 0-bound state.

While the single-molecule experiments described above allowed us to directly observe individual gp32-ssDNA association and dissociation events, it was not possible to detect intermediates that persisted for times shorter than the 30 msec instrument resolution, or to distinguish between different types of bound ssDNA-$(gp32)_n$ states. Because the ssDNA segment of the p/t DNA construct could accommodate the binding of up to two gp32 monomers, it is reasonable to expect to detect up to three or more conformations with discrete E_{FRET} values, provided there is sufficient instrument time resolution. Our observation of only two conformations with well-separated E_{FRET} values suggested the presence of one or more short-lived intermediates, which might themselves have exhibited E_{FRET} values that would be difficult to distinguish from those of the 0- and 2-bound end-states. In the next section, we describe smFRET experiments on the same p/t DNA constructs using an apparatus capable of ~10 μs resolution.[16]

gp32-p(dT)15 binding involves template conformation dynamics that occur at two well separated time scales.

The 2nd-order TCF of the time varying signal $E_{FRET}(t)$ is the average product of two successive measurements made at times t_1 and t_2, which are separated by the interval $\tau = t_2 - t_1$

$$C^{(2)}(\tau) = \langle E_{FRET}(0)E_{FRET}(\tau)\rangle \qquad (3.1)$$

In Eq. (3.1), the angle brackets denote the average 'two-point product' that has been performed over all possible starting times, given by $C^{(2)}(\tau) = \int_{-\infty}^{\infty} E_{FRET}(t)E_{FRET}(t + \tau)dt$. If the longest relaxation time of the system exceeds the duration of an individual single-molecule trajectory, then Eq. (3.1) must be additionally averaged over a large number of data sets.

We performed microsecond smFRET experiments on the 3'-Cy3/Cy5-p(dT)15-p/t DNA construct in the absence of the gp32 protein. For these samples, we found that the 2nd-order TCF did not exhibit a decay that could be distinguished from instrument noise (see Fig. S3.1A). This observation suggests that conformational fluctuations of the p/t DNA substrate might occur much faster than the ~10 µs instrument resolution, and/or much slower than the 30 sec duration of a typical smFRET trajectory.

We next performed microsecond-resolved smFRET experiments on the 3'-Cy3/Cy5-p(dT)15-p/t DNA construct in the presence of 0.1 µM gp32. When viewed on a linear-log scale, the 2nd-order TCF exhibited an estimated $1/e$ decay correlation constant τ_c ~ 160 msec (see Fig. S3.1B). The same data plotted using a log-linear scale could be well fit to a bi-exponential decay, with decay constants τ_{fast} = 18.4 and τ_{slow} = 157.8 msec, with corresponding amplitudes A_{fast} = 0.146 and A_{slow} = 0.854 (see Fig. 3.3A and Table S3.2). When we performed experiments on the p/t DNA substrate in the presence of 1.0 µM gp32, we found that the decay of the 2nd-order TCF occurred more rapidly than at the lower (0.1 µM) protein concentration, with an average decay constant τ_c ~ 40 msec (see Fig. S3.1C). At this elevated protein concentration, the decay was similarly well described as bi-exponential, with decay constants τ_{fast} = 13.9 msec and τ_{slow} =

94.23 msec, and corresponding amplitudes A_{fast} = 0.531 and A_{slow} = 0.469 (see Fig. 3.3B, and Table S3.2).

The significance of the number of decay components of the 2nd-order TCF can be understood in the context of the theory of Markov chains, which describes the kinetics of an equilibrium chemical system that may undergo stochastic transitions between N discrete state.[36] Markov chain theory predicts that the 2nd-order TCF contains $N - 1$ decay components.[20] Thus, the bi-exponential decay we have observed for $C^{(2)}(\tau)$ implies that $N \geq 3$. Based on the magnitudes of the fast relaxation components we have measured at 0.1 and 1.0 μM gp32 (~ 10 msec), we concluded that there is at least one short-lived intermediate species that fluctuates on this time scale. Likely candidates for such species are the various singly-bound ssDNA-gp32 conformational states. As discussed in previous sections, depending on the loading position of a singly-bound gp32 molecule on the p(dT)$_{15}$ lattice, this state might (or might not) be highly reactive, due to its potential (or lack of potential) for cooperative interactions with a second gp32 molecule that could (or could not) subsequently bind to the lattice. To test this hypothesis, we introduced a short hexameric oligopeptide that inhibits the cooperative gp32 binding modality by competing with the binding interaction of the N-terminus of an adjacent (and potentially cooperatively bound) gp32 molecule.[37]

Disruption of the cooperative modality blocks the formation of doubly-bound states.

The peptide sequence KRKSTA, which is referred to as the 'LAST' sequence, was shown by Karpel and co-workers to disrupt the cooperative assembly mechanism of ssDNA-(gp32)$_n$ complexes by competitive binding to the core domain within the region responsible for binding to the N-terminus of an adjacent gp32 protein.[37] To confirm that the LAST peptide does indeed bind to the gp32 protein, we performed tryptophan-quenching studies in which solutions of gp32 were titrated against LAST. Intrinsic tryptophan residues on the gp32 protein were excited using 288 nm light, and the peak fluorescence was detected at 342 nm. When the LAST peptide was introduced into a solution containing 0.1 μM gp32, we observed that tryptophan fluorescence decreased rapidly, indicating a direct interaction between the LAST peptide and gp32 (see Fig. S3.2A). This interaction appeared to saturate near a 5:1 ratio of LAST peptide to gp32 protein. When we repeated these experiments using a solution containing 1.0 μM gp32, the saturation ratio increased to 10:1 (see Fig. S3.2B), which suggests that the LAST peptide is in competition with other gp32 proteins for its binding site. We performed similar bulk titration measurements on solutions containing 100 nM concentrations of the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct and either 0.1 μM (Figs. S3.3A and S3.3B) or 1.0 μM gp32 (Fig. S3.3C and S3.3D). We found that in the presence of gp32, the FRET efficiency of the p/t DNA construct was enhanced when the LAST peptide was titrated into the solution. These results demonstrate that the presence of the LAST peptide affects gp32 filament assembly, presumably by disrupting the cooperative gp32 binding mechanism.

In Fig. 3.4A, we show a representative smFRET trajectory (taken with 30 msec time resolution) of our 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct in the presence of 0.1 μM gp32 and 0.5 μM LAST (a saturating quantity). We see that the influence of the LAST peptide is to suppress stochastic transitions between conformational states with high- and low-E_{FRET} values. The effect of the LAST peptide on the smFRET trajectories is striking when compared with those obtained using the same 0.1 μM gp32 concentration in the absence of LAST peptide (see Fig. 3.2A, middle panel). We used the HMM method[19] to determine the dwell times associated with the high-E_{FRET} (0-bound) state and the low-E_{FRET} (2-bound) state (see Fig. S3.4). We note that the dwell time of the 2-bound state appears to be insensitive to the presence of the LAST peptide (Fig. S3.4B), suggesting that the presence of the LAST peptide does not perturb the dissociation from the p(dT)$_{15}$ template strand of gp32 dimers that have been successfully cooperatively bound.

In Fig. 3.4B, we show the histogram of E_{FRET} values of the p/t DNA substrate in the presence of 0.1 μM gp32 and 0.5 μM LAST. For samples incubated in the presence of LAST, the relative intensity of the low-E_{FRET} feature at ~0.56 is reduced in comparison to the case when LAST is absent (compare to Fig. 3.2B, middle panel). In Fig. S3.2C, we plot the relative intensities of the low-E_{FRET} (~0.56) to high-E_{FRET} (~0.81) histogram features versus LAST concentration. This ratio decays with increasing concentration of LAST, with a similar trend to that of our bulk tryptophan quenching experiments (see Fig. S3.2A). As shown with our bulk measurements (Fig. 3.1C), we found that saturation at this gp32 concentration occurred at values near a 5:1 ratio of LAST to gp32.

Our finding that the effect of the LAST peptide was to increase the dwell time of the high-E_{FRET} state (with an E_{FRET} value of ~0.81) provides us with a key insight into the mechanism of p(dT)$_{15}$-(gp32)$_n$ filament assembly. The activity of the LAST peptide is known to disrupt the cooperative interaction between adjacently bound gp32 proteins.[37] We thus conclude that the low-E_{FRET} state (with an E_{FRET} value of ~0.56) must correspond to a doubly-bound ssDNA-(gp32)$_2$ complex, since its formation is blocked by the activity of the LAST peptide. Because the LAST peptide only affects the cooperative binding mode of the gp32 protein, we further anticipate that the formation of the singly-bound ssDNA-(gp32)$_1$ complex should be unaffected by the presence of LAST. Our results suggest that the value of E_{FRET} for the singly-bound species is either indistinguishable from that of the zero-bound species (with E_{FRET} ~ 0.81), or that the persistence time of the singly-bound complex is shorter than the 30 msec resolution of our split-screen emCCD smFRET instrument.

We next examined our microsecond resolved smFRET experiments on the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct in which we used the LAST peptide to disrupt the cooperative binding mechanism of the gp32 protein. In Fig. 3.3C, we show the 2nd-order TCF that was constructed from samples incubated in the presence of 0.1 μM gp32 and 1.0 μM LAST. This function was constructed from an average of 21 individual smFRET trajectories. Although the dwell time of the high-E_{FRET} (~ 0.81) state is enhanced by the presence of the LAST peptide, the 2nd-order TCF is represented by a bi-exponential decay with time constants τ_{fast} = 24.3 msec and τ_{slow} = 184 msec, and amplitudes A_{fast} = 0.646 and A_{slow} = 0.354, respectively. We note that the two decay constants are strikingly similar to those found in the absence of LAST, further supporting our conjecture that the experimental system consists of a network of three or more microstates in dynamic equilibrium.

The results we have presented thus far suggest a mechanism for the ssDNA + 2 gp32 ⇄ ssDNA-(gp32)$_2$ assembly reaction in which a 2-bound state is rapidly formed following the initial creation of a short-lived, 1-bound intermediate. Nevertheless, the above results are not sufficient to distinguish between different models in which various (productive and unproductive) 1-bound states might (or might not) interconvert. In the remainder of this paper, we consider the effects of higher-order correlations on smFRET trajectories to ask if different types of short-lived intermediates are important in the assembly mechanism.

Analysis of smFRET trajectories using 4th-order TCFs. Our TCF analysis described above indicated that the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct, in the presence of the gp32 protein, supports two fundamental relaxation processes, one of which occurs on a tens-of-milliseconds timescale, and the other on a hundreds-of-milliseconds timescale. These relaxations partially characterize the dynamics of the p/t DNA construct as it undergoes stochastic transitions between its various 0-, 1- and 2-bound states. Nevertheless, information provided by the 2nd-order TCFs cannot by itself determine whether the states visited during a single-molecule trajectory occur independently, or if they are connected through a 'pathway' of correlated sequential events. One can imagine that a particular fluctuation must occur initially in order for a subsequent fluctuation to be possible. For example, the reaction scheme depicted in Fig. 3.1A illustrates a system of three coupled chemical species, in which the 0-bound state is connected to the 2-bound state through a single 1-bound intermediate. If the system is initially in the 0-bound state, one cannot observe a transition from the 1-bound to the 2-bound state without previously observing a transition from the 0-bound to the 1-bound state.

To distinguish between different chemical reaction schemes that describe the p(dT)$_{15}$-(gp32)$_2$ assembly process, we implemented a combined analysis based on 2nd- and 4th-order TCFs of microsecond smFRET trajectories. This analysis can, in principle, distinguish between models of varying levels of complexity. We define the 4th-order TCF according to

$$C^{(4)}(\tau_1, \tau_2, \tau_3) = \langle E_{FRET}(0) E_{FRET}(\tau_1) E_{FRET}(\tau_2) E_{FRET}(\tau_3) \rangle \qquad (3.2)$$

The 4th-order TCF is the product of four sequentially sampled data points from a smFRET trajectory, which are separated in time by the intervals τ_1, τ_2 and τ_3. The angle brackets in Eq. (3.2) have the same meaning as those in Eq. (3.1).

The 4th-order TCF contains information about the statistical weights of time-ordered events that occur within an individual smFRET trajectory. To illustrate this, we consider the three-state system of 0-, 1- and 2-bound states (see Fig. 3.1A), and assign to these the observable E_{FRET} values E_0, E_1 and E_2, respectively. The 4th-order TCF depends on the weighted observations of each possible time-ordered sequence of E_{FRET} values for a particular set of intervals. For example, we might observe the sequence $E_0E_1E_1E_2$ at the four times sampled. If we were to observe this sequence with greater statistical weight than ascribed to sequences containing E_0 followed by E_2, we might conclude that direct transitions between the 0-bound and 2-bound states are unlikely, and that the reaction must proceed through an intermediate 1-bound state. Because the timescales of transitions between states have well-defined values, certain sequences will be more likely to occur at short time intervals, while other sequences will become more prevalent at long time intervals. Thus, the information encoded in the 4th-order TCF provides direct insight into the reaction scheme that governs the time-ordered fluctuations of the system.

It is convenient to visualize the 4th-order TCF as a two-dimensional (2D) contour plot, with horizontal and vertical axes given by the variables τ_1 and τ_3, while the second time interval τ_2 (referred to as the waiting time) is held fixed. The 4th-order TCF describes the presence of correlation between transitions that occur during the interval τ_1 and those that occur during the interval τ_3. In the absence of 4th-order correlation, such temporally separated transitions are statistically independent. Examination of a series of 4th-order TCFs as a function of the waiting time τ_2 makes it possible to determine the average timescale over which pairs of successive transitions are correlated. In the limit that the waiting time τ_2 becomes very long, or that 4th-order correlations are short lived, we see from Eq. (3.2) that $\lim_{\tau_2 \to \infty} \langle E_{FRET}(0)E_{FRET}(\tau_1)E_{FRET}(\tau_2)E_{FRET}(\tau_3)\rangle = \langle E_{FRET}(0)E_{FRET}(\tau)\rangle^2$. In this limit, the 4th-order TCF is equal to the square product of the 2nd-order TCF defined by Eq. (3.1).

Even for relatively short waiting periods, the presence of 4th-order correlation must be detected above the level of 2nd-order 'background' correlation. It is useful to consider the 4th-order TCF in terms of the smFRET signal fluctuation: $\delta E_{FRET}(t) =$

$E_{FRET}(t) - \langle E_{FRET} \rangle$. We therefore reframe Eq. (3.2) in terms of the signal fluctuations

$$\bar{C}^{(4)}(\tau_1, \tau_2, \tau_3) = \langle \delta E_{FRET}(0) \delta E_{FRET}(\tau_1) \delta E_{FRET}(\tau_2) \delta E_{FRET}(\tau_3) \rangle - \bar{C}^{(2)}(\tau_1) \bar{C}^{(2)}(\tau_3) \tag{3.3}$$

where we have similarly reframed the 2nd-order TCF, in terms of signal fluctuations, according to $\bar{C}^{(2)}(\tau) = \langle \delta E_{FRET}(0) \delta E_{FRET}(\tau) \rangle = \langle E_{FRET}(0) E_{FRET}(\tau) \rangle - \langle E_{FRET} \rangle^2$. Eq. (3.3) decays with τ_2 from its maximum value $\langle \delta E_{FRET}(0) [\delta E_{FRET}(\tau_1)]^2 \delta E_{FRET}(\tau_3) \rangle$ to zero, over the time scales for which 4th-order correlation exists. We note that subtraction of the asymptotic ($\tau_2 \rightarrow \infty$) value $\bar{C}^{(2)}(\tau_1) \bar{C}^{(2)}(\tau_3)$ from the 4th-order TCF serves to isolate 4th-order correlations from 2nd-order 'background' correlation.

Both 2nd- and 4th-order TCFs can be modeled using the theory of Markov chains [20,36]. This approach assumes a kinetic scheme in which *N* states are interconnected by elementary chemical steps, as depicted in Figs. 3.1A and 3.1B. The input parameters of the Markov model are the observed fluctuation values δE_i assigned to state-i, and the rate constants k_{ij} associated with forward and backward transitions between state-i and state-j. The values of k_{ij} must satisfy the principles of detailed balance[20]. Expressions for calculated 2nd- and 4th-order TCFs are derived in the SI section. These expressions depend on: i) the equilibrium (time-independent) probability p_i^{eq} to observe the system in the *i*th state; and ii) the conditional probability $p_{ij}(\tau)$ that the system will be in the *j*th state at a time interval τ after it was initially observed in the *i*th state.

The 2nd-order TCF [see Eq. (S3.4)] is composed of $N - 1$ exponentially decaying terms, each with characteristic decay rates λ_1, λ_2, …, λ_{N-1}, and amplitudes $\mathcal{A}_1$, $\mathcal{A}_2$, …,$\mathcal{A}_{N-1}$. The Markov chain model expresses these parameters in terms of the rate constants k_{ij}.[20] The 4th-order TCF is composed of $(N-1)^2$ terms [see Eq. (S3.5)], each with an amplitude $\mathcal{A}_{n,m}$ $[n, m \in \{1, 2, \ldots, N-1\}]$ that depends on the rate constants and the waiting time τ_2. For a fixed waiting time, the decay of the 4th-order TCF occurs in two dimensions, corresponding to the time intervals τ_1 and τ_3. The characteristic decay rates of the 4th-order TCF are the same as those of the 2nd-order TCF. The $N - 1$ terms with amplitudes $\mathcal{A}_{n,n}$ are 'diagonal' terms, which each depend on a single decay constant λ_n. The terms with amplitudes $\mathcal{A}_{n,m}$ (with $n \neq m$) designate 'off-diagonal' coupling terms, which each depends on two decay constants, λ_n and λ_m. For an equilibrium system, the principles of detailed balance require that $\mathcal{A}_{n,m} = \mathcal{A}_{m,n}$.[20]

For situations in which relaxations that occur during the upstream interval τ_1 are uncorrelated to those that occur during the downstream interval τ_3, the coupling terms $\mathcal{A}_{n,m} = \mathcal{A}_{m,n}$ vanish. Under such circumstances, the functional form of the 4th-order TCF is simply related to that of the 2nd-order TCF. [20] This scenario is expected if intermediate 1-bound states were to dissociate, or to interconvert with other 1-bound states on a faster time scale than the experimental sampling resolution. If on the other hand the dynamics of 1-bound intermediates were experimentally resolved, the aforementioned dissociation and / or sliding processes can lead to negative or positive correlations between upstream and downstream signal fluctuations on these rapid time scales. The latter situation is expected to lead to either positive or negative 4th-order correlation between successive transitions, which is reflected by non-zero cross-term amplitudes, $\mathcal{A}_{n,m}$ and $\mathcal{A}_{m,n}$. For negative (or positive) off-diagonal amplitudes, the 2D surface of the 4th-order TCF can exhibit contours with convex (or concave) curvature.[20] Due to the presence of noise in experimental data, this curvature might be difficult to detect by visual inspection alone. Furthermore, the off-diagonal amplitudes may be small in comparison to those of the diagonal terms. Nevertheless, we find that a sufficiently large data set can, in principle, provide the information needed to extract these values numerically from the experimental data.

The information contained within 4th-order TCFs can be intuitively understood by representing these functions in the rate domain through an inverse Laplace transformation (ILT): $\bar{C}^{(4)}(\tau_1, \tau_3)]_{\tau_2\,fixed} \xrightarrow{\text{ILT},\tau_1,\tau_3} \hat{C}^{(4)}(k_1, k_3)]_{\tau_2\,fixed}$. In the rate domain, diagonal and off-diagonal terms that decay exponentially in time are represented as delta functions centered at values corresponding to the decay constants. These rate domain representations of the 4th-order TCFs are made in analogy to the frequency domain representations often used in 2D Fourier transform spectroscopies.[28,30] In our 2D rate domain spectra, diagonal features represent the characteristic relaxation rates, while the cross-peaks represent the presence of couplings between these relaxations.

To perform the ILT, we carried out an algorithm based on the Tikhonov regularization method.[38-40] We verified that our algorithm produced accurate results for the noise level of our data [see SI and Fig. S3.5]. In Fig. 3.5, we present 4th-order TCFs and the associated 2D rate spectra constructed from our experimental smFRET trajectories of the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA in the presence of 0.1 μM gp32. These and additional data for the 1.0 μM gp32 and 0.1 μM gp32 + 1 μM LAST peptide conditions are plotted for several values of the interval τ_2 in Fig. S3.6. We note that the decays of the 4th-order TCFs with respect to the intervals τ_1 and τ_3 occur on the same characteristic time scales as the 2nd-order TCFs (see Fig. 3.3 and Table S3.2). Additional information about the coupling between characteristic modes can be obtained from the associated 2D rate spectra. We performed the ILT of the 4th-order TCFs by using the information provided by the 2nd-order TCFs. Since the 2nd-order TCFs exhibited only two characteristic decay components ($\lambda_{slow} = \tau_{slow}^{-1}$, and $\lambda_{fast} = \tau_{fast}^{-1}$), we assumed a minimal model using an $N = 3$ scheme. We thus obtained the 2D rate spectra, as shown in Fig. 3.5 and Fig. S3.6. Each 2D rate spectrum exhibited two diagonal features, one centered at $(k_1, k_3) = (\lambda_{slow}, \lambda_{slow})$ and the other at $(k_1, k_3) = (\lambda_{fast}, \lambda_{fast})$. Off diagonal coupling features appeared at $(k_1, k_3) = (\lambda_{slow}, \lambda_{fast})$ and $(k_1, k_3) = (\lambda_{fast}, \lambda_{slow})$. The amplitudes of these components $A_{slow,slow}$, $A_{fast,fast}$ and $A_{slow,fast} = A_{fast,slow}$ varied as a function of the waiting time interval τ_2.

In Fig. 3.6, we plot the amplitudes of the rate domain peaks and cross-peaks as a function of the interval τ_2. We note that for all three sets of conditions, the amplitude of the faster characteristic rate constant $A_{fast,fast}$ decreased more rapidly with increasing τ_2 than the amplitude of the slower characteristic rate $A_{slow,slow}$. The fast and slow diagonal components exhibit very similar amplitudes and τ_2-dependent behaviors for the 0.1 μM gp32 data sets, both in the presence and absence of the LAST peptide (Figs. 3.6A and 3.6B). While the amplitudes of off-diagonal coupling features for our 0.1 μM gp32 data were negligibly small (see Fig. 3.6A), the amplitudes of these same features for both the 0.1 μM gp32 + 1.0 μM LAST and 1.0 μM gp32 samples decayed from finite positive values to zero on the characteristic time scales (see Fig. 3.6B and 3.6C, respectively).

The above results show that the fast and slow fluctuating modes we have observed in the p(dT)$_{15}$-(gp32)$_n$ system are interdependent processes. Moreover, the presence of the LAST peptide influences the degree of coupling between these processes, as we might expect due to its interference with the cooperative binding mechanism of the gp32 protein. In the remainder of this work, we devise a simple Markov chain model that accounts for our observations of the kinetic and thermodynamic behaviors of the system under the experimental conditions we studied.

Optimization of N = 4 Kinetic Scheme to Microsecond smFRET Data.

To take greater advantage of the information contained within our microsecond smFRET data, we considered kinetic schemes from which we could calculate histograms of E_{FRET} values, 2nd-order TCFs, and 4th-order TCFs. Through a direct comparison between these theoretical and experimental approaches we implemented a multi-parameter optimization to determine the model most consistent with our data. The $N = 3$ scheme discussed above (see Fig. 3.1A) is the simplest that can account for the known cooperative binding mechanism of the gp32 protein.[8,9,15,20] We deduced the existence of at least three states from our observation of two decay components in the 2nd-order TCFs. These three states likely represent the 0-,1- and 2-bound ssDNA-(gp32)$_n$ conformations. Nevertheless, as we show below, this basic $N = 3$ model could not adequately account for all aspects of our experimental results.

We therefore considered the $N = 4$ scheme shown in Fig. 3.1B, which represents an incrementally elevated level of complexity. As discussed above, the $N = 4$ scheme allows for two different categories of 1-bound intermediates, an unproductive intermediate (labeled state-1) and a productive intermediate (labeled state-1’). To simplify our model as much as possible, we assumed that the five 1-bound conformations corresponding to unproductive intermediates could be treated as experimentally indistinguishable. Likewise, we made this same assumption for the four 1-bound conformations that function as productive intermediates. Finally, we made the approximation that the E_{FRET} values corresponding to all 1-bound intermediates (productive and unproductive) were identically equal.

To compare specific kinetic schemes to our experimental data, we input 'fit parameters' – i.e., the forward and backward rate constants k_{ij}, k_{ji} $[i \neq j \in \{0,1,1',2\}]$ and E_{FRET} values $E_1 = E_{1'}$ – into Eq. (S3.3), whose solution provides the conditional probabilities $p_{ji}(\tau)$ and the equilibrium distribution of states, p_i^{eq}. Substitution of these solutions into Eqs. (S3.1) and (S3.2) allowed us to calculate the 2nd- and 4th-order TCFs, respectively, in addition to the E_{FRET} histograms. We constrained our solutions to ensure consistency with our previous findings, such that the observable values of the 0- and 2-bound states were E_0 = 0.81 and E_2 = 0.56, respectively. As described below, we generated initial values for the rate constants k_{ij} and the observable values $E_1 = E_{1'}$ to calculate the experimentally derived quantities. We then iteratively refined these values until we obtained optimal agreement with the experimentally derived functions. Further details of our multi-dimensional optimization procedure are given in the Appendix A Chapter III Supporting Information.

As mentioned previously, we initially attempted to fit our data to the N = 3 scheme depicted in Fig. 3.1A. For completeness, we considered the possibility that direct transitions between the unbound and doubly bound states are accounted for, although independent studies suggest that the gp32 protein cannot bind to the lattice directly as a dimer.[35] In Table S3.3, we list the optimized parameter values corresponding to the best fits we obtained using the N = 3 scheme. Although the 0.1 μM gp32 data could be reasonably well fit to the N = 3 model, the data collected under the 1.0 μM gp32 and 0.1 μM gp32 + 1.0 μM LAST conditions could not be similarly explained.

We understand the failure of the N = 3 model to describe our data as an indication that rapid interconversion between the nine possible singly bound conformations does not occur on time scales faster than the ~ 10 μs instrument resolution. This implies that singly bound intermediates must dissociate from the ssDNA lattice on a time scale much faster than it may 'slide' between different lattice positions. If all singly-bound sites are thermodynamically equivalent, a gp32 protein has a $5/9$ chance to initially bind to the ssDNA lattice at a nonproductive position, which precludes the cooperative binding of a second gp32 protein to the lattice. In this case, the time to undergo a transition from a singly-to-doubly bound state will be prohibitively long. Alternatively, a singly bound gp32 protein has a $4/9$ chance to initially bind at a productive position, which would provide room on the lattice for a second gp32 monomer to bind. The cooperative mode of binding would then result in a very rapid transition from the productive singly bound state to the doubly bound state. Because the $N = 3$ scheme does not allow for the simultaneously fast and slow association / dissociation pathways expected to occur in the presence of both productive and unproductive singly bound intermediates, this model cannot capture the analytical behavior of the experimentally derived functions under the various solution conditions.

The $N = 4$ scheme, on the other hand, can account for the presence of an assembly pathway that involves both productive and unproductive singly-bound intermediates (see Fig. 3.1B). In the $N = 4$ model, the unbound state can undergo a transition to the unproductive state-1, from which a direct transition to state-2 is not possible. Alternatively, the unbound state can also undergo a transition to the productive state-1', which may subsequently undergo a direct transition to state-2. For completeness, we allow for transitions between state-1 and state-1', which would account for sliding of the gp32 monomer along the ssDNA lattice.

Following the algorithm described above, we performed a search of the parameter space for the $N = 4$ scheme from which we obtained global solutions. For each of the experimental conditions we investigated, the solution is an optimized set of eight rate constants and a single (degenerate) FRET efficiency value for the two singly bound states. In Fig. 3.7, we present a comparison between the experimentally derived functions and their corresponding theoretical fits for the 0.1 μM gp32 conditions. In the SI section, we present full comparisons for the 0.1 μM gp32 condition (see Figs. S3.7 and S3.8), the 1.0 μM gp32 condition (see Figs. S3.9 and S3.10) and to the 0.1 μM gp32 + 1.0 μM LAST condition (see Figs. S3.11 and S3.12). The optimized values of the parameters used for these fits are given in Table 3.1. We note that for each of the three independently run optimizations, all resulted in very similar values of the FRET efficiencies, $E_1 = E_{1'} \approx 0.68$, for the singly bound states. For each set of conditions, we confirmed that the optimized values so obtained do indeed correspond to a global minimum of the multi-dimensional parameter surface (see SI section, Fig. S3.13).

To more closely examine the reasons that singly-bound intermediates are difficult to detect by inspection of smFRET trajectories that use the 30 msec time resolution of standard experiments, we performed computer simulations on these time scales. The simulations used as input the optimized rate constants and E_{FRET} values obtained from our $N = 4$ optimizations (listed in Table 3.1). We present the details of these calculations in the SI section. These trajectories were calculated using 1 msec time increments, and subsequently averaged to simulate single-molecule time traces with 30 msec resolution.

In Fig. S3.14, we present examples of our calculations for the three different experimental conditions investigated. For each case, the simulated trajectories appear to qualitatively reproduce the characteristic behavior of the experimental trajectories (compare Figs. S3.14A and S3.14B to Figs. 3.2A middle and bottom panels, respectively; and compare Fig. S3.14C to Fig. 3.4A, bottom panel). These favorable comparisons between simulated and experimental trajectories suggests that the four-state model can describe the dynamics of the ssDNA-(gp32)$_n$ system remarkably well. Nevertheless, the presence of the short-lived intermediates is not obvious by direct inspection of the trajectories alone. This comparison clearly shows that singly-bound intermediates are too short-lived to produce a distinguishable signal for experiments that use the standard 30 msec resolution. Thus, the sub-millisecond methods that we have demonstrated, in combination with our generalized TCF analysis, are necessary to detect and assign the roles of these short-lived intermediates.

Discussion

The results of our analysis permit us to derive some understanding of the basic assembly mechanism of the cooperatively bound T4 bacteriophage 3'-p(dT)$_{15}$-(gp32)$_2$–p/t DNA complex, while also providing new general insights into the ways that replication cofactors of the ssb type may function in replication, recombination and repair. As shown above, our experiments demonstrate that the system is well described by an $N = 4$ reaction scheme involving at least two categories of short-lived, singly-bound intermediates. These are the 'non-productive' and 'productive' states that we refer to as state-1 and state-1', respectively (see Fig. 3.1B). Moreover, these two types of singly-bound intermediates do not rapidly interconvert – i.e. a singly-bound gp32 monomer does not ‘slide’ between non-productive and productive binding sites. If this were not true, the two species in our experiments would be indistinguishable and, as a consequence, the system would have been well described by an $N = 3$ kinetic scheme. The lifetimes of singly-bound states are on the order of ~20 msec, and they decay either by dissociation to the unbound state-0 or react by the cooperative binding of a second gp32 protein to form a stable doubly-bound dimeric gp32 cluster (state-2). This finding is further corroborated by the prohibitively long exchange times that we find for the ‘sliding’ between non-productive and productive intermediates ($k_{11'}^{-1}$ and $k_{1'1}^{-1}$ ~ many seconds). While the very long times determined from our analysis (i.e., those greater than ~ 1 sec) are not quantitatively meaningful, we interpret these results to indicate that monomer sliding rates are at least several orders of magnitude slower than the dissociation rate.

The inverse rate constants (i.e. time constants) we determined for each of the three different solution conditions investigated are listed in Table 3.1. We compare the assembly dynamics seen at a relatively low protein concentration (0.1 μM gp32) to those observed at a relatively high protein concentration condition (1.0 μM gp32). We also compare the 0.1 μM gp32 concentration condition to the same gp32 concentration in the presence of 1.0 μM LAST peptide, which we have confirmed functions to inhibit the cooperative binding mode of the gp32 protein[37]. We see that the loading times k_{01}^{-1} and $k_{01'}^{-1}$ of the non-productive and productive singly-bound states scale inversely with concentration, as expected for a diffusion-limited reaction. On the other hand, the transition time $k_{1'2}^{-1}$ associated with the elementary dimerization step (state-1’ → state-2) is independent of concentration, and is on the order of ~20 msec. Once the singly-bound protein is situated at a productive binding site, the cooperative binding of a second gp32 to form a dimeric gp32 cluster appears to dominate the rate at which the dimerization step proceeds, consistent with the conclusions of prior bulk spectroscopic studies.[9]

We note that the value of $k_{1'2}^{-1}$ is slightly increased in the presence of the LAST peptide, as expected given that the peptide appears to function by interfering with the cooperative gp32-gp32 subunit binding mechanism. However, the presence of LAST has the most pronounced effect on the loading times of the non-productive and productive intermediates, k_{01}^{-1} and $k_{01'}^{-1}$, respectively. In the presence of LAST, we see that the loading time of the non-productive intermediate is decreased, while that of the productive intermediate is increased. We may understand this effect in the following way. Proteins bound to the competitive inhibitor will have their cooperative binding modality suppressed, while those proteins that are not bound to the LAST peptide will still participate in the cooperative binding step to form the doubly bound state-2. The overall effect is to reduce the population of proteins that can participate in the ‘productive assembly pathway,’ while increasing the population of proteins that can participate in the ‘non-productive assembly pathway.’ Because these loading steps are diffusion-limited, the transition time scale with these changes in the effective population of ‘functional’ gp32 proteins.

Although both productive and non-productive singly-bound states exhibit indistinguishable FRET efficiency values (the sole observable in our experiments), our generalized TCF analysis has allowed us to determine the dynamics of these states and to observe their presence individually. In all cases, the loading transition of the non-productive state occurred more rapidly than that of the productive state, which is consistent with the notion that there are more ways that the protein may bind to the $p(dT)_{15}$ lattice in a non-productive manner than in a manner leading to cooperative gp32 cluster formation.

In summary, our results allow us to deduce the mechanism by which the fully assembled $3'$-$p(dT)_{15}$-$(gp32)_2$–p/t DNA complex is formed. When a gp32 protein initially binds to a nonproductive ssDNA lattice site, it will dissociate rapidly (within ~20 msec). This permits the same (or a different) gp32 protein to undergo multiple binding events at various positions until a productive state is achieved that will allow for the adjacent binding of a second gp32 protein within the binding lifetime of the ‘productive singly-bound’ state. The cooperative binding modality of the protein leads to the relatively fast dimerization step, which also occurs on an ~20 msec time scale. The alternative mechanism considered, in which an initially bound gp32 protein binds at any position and then rapidly ‘slides’ along the ssDNA by means of a one-dimensional random walk process until it reaches a productive lattice site and there ‘trapped’ by a second gp32 monomer, is inconsistent with our analysis. Thus, the rate of sliding is too slow to compete with the unbinding-rebinding mechanism of singly-bound proteins under the conditions we examined in this study.

Conclusions

In this work, we have applied a novel analysis of microsecond smFRET experiments using a generalized TCF approach. These experiments have allowed us to extract detailed kinetic information about the assembly pathway of the T4 ssDNA binding protein gp32 with a 15-nt ssDNA template. We find that short-lived intermediates, which are otherwise difficult to detect using standard smFRET techniques (with ~30 msec resolution), can be clearly identified and studied. Indeed, the application of HMM analyses to the same system studied using standard smFRET methods could not be used to measure the kinetics of short-lived singly-bound intermediates, although their presence could be inferred indirectly.[15] The observation of singly-bound states requires the ability to detect signal fluctuations on microsecond time scales. Yet, such rapidly sampled smFRET experiments lead to sparse and noisy signal trajectories, which cannot be analyzed using HMM techniques. We anticipate that future microsecond-resolved experiments might be treated in similar ways to study individual steps within the biochemical reaction pathways of other complex biochemical networks.

It was recently confirmed by bulk solution studies using base analogue probes that 'clusters' of cooperatively bound gp32 proteins can slide along a ssDNA lattice[9] as inferred in earlier studies by Kowalczykowski and Lohman and their co-workers.[41,42] However, it was unknown whether this might also be the case for a singly-bound gp32 protein on a ssDNA template. The current study indicates that gp32 monomers do not slide along ssDNA, but instead dissociate within ~20 msec. Nucleation of the formation of cooperatively-bound clusters of gp32 proteins must therefore require that an initially bound gp32 monomer occupy a position that can accommodate the binding of a second gp32 protein. This difference in behavior between a singly-bound gp32 protein and a cluster of cooperatively bound proteins likely reflects the weak binding constant of gp32 in its non-cooperative binding mode. Upon the initial binding of a gp32 protein, a second gp32 monomer must immediately bind in order for the ternary complex to become stable. Otherwise, the initial singly-bound protein will rapidly dissociate. It is anticipated that multiple cooperatively-bound gp32 proteins must bind a ssDNA lattice with significantly higher affinity, so that the dissociation time constant of the assembled cluster increases, and sliding becomes a much more likely mechanism for cluster translocation to a new position on the ssDNA lattice. Other approaches will be needed to determine the limits of the lengths of gp32 clusters that can effectively translocate along ssDNA by sliding mechanisms.

Bridge to Chapter IV

We have established a binding mechanism that describes gp32 dimer assembly onto ssDNA templates near ss duplex junctions. Although this work used microsecond resolution smFRET measurements, the single-molecule fluorescence did not have enough signal-to-noise to facilitate sub-millisecond analysis. After various improvements to the optical geometry, we managed to dramatically increase the signal-to-noise of the single-molecule experiment which ultimately led to a series of microsecond-resolution measurements that could be analyzed at sub-millisecond timescales. In the subsequent chapter, we report on an analysis of four ssDNA p/t junction constructs that differ in length and polarity. We show that ssDNA has conformational fluctuations that occur on time scales spanning tens-of-microseconds to hundreds-of-milliseconds. Our results suggest that the ssDNA backbone of the p/t junction fluctuates between compact and extended conformations. These thermally driven fluctuations may play a role in the initial nucleation of gp32 monomers onto the ssDNA lattice.

CHAPTER IV

SUB-MILLISECOND CONFORMATION FLUCTUATIONS OF SINGLE-STRANDED DNA NEAR PRIMER-TEMPLATE (P/T) DNA JUNCTIONS BY MICROSECOND-RESOLVED SINGLE-MOLECULE FRET

The studies described in this chapter were performed by a number of lab members including myself, Claire Albrecht, Anson Dang, and Megan Barney. Claire Albrecht contributed substantially to this work by participating in the development of mathematics used to model the data. Anson Dang and Megan Barney contributed by assisting me in making samples and helping take single-molecule data. I was the primary author for this manuscript with Claire Albrecht being highly involved. Myself, Claire Albrecht, and Professors Andrew H. Marcus and Peter H. von Hippel have been involved in the drafting of the manuscript. This was work was supported by grants from the National Institutes of Health (molecular biology and biophysics training grant), the National Science Foundation (Chemistry of Life Processes Program grant CHE-1608915 to A.H.M.), and the National Institutes of Health (NIGMS Grant GM-15792 to P. H. v. H. and A.H.M.). Andrew H. Marcus was the principal investigator for this work.

Introduction

It is well known that the thermodynamically stable form of double-stranded (ds) DNA is the right-handed double-helix, which was famously deduced from the analysis of x-ray crystallographic studies.[1] In living cells, the genetic information is encoded as single-stranded (ss) DNA base sequences, which is often in complexation with proteins that serve to package and protect the information in the form of chromatin.[2] In order for gene sequences to be utilized by proteins that carry out biological functions, DNA must be able to transform to more accessible, albeit less stable, 'open' conformations. For example, when the two complementary strands of DNA separate enough to permit access to replication polymerases and helicases, the complex chemistry of DNA replication can occur. Duplex DNA fluctuates into such open states due to thermally-induced fluctuations of solvent molecules [3,4] a process termed DNA "breathing." The open DNA conformations exhibit contiguous sections of nucleotides, which are unconstrained by

Watson-Crick (W-C) hydrogen bonding. Such transiently exposed single-stranded (ss) DNA segments may undergo rapid conformational rearrangements between various conformational macrostates, which may themselves represent potential binding sites for DNA associating proteins.

Each macrostate contains many microstates, which account for all possible orientations in space and small perturbations to the structure. The free energy barriers between microstates are small relative to the barriers between macrostates, so the system can explore many microstates within a macrostate without significant change to the measured value of the observable. Note that the term macrostate and state are used interchangeably henceforth.

The direct interaction between proteins and ssDNA is a fundamental part of many biological processes. Besides being sites of DNA replication, ssDNA may also serve as templates for DNA polymerases.[5] The ssDNA molecule is a dynamic molecule with many degrees of freedom, in which thermal fluctuations cause it to sample various conformational states.[6] The exploration of conformational states is an essential part of what allows ssDNA to be available for protein binding.[7]

The T4 bacteriophage is a model system for DNA replication.[8] When transcription factors, necessary for cellular differentiation, or replication proteins, essential for genome duplication, approach a nucleic acid lattice, they will not bind unless the ssDNA is prepared to accommodate them. In order for the T4 bacteriophage ssb protein gp32 to bind to ssDNA, a contiguous sequence of seven nucleotide residues needs to be available[9] in a conformational macrostate that is suitable for binding. Before proceeding to investigate the dynamics of such protein binding events, it will be necessary to understand the subtleties of ssDNA alone.

To address this problem, we use the inverse problem approach to estimate 1) the number of conformational macrostates, and 2) the rates of interconversion for ssDNA between those states. An inverse problem is where you start with the measurement and apply a model to estimate the underlying parameters that must exist in order to produce the measurement.[10] Our goal is to construct a model that is the best approximation of the underlying polymer physics which the experimental single molecule data is derived from.

Here we report on a series of microsecond-resolved single-molecule Förster Resonance Energy Transfer (smFRET) studies of the backbone fluctuations of short single-stranded (ss) oligo- deoxythymidine [$o(dT)_n$] templates of varying length (n = 14 and 15) and polarity (3' versus 5') near ss—double-stranded (ds) DNA primer-template (p/t) DNA junctions. We analyzed our single-molecule data by calculating the 2nd- and 4th-order time correlation functions (TCFs) as well as the equilibrium distribution of conformational macrostates.

Our findings suggest a cyclical three-state model can be used to adequately describe these systems. We obtain quantitative agreement with our data by performing numerical optimizations of the transport master equation, from which we determine the rates of interconversion between conformational macrostates. Obtaining information about the nature of these functionally relevant DNA conformations, in addition to the time scales of their inter-conversion, is critical to understanding the detailed molecular mechanisms of protein-DNA interactions

The chapter is organized as follows: 1) the materials and single-molecule methods used to complete the experiment; 2) we review the theory of multi-order TCF analysis of smFRET data [11]; 3) an explanation of how to analyze the single molecule data; 4) we present our numerical results; 5) a discussion; and 6) and our conclusions and implications for the biophysics of protein-nucleic acid interactions; 7) lastly is the bridge to the next chapter.

Materials and Methods

DNA Primer-Template Constructs

We performed a series of phase sensitive, microsecond resolution single-molecule FRET experiments on four primer-template (p/t) DNA constructs that vary in length and polarity (Figure 4.1, found in Appendix A Chapter IV).

The p/t DNA constructs are labeled by the cyanine dyes Cy3 and Cy5, which act as the donor and acceptor of a FRET pair. The Förster length of the pair is 60 Å.[12] The average distance between base-pairs on ssDNA is 0.63 nm [13], giving the 14-nucleotide constructs a maximum contour-length of 8.8 nm, and the 15-nucleotide constructs a maximum end-to-end length of 9.5 nm.

The DNA constructs were purchased as dehydrated solids composed of nanomoles of individual single-strands (Integrated DNA Technologies *IDT*, Coralville, IA, USA), Table 4.1. The ssDNA solids are each rehydrated to reach 100 µM in a buffe*r* which consists of a similar ion concentration found in living cells: 100 mM NaCl and 6 mM $MgCl_2$. The buffer also contains 10 mM Tris buffer (pH 8.0) to reduce changes in pH.

The 100 µM ssDNA are annealed in house to form the DNA (p/t) junctions used in the experiment. To anneal the ssDNA, the 100 µM Cy3-containing strand is added to buffer to reach 150 nM concentration, and the complementary 100 µM Biotin/Cy5-containing strand is added to the same buffer to reach 100 nM concentration. The 1:1.5 molar ratio is such to ensure that each strand that is bound to the slide likely has a partner. The ssDNA strands are annealed together by heating the DNA constructs in the same vial together well above the constructs melting temperature (~60°C) at 94°C for 2 minutes, followed by gradual cooling to room temperature (22℃). The annealed p/t DNA constructs are stored at 4℃ until ready to use, or frozen at -4°C for extended periods of time.

Before the annealed DNA solution is added to the modified microscope slide (Figure 4.2), it is diluted 1000-fold to 100 pM in whatever condition solvent the experimenter is interested in. In this study all samples were prepared in the same buffer used for resuspension of the oligonucleotides.

The Microfluidic Sample Chamber

The microscope slide (Figure 4.2) has been chemically modified by a procedure described in Chandaross et. al.[14]. In short, the sample chambers have been thoroughly cleaned of contaminants with acetone, KOH, piranha-solutions (Nitric Acid + Hydrochloric Acid), and methanol. The piranha solution serves to both clean the surface of the slide, as well as to reduce the silicon-oxide to Si-OH. The reduced quartz-silica is treated with a methanol solution of 3-aminopropyl-trimethoxy silane + acetic acid, and subsequently pegylated (in a 0.1 M sodium-bicarbonate solution) to become less reactive to proteins. A small fraction of the PEG polymers are labeled with a biotin moiety. Neutravidin (Thermos Fisher) is a tetramer with a very strong affinity to biotin.[15] It can bind to the biotin attached to the slide as well as to the biotin-covalently attached to the ends of the p/t DNA constructs.

Oxygen Scavenging Buffer

An oxygen scavenging solution consisting of glucose oxidase, catalase and glucose are used to react with free oxygen in the solution. This solution is flowed into the sample cell before single-molecule data is taken. A series of chemical reactions result in the net loss of oxygen from the buffer.[16] Left unchecked, the dissolved oxygen in the system will rapidly react with the excited-state fluorophores and create a non-fluorescent molecule. When this occurs, the molecule is said to have photobleached.

The triplet state quencher Trolox (Sigma Aldrich) is also used to prevent photoblinking, the process where a chromophore reversibly enters the triplet state, for many milliseconds at a time.[17] Return to the ground state from the triplet is a spin-forbidden process, so fluorescence from this state is very slow. Keeping the molecule out of the triplet state avoids the molecule wasting time in states where it cannot absorb photons. The Trolox oxygen scavenging system (Trolox, oxidase, catalase, and glucose) is an efficient method to reduce the frequency of photobleaching.

Sub-Millisecond Single-Molecule FRET Spectroscopy

Single molecules have a fundamental limit to how much fluorescence they can emit in a given time. To perform microsecond-resolved single-molecule FRET microscopy, much care needs to be taken to ensure that this limited signal is captured in its entirely. Figure 4.3 shows the instrument set up we used to perform single-molecule FRET measurements.

Data was collected over a thirty-second interval for each molecule. We used total internal reflection fluorescence (TIRF) to excite molecules bound to a quartz slide. The microscope slide containing the sample chamber was attached a nanometer-precision computer-controlled stage (NPS-XY-100A with NPS3330 controller; Queensgate UK) mounted an inverted microscope (Nikon 2000 TE Eclipse). Using a prism-based TIRF geometry, a laser centered at 532-nm was focused with a 100 mm plano-convex lens onto a small portion of the sample centered over a 100x oil-based immersion objective. The size of the excitation is kept small to reduce the extent of photobleaching. The fluorescence from single molecules was sent though a 100 μm pinhole (Thor Labs, USA) mounted at the imaging plane directly downstream of the 200 mm tube lens. Next the fluorescence is collimated using a 75 mm biconvex lens placed its focal length away from the pinhole. The fluorescence is spectrally filtered into a donor and acceptor channel using a 635 nm dichroic beam splitter (Semrock, USA). The separated two-channel fluorescence is focused onto two avalanche photodiodes (APDs, SPCM AQR-13, Perkin-Elmer) using 60x Plano objectives (Olympus).

Single-molecule fluorescence data was acquired at a resolution of 1 μs and converted into 100 ns resolution post-acquisition by a phase-tagging procedure. We used a deconvolution procedure[18,19] to remove the instrument response function from our data at sub-microsecond resolution. The deconvolution procedure removed any artifacts that occur faster than two μs.

To ensure the molecules being imaged are not in too crowded of an environment and that the sample-chamber is scratch-free, we use an electron multiplying (em) CCD (iXon, Andor) to take wide-field images at 30 ms resolution. Figure 4.4 shows a typical image visible when the slide is viewed through an EMCCD. Each of the spots on the split-screen image is an individual DNA construct

Theory

Stochastic Model of a Fluctuating System

smFRET was used to measure the stochastic fluctuations between discrete conformational states of the single-stranded region of a p/t DNA construct. We apply a memory-less master equation, where the subsequent fluctuations between states only depends on the current state, as required for Markov state models. The ssDNA molecule can occupy a vast number of states, which are grouped into macrostates based on a common observable, in this case the FRET value.

The conformational disorder within each macrostate is reflected in the number of microstates it contains. In general, the more disorder that is within a state, the broader the distribution of observable values that will be associated with that state. The distribution of observable values reflects the probability of the system occupying that state. From this, we find the relative Boltzmann weight of each of our states is given by

$$\frac{P_i}{P_j} = e^{-\frac{G_i - G_j}{k_B T}} \tag{4.1}$$

where P_i is the integrated probability density of macrostate-i, G_i is the free energy of macrostatestate-i, k_B is Boltzmann's constant and T is the temperature.[20] The Boltzmann weight allows us to find the relative free energy of each macrostate sampled by single-stranded DNA molecules, using

$$\Delta G_{ij} = -k_B T \ln\left(\frac{P_i}{P_j}\right). \tag{4.2}$$

Thus, this analysis is sensitive to the relative free energy between macrostates ΔG_{ij}.

The activation barrier between states i and j is related to the observed rate of conformational transition between states, k_{ij}. By analyzing the dynamics of the system, we estimate transition rates between each state, giving us a sense of the relative height of the barriers that separate each pairs of states. These results provide an approximation of the energy landscapes that are explored within this set of ssDNA molecules.

Determination of the Probability Distribution of FRET States

The histogram $H(E)$ contains information about the distribution of FRET states E explored by the system. The experimentally measured distribution of FRET states is one of the three surfaces which we fit our model to. To construct $H(E)$ from smFRET data, the frequency of occurrence of each FRET state is plotted against the FRET value, as demonstrated in Figure 4.5.

To simulate the FRET distribution from a Markov model, we use equation 4.3 to calculate $H(E)$ as a sum of gaussian, one for each of the N conformational macrostates. Each state is centered at its FRET value, with an amplitude proportional to the probability of observing the system in that state given by

$$H(E) = \sum_{i=1}^{N} \frac{p_i^{eq}}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{E - E_i}{\sigma_i}\right)^2} \tag{4.3}$$

where σ_i is the standard deviation of the FRET values of state-i, E is the array of FRET values that the single-molecule data is binned into, and E_i is the expected FRET value of state-i. The equilibrium probability distribution, P^{eq}, is found by solving the transport master equation, explained in subsequent sections. The agreement between the simulated and measured $H(E)$ is a good indicator of the overall goodness-of-fit between our models and the experimental conditions.

2nd Order and 4th Order Time Correlation Functions (TCFs)

We analyze the time correlation functions of single molecule FRET trajectories to determine the number of macrostates, and the characteristic timescales of the system. Analysis of the TCFs ultimately yield the average rates of fluctuations between each of the N conformational macrostates. Experimental time correlation functions are calculated directly from single molecule FRET trajectories as described below. These experimentally derived correlation functions are compared to simulated correlation functions calculated with solutions to a transport master equation using a Markov model.

The two-point time correlation function contains information about the number of states, their connectivity, and the timescales of interconversion between each pair of states. It addresses the question: if a system is in state-i, what is the average likelihood that it will transition to any other state-j in time τ? Using angled brackets to imply an average over all combinations of points, we can write the experimental two-point time correlation function as

$$\bar{C}^{(2)}(\tau) = \langle\, \delta E(0)\, \delta E(\tau)\, \rangle \tag{4.4}$$

Where $\delta E(\tau) = E(\tau) - \langle E \rangle$. Statistical correlation functions are calculated for each molecule separately and averaged into a single correlation function which describes the ensemble of molecules for each data set. If the time series FRET data, $E(t)$, is continuous, one may use fast numerical algorithms to compute autocorrelation functions based on the Wiener-Khintchine theorem,

$$\bar{C}_E^{(2)}(\tau) = \frac{F^{-1}\left(F\left(E(t)\right)\times F\left(E(t)\right)^{*}\right)}{M^2}, \tag{4.5}$$

where M is the number of measurements in $E(t)$, $F\left(E(t)\right)$ denotes the fast-Fourier-transform of $E(t)$, F^{-1} denotes the inverse-fast-Fourier-transform, and $F\left(E(t)\right)^{*}$ is the complex-conjugate of $F\left(E(t)\right)$.

However, if the data series is not continuous, to calculate the autocorrelation function one must resort to a brute-force algorithm which steps through the data one measurement at a time. This is normally the case when the sampling frequency is much faster than the rate of fluorescence. Given sufficient signal-to-noise ratio reduction and high-resolution measurements, researchers will inevitably encounter such situations.

To calculate the time correlation function from sparse data, we account for the pseudorandom distribution of measurements in time, $N_{pairs}(\tau)$. Let t_{res} be the resolution of the trajectory in seconds and let τ_{max} be the maximum domain of the time correlation function. Starting at time $t = 0$, we calculate the product of all FRET pairs separated by a set of lag-times that span $\tau = 0$ to $\tau = \tau_{max}$. We write the experimental 2nd order time correlation function as

$$\bar{C}_E^{(2)}(\tau) = \frac{1}{N_{pairs}(\tau)} \prod_{t=0}^{M-1} \{\delta E(t) \delta E(t+\tau)\}, \tag{4.6}$$

where $N_{pairs}(\tau)$ is the exact number of measurements separated by τ. A consequence of the sparse sampling is that $N_{pairs}(\tau)$ is non-uniform across the domain. We normalize $\bar{C}_E^{(2)}(\tau)$ by explicitly accounting for this distribution of measurements. This procedure gives the same result as Eq. 4.5 in the limit of continuous measurement.

Given solutions to the master equation, we can write the analytical form of the 2nd order TCF as

$$\bar{C}^{(2)}(\tau) = \sum_{i,j=1}^{N} \delta E_j \, p_{ji}(\tau) \; \delta E_i \, p_i^{eq} \tag{4.7}$$

Where p_i^{eq} is the equilibrium probability of state-i. δE is the fluctuation of the FRET observable given by $\delta E_i = E_i - \sum p_i^{eq} E_i$, τ is the lag-time between measurements, $p_{ji}(\tau)$ is the conditional probability that the system transitions to state-j from state-i in the time interval τ.

While two-point correlation functions tell us about the average fluctuation timescale between any two states in the trajectory, higher order time correlation functions reveal additional information. We use the higher order time correlation functions to determine the pathways the system uses to go from one state to another, and the relative probabilities with which each pathway is used. Here, we use four-point time correlation functions to probe the question: on average, if we observe a transition from state-i to state-j how likely are we to observe a transition from state-k to state-l after an interval τ_2, for all possible combination of states? The experimental four-point time correlation function reveals the timescales of fluctuations across four measurements and can be written as

$$\bar{C}^{(4)}(\tau_1,\tau_2,\tau_3) = \langle \delta E(\tau_3)\, \delta E(\tau_2)\, \delta E(\tau_1)\, \delta E(0)\rangle \qquad (4.8)$$

where $\tau_1 = t_2 - t_1,\ \tau_2 = t_3 - t_2$, and $\tau_3 = t_4 - t_3$.

Like the 2nd order TCF, this product needs to account for the random distribution of measurements using the following formula

$$\bar{C}_E^{(4)}(\tau_1,\tau_2,\tau_3) = \frac{1}{N_{pairs}(\tau_1,\tau_2,\tau_3)} \prod_{t=0}^{M-1}\{\delta E(t)\delta E(t+\tau_1)\delta E(t+\tau_1+\tau_2)\delta E(t+\tau_1+\tau_2+\tau_3)\} \qquad (4.9)$$

where $N_{pairs}(\tau_1,\tau_2,\tau_3)$is the exact number of measurements separated by τ_1, τ_2, and τ_3. Again, once we have solutions to the transport master equation we can simulate the 4th order time correlation function with

$$\bar{C}^{(4)}(\tau_1,\tau_2,\tau_3) \ = \sum_{i,j,k,l=1}^{N} \delta E_l\, p_{lk}(\tau_3)\, \delta E_k\, p_{kj}(\tau_2)\, \delta E_j\, p_{ji}(\tau_1)\, \delta E_i p_i^{eq} \qquad (4.10)$$

where $\tau_1 = t_2 - t_1,\ \tau_2 = t_3 - t_2$, and $\tau_3 = t_4 - t_3$.

The experimental 2-point and 4-point TCFs do not fully decay to zero, but rather asymptotically approach it. To account for this in our model, we add an experimentally determined y- and z- offset to the simulated $C^{(2)}$ and $C^{(4)}$, respectively. The value used for the y- and z-offsets are the absolute value of the mean of the last 3 measurements of each surface, when the decay is essentially complete.

To calculate the theoretical conditional probabilities for our system, $p_{ji}(\tau)$, we borrow a well-known tool from chemical kinetics: the transport master equation.

Transport Master Equation

The transport master equation describes how the population in each state, $p_i(t)$**,** changes with time based on the connectivity and relative populations of the states. We write this expression as

$$\dot{\mathbf{p}}(\mathrm{t}) = \mathrm{K}\, \mathbf{p}(\mathbf{t}) \qquad (4.11)$$

where K is the rate matrix (Eq. 4.12), and $\dot{\boldsymbol{p}}(t)$ is the time-derivative of the population state vector. The solutions to this equation are the state-to-state conditional probabilities and equilibrium probabilities that are required to simulate the time correlation functions.

For a system with N-states, we write an $N \times\ N$ rate matrix, K, containing the rate constants k_{ij} (Eq. 4.12). The rate constants are inversely proportional to the time required for interconversion between macrostates, e.g. k_{ij} is the rate constant between state-i and state-j. $K_{ij} = k_{ji}$ if a transition between states-i and-j exist. If the model has no connection between states i and j, k_{ij} is set to zero. The diagonal elements correspond to the rate of population leaving each state, and the off diagonal elements describe population entering each state. Ultimately, determining the values of the rate constants will provide information about the energy landscape of the molecule.

$$K = \begin{pmatrix} -\sum_{i=1}^{N} k_{1,i} & k_{2,1} & \cdots & k_{N,1} \\ k_{1,2} & -\sum_{i=1}^{N} k_{2,i} & \ddots & \vdots \\ \vdots & \ddots & \ddots & k_{N,N-1} \\ k_{1,N} & \cdots & k_{N-1,N} & -\sum_{i=1}^{N} k_{N,i} \end{pmatrix} \tag{4.12}$$

After building the K matrix for a chosen model, we find the set of scalar eigenvalues $\{\lambda\}$ and corresponding eigenvectors $\{\boldsymbol{v}\}$ by solving the eigenvalue equation $K\boldsymbol{v} = \lambda\boldsymbol{v}$. The eigenvalues of K are the characteristic timescales of the system and the eigenvectors describe the normal modes of the dynamics. For an N-dimensional matrix, there are N non-trivial eigenvectors, each with a corresponding eigenvalue. Since the sum of each column of the rate matrix is zero, one eigenvalue is always zero and the rest are negative [21].

We use the following general solution for a set of coupled, first-order, linear, ordinary differential equations

$$\boldsymbol{p_i}(t) = c_1^i \boldsymbol{v}_1 e^{\lambda_1 t} + c_2^i \boldsymbol{v}_2 e^{\lambda_2 t} + \cdots + c_N^i \boldsymbol{v}_N e^{\lambda_N t} \tag{4.13}$$

where $\boldsymbol{p_i}(t)$ is the solution to the rate equation for each initial condition and c_n^i are the expansion coefficients, indexed by i corresponding to the initial condition [22].

The equilibrium population of state-i, $\boldsymbol{p}_i^{eq}$, is given by $c_1^i \boldsymbol{v}_1^i$, where $\boldsymbol{v}_1^i$ is the eigenvector associated with the eigenvalue $\lambda_1 = 0$, so we can rewrite Eq. 4.13 as

$$\boldsymbol{p_i}(t) = \boldsymbol{p}^{eq} + c_2^i \boldsymbol{v}_2 e^{\lambda_2 t} + \cdots + c_N^i \boldsymbol{v}_N e^{\lambda_N t}. \tag{4.14}$$

Using the initial conditions, we solve the system of equations to calculate the expansion coefficients. This gives the conditional probabilities $\boldsymbol{p}_{ji}(t)$ that describe the probability of transitioning from state-i to state-j in time t.

$$p_{ji}(t) = p_j^{eq} + c_2^i v_2^j e^{\lambda_2 t} + \cdots + c_N^i v_N^j e^{\lambda_N t} \tag{4.15}$$

Here, we see that the expansion coefficient, c, depends on the initial condition and the eigenvector component, v, depends on the final condition, indexed by i and j, respectively. Given that all $\lambda_n \leq 0$, note that as $t \rightarrow \infty$, or $t \gg t_{max}$ where t_{max} is the longest relaxation time in the system, the conditional probability $p_{ji}(t)$ evaluates to p_j^{eq}. In other words, if you wait long enough, the probability of going from state-i to state-j is simply the equilibrium population of state-j.

If the system is relatively simple, $p_{ji}(t)$ can be solved as a set of linear, coupled equations for the expansion coefficients; however, as the system grows increasingly complex, it becomes infeasible to solve these equations analytically. For large models, using numerical techniques decreases computation time. Instead of solving a large system of coupled equations, we can transform the rate matrix to its eigen-basis and solve for the conditional probabilities directly. Appendix C, Chapter IV contains the derivation of solving the master equation with numerical methods. In short, we can write

$$P(t) = U\, e^{\Lambda\, t}\, U^{-1}\, P_0 \tag{4.16}$$

Where U is the modal matrix composed of the eigenvectors of K, $e^{\Lambda\, t}$ is the exponential of the spectral decomposition of K (i.e. the diagonal elements are functions of the N eigenvalues, $e^{\lambda_n t}$), and P_0 the $N \times N$ identity matrix.

At this point, the framework for simulating time correlation functions of a chosen model is complete. We now discuss the specifics of the ssDNA system in the context of this framework.

Results

Fits to the 2nd Order Time Correlation Functions

The 2nd order time correlation function of each construct exhibited a decay on the order of several microseconds as shown in Figure 4.6. We expect this relaxation timescale is photophysical in nature related to the timescale of cis/trans isomerization of the fluorophores[23,24] and transfer to the triplet state[25]. We set out to learn about the nature of conformational fluctuations, not photophysics. Therefore, we started fitting the autocorrelation function beginning at 20 microseconds, past the time of photophysics.

Analysis of the 2nd order time correlation functions beyond 20 μs revealed the presence of at least two eigenmodes (Figure 4.7) in each construct. This implies that there are at least three states that the DNA is interconverting between on those timescales. See Appendix C, Chapter IV part 2: Exponential form of the TCFs, for an explanation of why this is true. Therefore, we assumed the simplest model that could explain our data consists of three interconverting states.

In addition to the time correlation function indicating three conformational macrostates, the distribution of FRET states (Figure 4.5) also suggest that a single mode will be unable to capture the full range of polymer conformations.

Model for conformation fluctuations of single-stranded DNA

For this experiment, we have built a model with three macrostates: a compact state, an extended state and an intermediate (Figure 4.8). The compact macrostate features the ssDNA tightly coiled or folded, allowing the dye molecules to be close together, yielding a high FRET efficiency. The extended macrostate features the ssDNA stretched out, yielding a lower FRET efficiency. And the intermediate state is a partially compact structure, giving a FRET value between the extended and compact states.

We propose that each state can interconvert with the other two states, resulting in six rate constants in the model (Figure 4.9). However, since the system features a closed loop, it must obey detailed balance, which requires that one rate depends on all the others; this leaves five free rate parameters [26].

$$k_{32} = \frac{k_{31}\, k_{12}\, k_{23}}{k_{21}\ k_{13}} \tag{4.18}$$

In addition to the unknown rates, we also need to determine the FRET value E_i for each macrostate; giving a total of eight free parameters in our model.

Now that we have built a model, we construct the corresponding rate matrix, K

$$K = \begin{pmatrix} -(k_{12}+k_{13}) & k_{21} & k_{31} \\ k_{12} & -(k_{21}+k_{23}) & k_{32} \\ k_{13} & k_{23} & -(k_{31}+k_{32}) \end{pmatrix} \tag{4.19}$$

according to Eq. 4.12.

The magnitudes of the eigenvalues of K are the timescales of the normal modes of the rate matrix. These timescales directly related to the systems dynamics. The eigenvalues and eigenvectors of K are used in Eq. 4.15 to solve for the state-to-state conditional probabilities, $P_{ji}(t)$. With the $P_{ji}(t)$ vector solved for, we simulate the two-point and four-point time correlation functions making use of Eq. 4.7 and Eq. 4.10, respectively. Figure 4.10a shows a set of simulated two-point correlation functions for randomly chosen sets of rate constants for the three-state model. Simulated examples of the four-point correlation function for different inputs are shown in Figure 4.10b. These plots show that the correlation functions are sensitive to the input parameters.

The sensitivity of the simulated two-point and four-point correlation functions is essential to our analysis of the rate constants between our macrostates. To proceed, we solve a version of the inverse problem. First, we construct simulated correlation functions based on a guessed set of input rate constants. Then we vary the set of input rates until we have minimized the difference between the simulated and experimental correlation functions using Eq. 4.20. The set of rates that create correlation functions that most resemble the data are our estimates for the rate constants of interconversion between macrostates.

How to Determine Model Parameters

To find the optimal set of model parameters, we performed a multi-dimensional optimization procedure, similar to the one outlined by Phelps et al.[11,27]. The purpose of the optimization is to find the set of input parameters that best reproduce the experimentally measured FRET probability distributions and time correlation functions.

We introduce the weighted error function χ^2 as a metric to assess the degree to which our models can reproduce statistics from experimentally measured data. The error function is defined to be the cumulative squared-difference between the measured quantities y and the predicted ones g,

$$\chi^2 \ = \ \sum_{q=1}^{3} w_q \sum_i \left(y_i - g_i(\{k_{lm}, A_l\})\right)^2 \tag{4.20}$$

where q enumerates the histogram, 2nd order TCF and the 4th order, and i is the domain index for each function. The weighting coefficients, w_n, are chosen such that the histogram, and time correlation functions contribute approximately equal amounts to the error function. Additionally, w_n is a function of lag times τ, such that more weight is given to the early times of the time correlation function.

The parameters of the model are: (1) the elements of the rate matrix, i.e. the set of rates $\{k_{ij}\}$ that describe the inverse transfer-time between states i and j, and (2) the values of the FRET states $\{A_i\}$. A genetic algorithm (GA) was used to search for an *approximate* solution to the least-squares fitting problem by finding a set of parameters that produced a low value of χ^2. A schematic demonstrating how the GA optimization procedure works is shown in Figure 4.11. Next, a constrained nonlinear multivariable problem solver, *patternsearch* from MATLAB (The MathWorks, USA), is used to minimize the error function Eq. (4.20) in order to find the global solutions reported in Table 4.3.

Best Fits to the Data

The simulations in Figure 4.12 show the best fit to the FRET histograms. The simulated lines were calculated with Eq. (4.7). Figure 4.13 and Figure 4.14 are the best fits to 2nd- and 4th- order time correlation functions, calculated with Eq. (4.7) and Eq. (4.11), respectively. An error analysis was conducted on each of the free model parameters (Figure 4.15). From the error analysis we see that if any of the optimized parameters are increased or decreased from their nominal values, the overall error function will increase.

Conditional Probabilities

The conditional probability $p_{ji}(t)$, given by Eq. (4.15), contains the underlying time dependence of the time correlation functions. These functions describe the average survival time of each macrostate as well as the relative likelihood of possible transitions. The survival probability of state-i is the probability of remaining in state-i at time t assuming the system was in state-i at time $t = 0$. Figure 4.16 shows the N^2 conditional probabilities for the $3'$-p(dT)$_{15}$ p/t DNA construct.

Despite the ssDNA being composed of poly-pyrimidines, which have minimal stacking interactions[28], we speculate that when the ssDNA backbone is in the extended conformation, the nucleotide bases are more stacked than they are in the other conformational states. The relatively long survival time of this conformational macrostate may be related to the enthalpic cost of unstacking the bases.

The partially- and fully-extended macrostates may be important intermediate states in protein binding mechanisms. Proteins can bind to these transiently exposed conformations which have enough room to accommodate one or more protein-monomers. Once a protein has bound to the ssDNA lattice, the overall free energy of the system has changed. In the absence of a binding event, we can describe the energetics of a system using a free energy surface.

Free Energy Surface

The free energy surface (FES) is a multidimensional representation the energy versus the various conformations of a molecule. The shape of the surface determines the rate of transfer between various conformational states, which are themselves determined by the minimal values of the FES. Measurements sensitive to the molecular conformation inform on what this surface looks like. A deep understanding of the FES enables the creation of detailed mechanistic models.

To illustrate the free energy landscape of ssDNA near a p/t junction (Figure 4.17), we calculated G_i, the free energy of each macrostate-i, using

$$G_i = -k_B T ln\left(P_i^{eq}\right), \tag{4.21}$$

Where k_B is Boltzmann's constant, T is the temperature, and P_i^{eq} is the equilibrium population of state-i. The most stable state is the compact-macrostate, followed by the intermediate-conformation and lastly the extended conformation state. All of the states in the FES are accessible by stochastic fluctuations at room temperature.

Principal Component Analysis

To quantify the variability between the data sets, we performed a principle component analysis (PCA) on the results presented in Table 4.3. A PCA illustrates the correlation between datasets by reducing their dimensionality. MATLAB's (MathWorks USA) PCA function we arrived at Figure 4.18, a visual representation of where each of the four constructs lie along the first two principal components.

The results of the PCA indicate that there is not much variability between the data sets. The principal component that explains by far most of the variability is the first principal component, Component 1. All the data sets lie on the same side of Component 1, with the biggest outlier being the $5' - p(dT)_{15}$ construct, as one might expect by examining the results in Table 4.3.

Discussion

Our results suggest that ssDNA near a primer-template junction is a dynamic system with several distinct modes whose time-scales span four orders of magnitude, from tens of microseconds to hundreds of milliseconds. This is in opposition with our previous assumptions of this system, where we had assigned a single conformational state to ssDNA.[27]

Using a three state Markov chain model, we simulated correlation functions and FRET probability distributions. Then we minimized the difference between the simulated and experimental surfaces to obtain a set of model parameters that represent the system. Our error calculations (Figure 4.15) demonstrate that our optimization has found a local minimum.

The optimized parameters averaged across the constructs contain a high FRET value ($\langle E_1 \rangle = 0.73 \pm 0.08$), a low value ($\langle E_3 \rangle = 0.18 \pm 0.07$), and an intermediate value ($\langle E_2 \rangle = 0.42 \pm 0.09$) (Table 4.3), which correspond to the compact, extended and intermediate states, respectively. The time for interconversion between states, i.e. the inverse of the rate constants, span several orders of magnitude (table 1). The fastest process (~20 μs) is the interconversion from the intermediate state to the compact state. This rapid rate suggests a low energy barrier separates these macrostates. The longest conversion times (hundreds of milliseconds) are the transitions from the compact state to the extended state, suggesting a relatively large concerted rearrangement of the ssDNA.

Overall, our results suggest ssDNA near a (p/t) junction tends to exist in the compact form and interchange frequently with the intermediate state. To reach the extended conformation, it typically proceeds through the intermediate state.

We found that the average rate of fluctuation from the partially-extended state to the compact state across the four DNA constructs was 23 μs, commensurate with the values found for ssDNA loop closure in the formation of a DNA hairpin composed of a $p(dT)_{16}$ oligo- deoxythymidine loop [29,30].

The conditional probabilities reflect the timescales over which one state transfers to another state (Figure 4.16). The conditional probabilities suggest if the system is in the extended state its lifetime is hundreds of microseconds, whereas the system is likely to have left the intermediate state within tens of microseconds. Generally, the lifetime of the survival probability is dominated by the fastest rate out of that state.

Using Eq. (4.21) we calculated the *relative* free energy for each state using the equilibrium probability values P_i^{eq} (Figure 4.17). These calculations are consistent with our analysis of the rates: the rate constants to jump downhill in energy are higher than to jump uphill. The transfers between states close together in free energy are more rapid than transitions between states with a larger energy difference.

From the principle component analysis (Figure 4.18) we see that the constructs with the same polarity are more similar than those with the same length. Overall, the PCA analysis suggests that there is very little variation between constructs.

Conclusion

We have determined that ssDNA templates near DNA (p/t) junctions exist in primarily three conformational macrostates: compact, intermediate-extension, and fully-extended. Fluctuations of the ssDNA backbone among these conformational macrostates occur on a range of time scales from tens-of-microseconds to hundreds-of-milliseconds. Our findings indicate that the polarity of the ssDNA strand near a (p/t) junction makes little difference in the timescales of conformational changes between ssDNA conformational macrostates.

The ssDNA is most likely to be in the compact state, which is reflected in our calculation of the compact state having the lowest free energy. The partially- and fully-extended conformational macrostates are likely the principle partners for protein-DNA binding interactions. Our results indicate that single-stranded polynucleotides in the vicinity of dsDNA can exist in a conformation amenable to protein binding upwards of a millisecond at a time.

Bridge to Chapter V

The T4 Bacteriophage ssb, gp32, has a binding site of seven contiguous nucleotides.[31,32] In order for ssb gp32 to bind, the ssDNA needs to be in a conformation with a sequence of seven exposed nucleotide residues. The next chapter explores modeling the gp32 dimer assembly mechanism with new insights from this study. Our results indicate that the ssDNA templates exist primarily as three distinct macrostate conformations: i) a majority-component ‘compact’ macrostate, ii) an intermediate-component ‘partially-extended’ macrostate, and iii) a minority-component ‘fully-extended’ macrostate. These macrostates can function as preferential binding sites for DNA associating proteins, such as the single-stranded DNA binding proteins (ssb). When the ssDNA fluctuates into a state that is amenable to binding by a protein, the protein can 'trap' the ssDNA into this configuration for extended periods of time [27]. We expect the partially- and extended-conformations are able to accommodate the initial binding of gp32, while the compact form of ssDNA will not be on path. Therefore, the models of gp32 dimer assembly presented in Chapter III can be revised to include multiple DNA states and become a more accurate depiction of mechanism of gp32 dimer assembly.

CHAPTER V

THE EFFECTS OF LENGTH AND POLARITY OF SSDNA ON THE ASSEMBLY MECHANISM OF DIMERS OF T4 BACTERIOPHAGE SSB GP32 ONTO SSDNA NEAR PRIMER-TEMPLATE DNA JUNCTIONS

The work described in this chapter was performed by myself, Claire Albrecht, Anson Dang, and Megan Barney. Anson Dang and Megan Barney contributed by assisting me in making samples and helping take single-molecule data. Claire Albrecht assisted with the modeling of the data. I was the primary author for this chapter. Claire Albrecht, Andrew H. Marcus and Peter H. von Hippel and I are involved in the drafting of the manuscript based on this work. These studies were supported by grants from the National Institutes of Health (molecular biology and biophysics training grant), the National Science Foundation (Chemistry of Life Processes Program grant CHE-1608915 to A.H.M.), and the National Institutes of Health (NIGMS Grant GM-15792 to P.H.v.H. and A.H.M.). Andrew H. Marcus was the principal investigator for this work.

Introduction

Single-stranded (ss) DNA binding proteins (ssb) bind to the ssDNA phosphate backbone in order to protect it from nucleases, melt unfavorable secondary structure, and configure it for interaction with other proteins[1]. The T4 bacteriophage ssb is the gene-protein 32 (gp32). It has a binding site footprint equal to a size of seven nucleotides [2]. Gp32 is made of three basic domains: the core of the protein which binds to ssDNA[3,4]; the C-terminal region, a negatively charged intrinsically disordered domain that binds to the gp32 core and protects it[5]; and the N-terminal domain of gp32 that is responsible for gp32 cooperativity[6,7]. To study the assembly mechanism of a dimer of gp32 onto ssDNA near a ssDNA-dsDNA junction we performed sub-millisecond single molecule FRET spectroscopy using four different DNA constructs differing in length and polarity (Figure 5.1).

The gp32 protein is an important component of DNA replication in T4 bacteriophage[1], and as such must bind to and unbind from the DNA as the replication fork advances, in addition to forming cooperatively bound assemblies that may slide along the ssDNA intermediates as the replisome advances.[8] The 15-nucleotide constructs are long enough to investigate some aspects of mechanisms and rates at which dimer clusters can slide. By studying different ssDNA polarities, we can learn about intrinsic differences in binding kinetics on the leading- versus the lagging-DNA strand, as well as probing the assembly and disassembly mechanism of dimers at the ss-dsDNA junction.

Materials and Methods

DNA Primer-Template Constructs

The four DNA (p/t) constructs were purchased as ssDNA from Integrated DNA Technologies (IDT, Coralville, IA, USA). First, the solid ssDNA pellets are each resuspended in a buffer (100 mM NaCl, 6 mM $MgCl_2$, 10 mM Tris-HCl pH 8.0) to reach a relatively high concentration of 100 μM. Next, complementary ssDNA oligonucleotides are hybridized to form the DNA constructs shown in Figure 5.1 (see Appendix A Chapter V for figures). The oligomers are not mixed at added at equal molar amounts: the Cy3-donor labeled oligo is added in excess at 150 nM concentration and the acceptor-labeled, biotin-tagged oligo is added at 100 nM concentration. As only the biotin-labeled ssDNA can bind to the microscope-slide, mixing the ssDNA strands with excess non-biotinylated strand will ensure that each construct bound to the slide is a dsDNA construct. The dilute (100 nM) ssDNA mixture is heated to 94° C for 2 minutes, and then allowed to slowly come to room temperature. The sequences of the strands are shown in Table 5.1 (see Appendix B Chapter IV for tables).

Details of the Sample Cell Chamber and Preparation

Single-molecule FRET spectroscopic measurements were taken on the p/t DNA constructs using microfluidic sample chambers (Figure 5.2). The structural elements of the chambers include a chemically modified quartz slide, a glass coverslip, binding epoxy, plastic tubing, and pipette tips. Once the chamber is constructed, the DNA constructs are immobilized on the slide using biotin-neutravidin chemistry. Finally, the Trolox oxygen scavenging solution is flowed into the chamber (see Chapter IV Materials and Methods for details). Control measurements are taken on the constructs before protein is added to the chamber. The solution is allowed to equilibrate for at least 10 minutes before further measurements are performed.

Microsecond-Resolved Single-Molecule FRET Spectroscopy

Figure 5.3 shows the instrument set up we used to perform single-molecule FRET measurements. Care was taken to ensure a minimum of environmental noise was present so that the signal at the detectors is as optimal as possible. This entails using all means possible to block background light and optimizing all the optics in the path length.

An Eclipse TE2000 inverted microscope (Nikon) was used to image individual molecules attached to a quartz slide using prism-based TIRF microscopy. The microscope slide containing the sample chamber was attached a nanometer-precision computer-controlled stage (NPS-XY-100A with NPS3330 controller; Queensgate UK) controlled by a custom program written in LabVIEW. A 50 mW continuous-wave laser centered at 532-nm (Coherent, Compass model no 215M) with approximately 10 mW power at the prism was focused onto our sample using a 100 mm plano-convex lens. The illuminated area is kept to a minimum to avoid photobleaching of nearby molecules not being immediately imaged.

Before the incident light hits the sample-chamber containing the DNA samples, it passes through a Pellin-Broca prism which changes the angle of the light at the air-quartz interface according to Snell's law, $n_1\theta_1 = n_2\theta_2$ where n is the index of refraction of refraction of the transmissive medium, and θ is the angle between the incident light and a line normal to the surface. Care must be made to ensure the angle of the prism is accounted for in determining the final angle at the quartz-aqueous solution interface to achieve total internal reflection. The angle of incidence between the excitation beam at the quartz-water interface is set just beyond the critical angle, $\theta_c = 65.8°$, such that the incident light is totally internally reflected and an evanescent field is produced. This electric field penetrates the sample chamber enough to excite the molecules of interest attached near the quartz surface, but does not extend far enough into the solution to generate significant background from out of focus molecules scattering the excitation beam.

A 100x infinity-corrected Plan Apo oil immersion objective (Nikon) with a numerical aperture of 1.40 was used to capture the fluorescence from the DNA constructs. The infinity-corrected objective sends a collimated image to the 200 mm tube lens of the inverted microscope, which further magnifies the molecule by a factor of 1.5

and forms a focused image at the image plane. That image contains the focused fluorescence of many molecules (~50-100 constructs) that are widely enough separated so that the point-spread-functions of individual constructs do not overlap. The diffraction limit dictates that the molecules will appear no smaller than $\lambda/(2 \times NA)$, where λ is the wavelength of emission and NA is the numerical aperture.[9] Using the emission maximum of Cy5, $\lambda = 650\ nm$, this makes the minimum size of the molecules in our system approximately 232 nm. After the 150x magnification, the size is about 35 microns at the image plane. A 100 μm pinhole (Thor Labs) placed at the image plane is used to block out the fluorescence from all molecules except the one in the center of the field-of-view. The size of the pinhole was chosen to maximize signal-to-noise (S/N) by maximizing the transmission from one central molecule, but minimizing chances of capturing fluorescence from non-interacting, neighboring molecules.

The surviving fluorescence after the pinhole is collimated using a 75 mm biconvex lens placed its focal length from the pinhole. It is important to collimate the emission to avoid loss of signal on the way to the detectors. A longpass Di03-R635 dichroic beamsplitter (Semrock) was used to separate the fluorescence from the Cy3/Cy5 chromophores at 635 nm to minimize the bleed-through of the donor into the acceptor channel. The longer wavelength acceptor emission passes though the dichroic while the shorter wavelength donor emission is reflected. Finally, the spectrally separated fluorescence from the DNA constructs is focused by two 60x objectives (Olympus) onto nanosecond-resolution photon-counting avalanche photo diodes (APDs, SPCM-AQR-16, Perkin-Elmer). The APDs relay the signal to a data acquisition (DAQ) board (National Instruments, NI PCIe-6535) that records the arrival time of the photon and passes this information on to our data storage computer for later analysis. Single molecule trajectories are recorded for each molecule for 30 seconds.

Results

The experimental 2nd-order time correlation functions, $C^{(2)}$, of the single-molecule FRET data of the (ss) oligo- deoxythymidine [o(dT)$_n$] templates (n = 14 and 15) and polarity (3' versus 5') near ss—dsDNA junctions are shown in Figure 5.4.

Fitting the 2nd order time correlation function of each construct to a sum of exponentials suggested the presence of multiple exponential decays not present in the DNA-only data (Figure 5.5). See Appendix C, Chapter IV for a review of the exponential form of time correlation functions

An initial exponential decay on the order of a few microseconds was discovered, but it likely corresponds to chromophore photophysics so was not included in this analysis. The same decay was observed in control experiments without the protein present (see Chapter IV).

Simulating Time Correlation Functions and Equilibrium Probability Distributions

A knowledge of the timescales of the decays of the experimental time-correlation function tells you approximately the minimal number of states that are needed to simulate a time-correlation function using a Markov-state model. In a Markov model, one describes a set of states with an associated observable value (in our case, FRET), and then a probability to go from one state to another in a given time (i.e. a state-to-state rate). Markov models are useful in making predictions and understanding the most likely pathways from one state to another.

The entire process can be summed up in an equation known as the transport master equation

$$\dot{\boldsymbol{p}}(t) = K\,\boldsymbol{p}(\boldsymbol{t}), \tag{5.1}$$

where $\boldsymbol{p}(t)$ is the $N \times 1$ population state vector for N states (Eq 5.1), and K is the rate matrix, shown in Eq. (5.4).

$$\boldsymbol{p}(\boldsymbol{t}) = \langle \boldsymbol{p_1}(\boldsymbol{t}), \boldsymbol{p_1}(\boldsymbol{t}), \cdots, \boldsymbol{p_N}(\boldsymbol{t}) \rangle, \tag{5.2}$$

The sum of $\boldsymbol{p}(t)$ is always 1. Even if the system starts out far from equilibrium, given enough time the population state vector will approach the probability distribution, $\boldsymbol{p}^{eq}$. The form of $\boldsymbol{p}(t)$ is equivalent to the equilibrium FRET probability distribution that can be measured experimentally.

$$\boldsymbol{p}^{eq} = \sum_{i=1}^{N} p_i^{eq}\, e^{-\left(\frac{E-E_i}{\sigma_i}\right)^2}, \tag{5.3}$$

where σ_i is the standard deviation of state-i, E is the array of FRET values 0-1, and E_i is the FRET value of state-i. Each FRET state-i is a macrostate, containing many microstates.

A macrostate is all molecular conformations (the individual microstates) with low internal barriers between them. In contrast, relatively high barriers separate one macrostate from another. As such, the system will fluctuate from one macrostate to another at a rate proportional to the heights of the relevant state-to-state barriers. The rate matrix encapsulates these barriers, which are defined by *rate constants*. The rate matrix has off-diagonal elements k_{ij} that are the rate constants with which system transfers from state-i to state-j. The diagonal elements of K are such that the sum of each column of the rate matrix is equal to zero.

$$K = \begin{pmatrix} -\sum_{i=1}^{N} k_{1,i} & k_{2,1} & \cdots & k_{N,1} \\ k_{1,2} & -\sum_{i=1}^{N} k_{2,i} & \ddots & \vdots \\ \vdots & \ddots & \ddots & k_{N,N-1} \\ k_{1,N} & \cdots & k_{N-1,N} & -\sum_{i=1}^{N} k_{N,i} \end{pmatrix} \tag{5.4}$$

Using numerical methods from linear algebra, commercial software such as MATLAB can quickly solve for the eigenvalues and eigenvectors of K, and we can write the solutions to the master equation as

$$P(t) = U\, e^{\Lambda\, t}\, U^{-1}\, P_0. \tag{5.5}$$

In this equation, $P(t)$ is the *conditional probability tensor*, with N^2 elements being the *conditional probability vectors* $p_{ji}(t)$. U is the modal matrix formed by combining each eigenvectors of K into a matrix of column vectors

$$U = (\,\vec{\boldsymbol{v}}_{\boldsymbol{1}} |\, \vec{\boldsymbol{v}}_{\boldsymbol{2}} |\, \dots \,|\, \vec{\boldsymbol{v}}_{\boldsymbol{N}}) = \begin{pmatrix} v_1^1 & v_2^1 & \cdots & v_N^1 \\ v_1^2 & v_2^2 & & v_N^2 \\ \vdots & & \ddots & \vdots \\ v_1^N & v_2^N & \cdots & v_N^N \end{pmatrix}. \tag{5.7}$$

Note that the paired eigenvectors and eigenvalues should be sorted in decreasing order starting with the highest value at $\lambda_1 = 0.$ U is by construction an invertible matrix, so U^{-1} is real-valued. $e^{\Lambda\, t}$ is a sparse matrix of exponentiated eigenvalues of K, i.e. ${e^{\Lambda\, t}}_{i,i} = e^{\lambda_i t}$, and ${e^{\Lambda\, t}}_{i,j} = 0, \forall\, i \neq j$. P_0 is the identity matrix of dimension N. The term $U^{-1}\, P_0$ gives the expansion coefficients in the solution for $P(t)$ in Eq. 5.1

Finally, once we have $P(t)$, and E, the array of FRET values for each state we can write the form of the 2nd and 4th order time correlation functions as

$$\bar{C}^{(2)}(\tau) = \sum_{i,j=1}^{N} \delta E_j \, p_{ji}(\tau) \; \delta E_i \, p_i^{eq} \tag{5.8}$$

and

$$\bar{C}^{(4)}(\tau_1, \tau_2, \tau_3) = \sum_{i,j,k,l=1}^{N} \delta E_l \, p_{lk}(\tau_3) \, \delta E_k \, p_{kj}(\tau_2) \, \delta E_j \, p_{ji}(\tau_1) \, \delta E_i p_i^{eq}, \tag{5.9}$$

where $\delta E_i = E_i - \langle E \rangle$, E_i is the FRET value of state-i, and p_i^{eq} is the equilibrium probability of being in state-i. In Eq.5.8, τ is the time between the two measurements; likewise, in Eq. 5.9, τ_3, τ_2, and τ_1 are the time intervals between all four measurements. Therefore, the 2nd and 4th order TCFs are expressions of all 2- and 4- step pathways among the different states.

Calculating Time Correlation Functions from Single-Molecule FRET Data

We calculate the 2nd order TCFs from experiments using MATLAB to step through the array of FRET values and compute

$$\bar{C}_E^{(2)}(\tau) = \frac{1}{N_{pairs}(\tau)} \prod_{t=0}^{M-1} \{\delta E(t) \delta E(t+\tau)\}, \tag{5.10}$$

where $N_{pairs}(\tau)$ is the exact number of measurements separated by τ, and M is the total number of measurements taken in the duration of the scan. A consequence of the sparse sampling is that $N_{pairs}(\tau)$ is non-uniform across the domain. We normalize $\bar{C}_E^{(2)}(\tau)$ by explicitly accounting for this distribution of measurements.

Similarly, we calculate the 4th order TCF as follows,

$$\bar{C}_E^{(4)}(\tau_1, \tau_2, \tau_3) = \frac{1}{N_{pairs}(\tau_1, \tau_2, \tau_3)} \prod_{t=0}^{M-1} \{\delta E(t) \delta E(t+\tau_1) \delta E(t+\tau_1+\tau_2) \delta E(t+\tau_1+\tau_2+\tau_3)\} \tag{5.11}$$

Multidimensional Parameter Optimization

To find a set of rate constants and FRET states that are able to simulate FRET probability distributions, and time correlation functions that resemble experiments, we used a combination of a non-deterministic genetic algorithm and a deterministic constrained nonlinear multivariable function optimizer, MATLAB's *ga* and *fmincon* functions, respectively. The procedure was to use the genetic algorithm many times to find a family of solutions that are near the global minimum. From there, *fmincon* was used to further refine each guess to the local minimum.

We introduce the weighted error function χ^2 as a metric to assess the degree to which our models can reproduce statistics from experimentally measured data. The error function is defined to be the cumulative squared-difference between the measured quantities y and the predicted quantities g,

$$\chi^2 = \sum_{q=1}^{3} w_i^q \sum_i \left(y_i^q - g_i^q(\{k_{lm}, A_l\})\right)^2 \tag{5.12}$$

where q enumerates the histogram, 2- and 4-point time correlation functions, i is the domain index for each function and w_i^q are the weighting functions of each of the q surfaces. w_i^q are chosen such that the histogram, and time correlation functions contribute approximately equal errors to the difference function. Additionally, w_i^q is a function of lag times τ, such that more weight is given to the early times of the time correlation function.

The set of parameters with the overall best agreement to the data was chosen to be the *best fit* for that construct using the particular model. We used a heuristic to determine which model was best for a given construct: if the χ^2 measure of fitness is lowest on the simplest model (lower is better) we choose that model. If the χ^2 is only marginally higher with a simple model, we choose the simpler model. If the χ^2 value is sufficiently lower with a more complex model, we choose the more complex model. The AIC and BIC criteria can also be used to select the best model for gp32 dimer assembly. Both information criteria reward accuracy of the goodness of fit but penalize for the entropic reduction by using more states, and as such can be thought of as *maximum entropy methods.*

Modeling the gp32 dimer assembly pathways

To properly simulate the time correlation functions for gp32 binding to the 14- and 15-nucleotide ssDNA constructs, we used the five- and six-state models shown in Figure 5.6. Each state of the model has a FRET value associated with it.

We tried several networks (mechanistic pathways) to model the dimer assembly process, but ultimately the two that proved to be most fruitful were the five- and six- state versions. Appendix C shows the best network out of all attempts to fit each construct with every model. The best results of fitting the models are shown in Figures 5.7 – 5.10, and the values are recorded in Table 5.2.

Motivated by studies of the single-molecule FRET results obtained with the DNA constructs in the absence of added gp32 (see Chapter IV for more details), we modeled the free DNA constructs as two-state systems. We decided to combine the intermediate and extended DNA only conformation for simplicity. State-1 is a macrostate that contains all conformations of ssDNA that were not appropriate for binding gp32 monomers, and state-2 is a macrostate that includes all the conformations of ssDNA that are suitable for gp32 monomer binding. The nucleotide binding-footprint of gp32 is seven nucleotide residues in length[2], so presumably state-2 features conformations of ssDNA with seven contiguous nucleotides in secondary structure conformations that do not preclude gp32-nucleic acid interactions. When the ssDNA fluctuates into a binding-possible state from a binding-impossible state, a nearby ssb protein can capture the ssDNA backbone and bind. Most of the ways the initial gp32 monomer can bind are not on-path for a second monomer to bind: these conformations are collectively designated as state-3. However, if a 'nucleating-monomer' binds right next to the ssDNA-dsDNA junction or at the terminal end of the ssDNA region, then (depending on orientation and proximity) a second monomer can also bind to the ssDNA backbone to form a gp32 dimer: these *on-path* monomer-bound states are conduits to forming a protein dimer, and are designated as state-4. Stated in another way, state-3 is unproductive for dimer formation while state-4 is productive for this process.

Lastly, our models feature a states for the gp32 dimers. We found that a single dimer state was sufficient to model the 14-mer constructs (state-5), while two non-degenerate states were necessary to model the dimer state in the 15-nucleotide constructs (states-5 and -6). Modeling the 14-mers with the six-state model as opposed to the five-state model resulted in poorer agreement between the simulated correlation functions and FRET distributions when compared to experimentally derived correlation functions and FRET distributions. In the $5'$-construct, the χ^2 measure of fitness was lower for a simpler model and only barely better for the $3'$-construct, such that an analysis with Bayesian information criterion (BIC) still yields a higher value. For the 15-nucleotide constructs the increase in agreement between simulations and experiment using a 6-state model was sufficient to account for the entropic cost of using a more complex model.

We fit to three surfaces in this study, the FRET probability distribution, the 2nd order TCF and the 4th order TCF. The fits to the FRET histogram, $C^{(2)}(\tau)$, and $C^{(4)}(\tau_1, \tau_2, \tau_3)$, are shown in Figures 5.8, 5.9, and 5.10. The values used to simulate all the surfaces are shown in Table 5.1 and again in network-form in Figure 5.7.

The fastest processes were unsurprisingly the interconversion between ssDNA states, which occur in tens-of-microseconds timescale. State-1 is the thermodynamically most stable DNA-only state, as reflected by the relative values of t_{12} and t_{21}. The slowest processes are typically the t_{24} monomer association rate, where one gp32 monomer has to bind to a particular stretch of ssDNA such that there is room for another neighbor to bind.

Figure 5.7 shows the best fit from each construct. As mentioned, the overall best fits for the 14-nucleotide constructs were accomplished with a 5-state model, and the best fits for the 15-nucleotide constructs were achieved with a 6-state model.

The probability FRET distributions reflect the equilibrium conformational states, shown in Figure 5.8.

Discussion

Informatively, in the two 14-nucleotide ssDNA constructs, the FRET states corresponding to the gp32-dimer were different: the 3′-construct had an even lower FRET efficiency state (E = 0.10) than its 5′ counterpart (E = 0.33). Since the 14-mer constructs can only hold one gp32 dimer, in register with nucleotide residues 1-14, we can interpret the difference in FRET values caused by polarity-driven asymmetries in the protein-nucleic acid interactions on the ssDNA portion of the construct. The fact that both constructs have a dimer in nucleotide positions 1-14, limit the likelihood that the FRET change is due to a conformational change of the ssDNA lattice, and suggest that maybe the change in FRET is, at least in part, dependent on the presence of the chromophores.

The ssDNA has a Cy-3 donor chromophore at the terminal-end which can be affected by PIFE, protein induced fluorescence enhancement[10,11]. Since the gp32 protein binds to ssDNA in an oriented manner, it is possible that the end of the protein facing the

5′-portion of ssDNA has a different interaction with the ssDNA than the end of the protein facing the 3′-portion.[12] So when the Cy3 chromophore is located at the 3′-end of ssDNA, there may be an interaction with gp32 that is not present when the Cy3 chromophore is at the 5′ end.[12] The resolution of the available crystal structure of gp32 with ssDNA was too low to determine which end of the protein faces which end of the ssDNA[3]. Nonetheless, our data suggest that there are distinct FRET states between the two polar constructs.

One end of the gp32 dimer has an intrinsically disordered, negatively charged C-terminal region[13,14], and the other end has a shorter, and also intrinsically disordered, N-terminal region responsible for gp32-gp32 cooperativity[15]. It is possible the C-terminal region of gp32 interacts with the Cy3 probe and causes the resulting fluorescence enhancement event. If this were true, it would suggest the polarity of the free N-terminal to free C-terminal axis of the gp32 dimer cluster aligns with the 5′ to 3′ axis of the ssDNA strand. This would explain why the PIFE-caused low-FRET dimer state is observed with the 3′ constructs, but not with the 5′ constructs. Conversely, if it were the free N-terminal domain of the dimeric gp32 cluster that causes the PIFE event, we would expect the free N-terminal to C-terminal gp32 cluster axis to align with the ssDNA 5′ to 3′ axis of the ssDNA lattice.

Further evidence for the 3′ PIFE effect is shown in the fits to the 15-nucleotide constructs. The best fits using the 6-state model also show a low FRET state for the 3′ constructs that is not present for the 5′ constructs.

Taking a closer look at the FRET values associated with the on-path monomer state (state-4), we see that it usually has a relatively higher FRET value, and so is probably not associated with monomer-binding to the end of the ssDNA tail where the Cy3 chromophore is attached, as we do not see a markedly lower value. Therefore, we hypothesize that state-4 usually involves the initial monomer binding contiguously to the ss-dsDNA fork junction.

Note that state-4 is not well defined because it encapsulates two very different monomer initiation sites: one near the fork and the other as far away from it as possible. Nonetheless, both states are potential targets for another monomer to bind to form a gp32 dimer. The rate at which a monomer binds to the ssDNA is typically measured in tens of milliseconds, and the rate of dissociation is usually faster than association (Table 5.2).

Once a dimer is bound to ssDNA, our fitting suggest it stays bound to the 5′ constructs longer than the 3′ constructs, possibly suggesting that the polarity of the gp32 binding interaction destabilizes some positions of the dimer cluster on the ssDNA lattice more than others. The rate of sliding appears to be faster toward the relatively high-FRET dimer state for both 15-nucleotide constructs, suggesting that gp32 dimer cluster sliding toward the ss-dsDNA junction is favored over sliding in the opposite direction.

Conclusion

We have shown that both the length and polarity of the ssDNA template effect the assembly mechanism of gp32 dimers. Our modeling suggests that dimer sliding occurs at the rate of about 1 nucleotide per ms towards the ss-dsDNA junction and much more slowly in the opposite direction. Our fits are also consistent with the fact that the gp32 dimer binds to ssDNA in a polar fashion, as reflected by the different FRET states of the 3′ and 5′ constructs that are likely caused by protein induced fluorescence enhancement when the donor is on the 3′ end of ssDNA.

CHAPTER VI

CONCLUDING SUMMARY

In this dissertation I have investigated the dynamics of DNA and of protein-DNA interactions primarily through single-molecule FRET spectroscopy. Single molecule experiments are windows into unfamiliar worlds, and the results are often difficult to interpret. In Chapter II I first present the methodology of modeling the dynamics of a complex chemical system using a relatively simple Markov state model. The work presented in this chapter was a collaboration with Carey Phelps. We show that using solutions of the transport master equation, one can simulate time correlation functions (TCFs) and probability distribution functions. These statistical functions can also be calculated from single-molecule experiments and compared to the simulations in order to identify the number of conformational states, the value of the observable associated with each state, and a set of kinetic parameters that describe the state-to-state interconversion. We hope that the biophysics field will adopt these powerful analytical techniques, which I and other lab members have continued to develop and make more accessible.

In Chapter III I demonstrated the TCF-analysis method on a single-molecule experiment. In this work, also with lab members Carey Phelps, Davis Jose, and Morgan Marsh, we examine the mechanism of gp32 dimer assembly onto a ssDNA lattice next to a ss-dsDNA junction. We found that a four-state model could describe the gp32 concentration-dependent effects on the kinetics of gp32 dimer assembly on timescales of a millisecond and beyond. The four-state model features a lone DNA conformational state, two mutually exclusive states corresponding to the DNA-gp32 monomer complex, and one state for the DNA-$(gp32)_2$ dimer complex. We found that a single-monomer of gp32 was more likely to unbind from ssDNA than to slide to another position along the lattice. The mechanism of gp32 dimer assembly involves a nucleation step with gp32 monomer randomly binding in a ‘productive’ position such that another gp32 monomer has room to bind cooperatively to it on an adjacent ssDNA site before the first gp32 has time to unbind. This finding reinforces the notion that DNA replication is stochastic and proceeds through a series of thermally driven molecular collisions without the need for a master coordinator.

In Chapter IV I present the results of a microsecond-resolution study of ssDNA conformational dynamics. Significant improvements to the single-molecule instrument, driven by an optical geometry not used in prior experiments, led to sub-millisecond experiments on variations the same constructs examined in Chapter III. Working with Anson Dang, and Megan Barney, we took single molecule FRET data on four different DNA constructs that differed in ssDNA length and polarity. I developed a novel method to calculate time-correlation functions that enables the calculation of experimental TCFs even if the FRET signal was not continuous. Claire Albrecht and I further refined the method to quickly calculate solutions to the transport master equation for arbitrarily complex networks using numerical methods.

We found that a cyclical three-state Markov model of ssDNA conformations was the minimal model that could reproduce experimental TCFs and FRET probability distributions. The three-state model features three ssDNA macrostates: a compact ssDNA conformation, an intermediate-extension conformation, and a maximally extended state. The timescale of interconversion between conformational states was on the order of tens of microseconds to milliseconds. Even with relatively unstructured oligo-deoxythymidine, we found that the extended state was stable for several milliseconds at a time. Our analysis thus suggests that large conformational fluctuations in ssDNA near ss-dsDNA junctions occur and are likely important states in protein binding mechanisms.

Chapter V examines the polarity-dependent binding mechanism of gp32 onto ssDNA regions near ss-dsDNA junctions. This work was also completed working in collaboration with Claire Albrecht, Anson Dang, and Megan Barney. We took microsecond-resolution single-molecule FRET measurements on the four p/t DNA constructs introduced in the last chapter. Motivated by the results of the studies of ssDNA conformational fluctuations presented in Chapter IV (and exponential fits to the 2nd order TCFs of the DNA + gp32 samples), I proposed a new model for gp32 dimer assembly onto ssDNA near ss-dsDNA junctions. This model extends the mechanism proposed in Chapter III by adding a non-productive DNA only state, and an extra state for the DNA:gp32 dimer complex. The new states added are key transition states in the assembly mechanism. They are very short short-lived, often lasting less than a millisecond, which is why the earlier study presented in Chapter III was not able to resolve them.

We found that the gp32 dimer can slide on the order of one nucleotide per ms. To our knowledge this is the first measurement of the kinetics of gp32 dimer sliding. Our results suggest the dimer slides more quickly toward the ss-dsDNA junction than away from it regardless of ssDNA polarity. Furthermore, based on our analysis of the 14-nucleotide constructs (which can only fit the dimer in one position) we concluded that the gp32 dimer causes a polarity-dependent PIFE effect with the Cy3 chromophore on the 3′ construct but not in the 5′ construct. We believe this is a very local effect, because in the 3′-p(dT)$_{15}$ construct (which can accommodate the dimer in two different positions) we see the presence of both low- and relatively high-FRET dimer states. This observation will be useful in future studies, and can even be used as a tool to further inform on the location of ssb protein on ssDNA lattices.

The gp32 dimer-sliding rates could be influenced by end-effects (i.e. the interaction of the gp32 dimer with end of the ssDNA and by interactions at the ss-dsDNA junction). If one were to repeat these studies on longer ssDNA constructs, it may be possible to extract kinetic parameters of sliding with fewer complicating interactions. Of course, the ss-dsDNA junction is the nexus at which much of fundamental biology occurs including DNA replication and repair, so learning more about how proteins interact with it is a valuable exercise.

Altogether, the findings presented in this dissertation suggest that single-molecule FRET spectroscopy is a formidable tool for the molecular biologist. Future studies will benefit from an increased understanding of the connections between a Markov state model and time correlation functions. DNA is a remarkable polymer. It may be the single most important molecule on the planet: if DNA didn’t exist, no one would be around to notice.

APPENDIX A

FIGURES

Figures for Chapter II

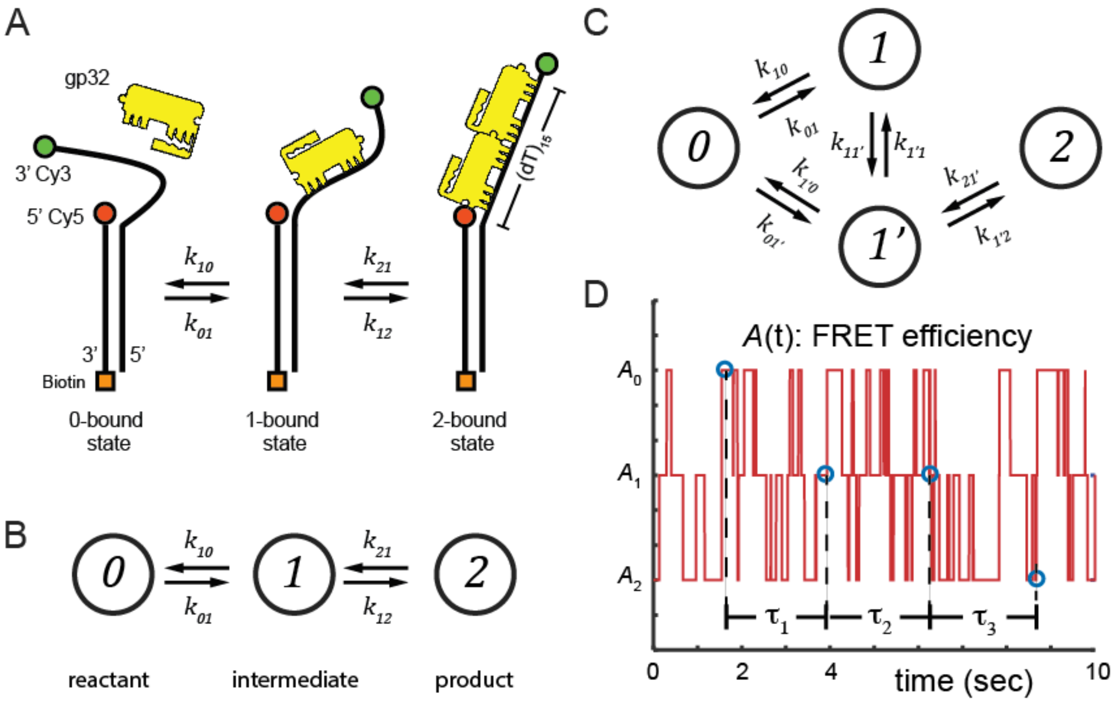


Figure 2.1. Models of Dimer Assembly and a single-molecule Trajectory
(*A*) A hypothetical 3-state reaction scheme for the ssDNA binding protein gp32, which can bind up to two proteins to the p(dT)$_{15}$ 'tail' region of a p/t DNA construct. FRET donor and acceptor chromophores (depicted as green and red circles) label the 3' end of the ssDNA region and the p/t junction, respectively. The gp32 protein is shown in yellow. (*B*) The 0-, 1- and 2-bound states of the *N* = 3 system shown in Panel (*A*) are depicted as a linear reaction scheme, in which the reactant (state-0) and product (state-2) are coupled by a single intermediate (state-1). (*C*) The reaction is depicted as an *N* = 4 system, in which the conformational end-states are inter-connected by a 'non-productive' intermediate (state-1) and a 'productive' intermediate (state-1'). Stochastic transitions from state-i to state-j occur with probabilities determined by the rate constants k_{ij}, where $i, j \in \{0, 1, \ldots, N-1\}$. (*D*) A simulated trajectory of the stochastic variable $A(t)$ is shown for the *N* = 3 system. Here we have assigned the three states to the resolvable values A_0 = 0.8, A_1 = 0.5, and A_2 = 0.2, and we have used the transition rates k_{01} = k_{21} = 5 s^{-1}, and k_{10} = k_{12} = 10 s^{-1}. An example of a four-point sequence of data points are shown corresponding to the time intervals τ_1, τ_2, and τ_3. Figure partially adapted from chapter II reference 28.

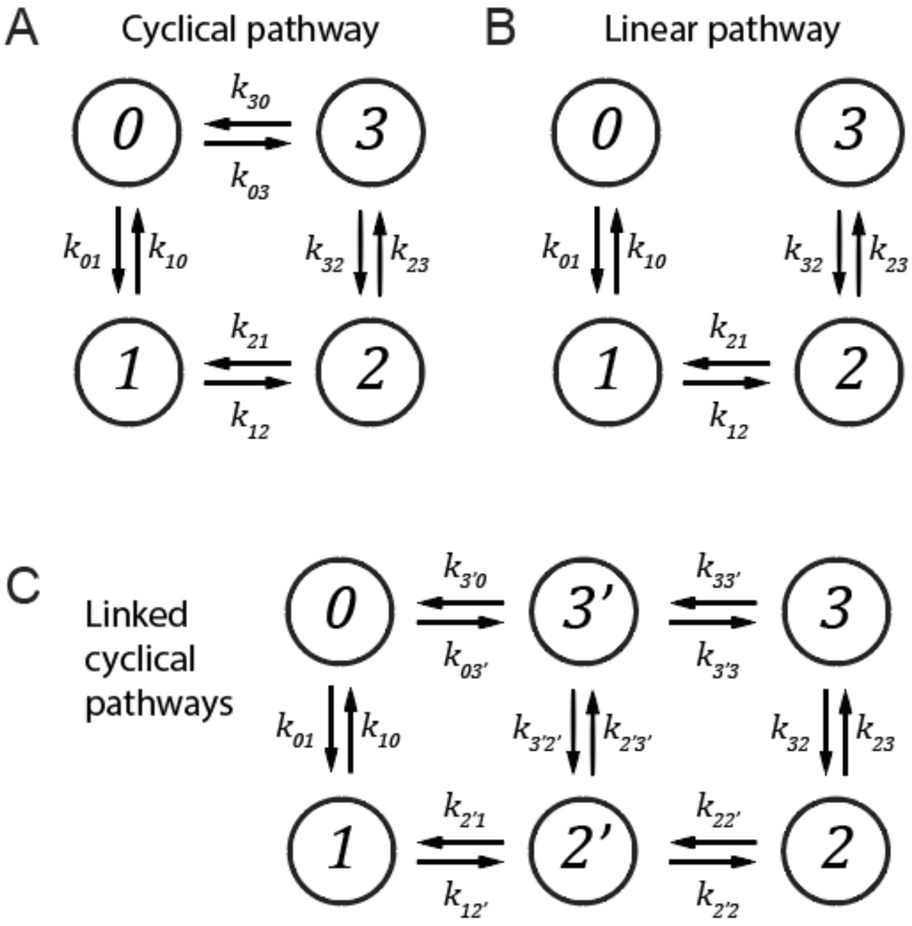


Figure 2.2. Example Kinetic Schemes
Example kinetic schemes for which the detailed balance condition requires different constraints to be applied to the rate constant relationships due to the presence or absence of cyclical pathways. (*A*) Single cyclical pathway. (*B*) Linear pathway. (*C*) Linked cyclical pathways.

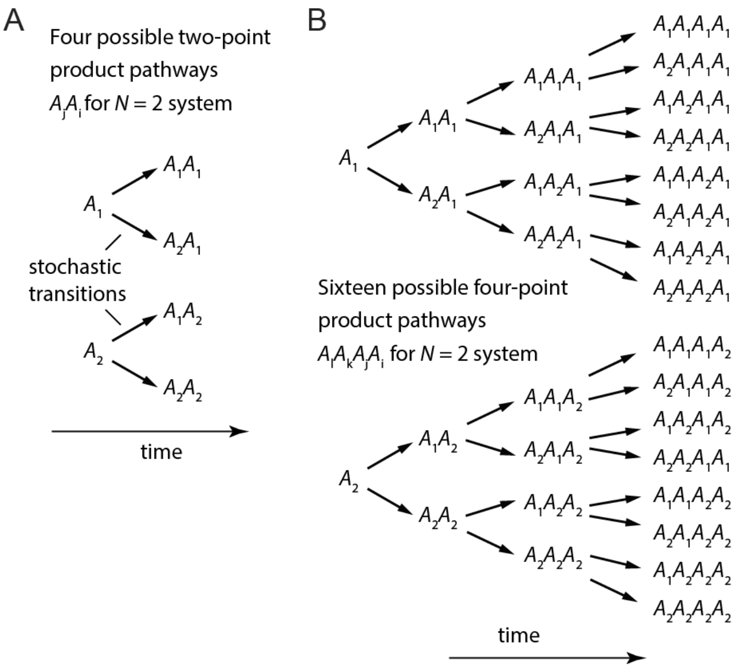


Figure 2.3. Transition pathway contributions to 2nd- and 4th-order TCFs for two-state (N = 2) system.
(*A*) There are $N^2 = 2^2 = 4$ possible outcomes of a time-ordered two-point product of the observable $A(\mathrm{t})$, which are used to construct the 2nd-order TCF $\bar{C}^{(2)}(\tau)$. (*B*) There are $N^4 = 2^4 = 16$ such sequences for the four-point product that is used to construct the 4th-order TCF $\bar{C}^{(4)}(\tau_1, \tau_2, \tau_3)$. The conditional probability $p_{ji}(\tau)$ that a stochastic transition will occur from state-i to state-j within the time interval τ is given by Eq. (10).

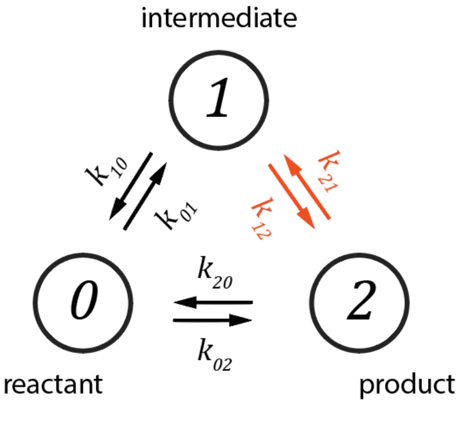


Figure 2.4. Three-State Model for gp32 Dimer Assembly
The $N = 3$ reaction redrawn from Fig 2.1 as a cyclical scheme. This allows for the product state-2 to form either directly from the reactant state-0, or through the intermediate state-1. The 'coupling step' is indicated in red.

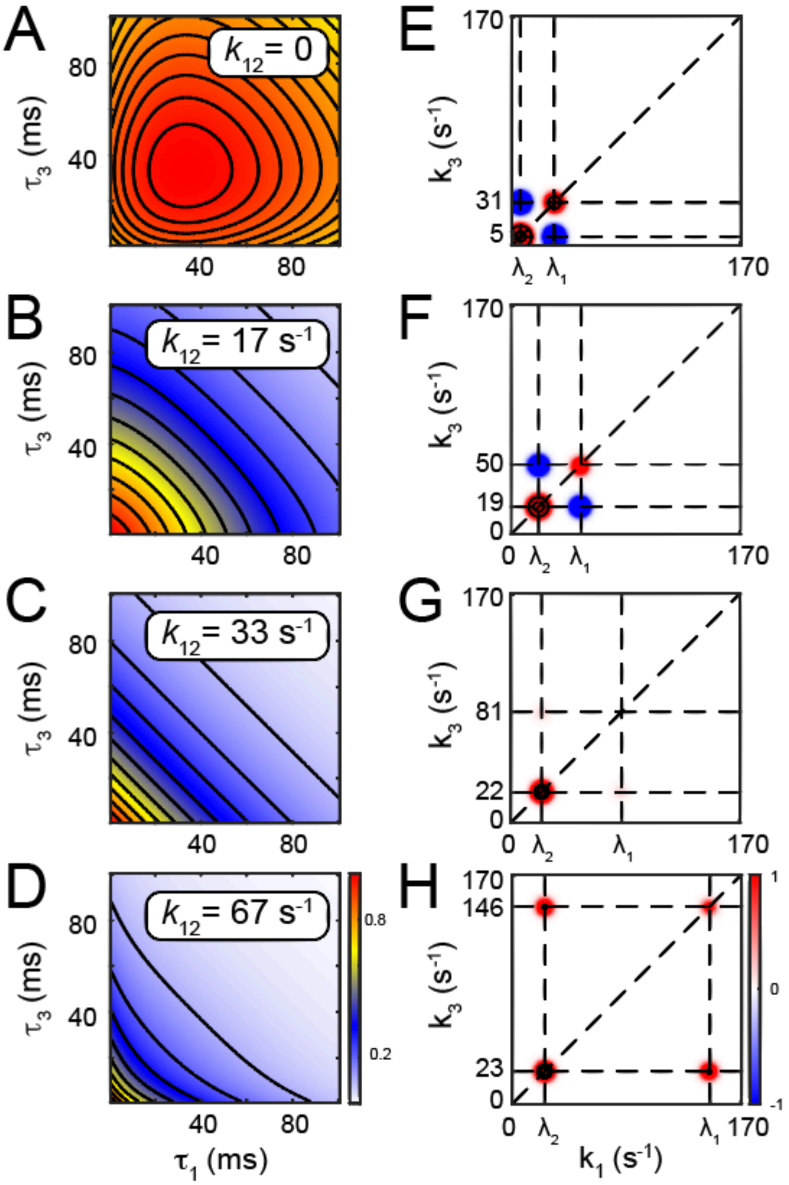


Figure 2.5. Simulations of N=3 4th-order TCFs and 2D rate spectra
Calculated 4th-order TCFs (panels *A – D*) and associated two-dimensional (2D) rate spectra (panels *E – H*) for the cyclical *N* = 3 system shown in Fig. 2.4. Here we have taken the waiting time interval τ_2 = 1 ms, and the rate constants k_{01} = 10 s^{-1}, k_{10} = 20 s^{-1}, k_{02} = 2 s^{-1}, and k_{20} = 4 s^{-1}. The TCFs are described by Eq. (2.17) and the 2D rate spectra by Eq. (2.18). The rate constants of the 'coupling step,' k_{12} = k_{21} are adjusted over the range (*A* and *E*) 0, (*B* and *F*) 16.7 s^{-1}, (*C* and *G*) 33.3 s^{-1}, and (*D* and *H*) 66.7 s^{-1}. For each of these conditions, values of the self- and cross-term amplitudes $\mathcal{A}_{11}$, $\mathcal{A}_{22}$, $\mathcal{A}_{12}$ = $\mathcal{A}_{21}$, respectively, are given in the text.

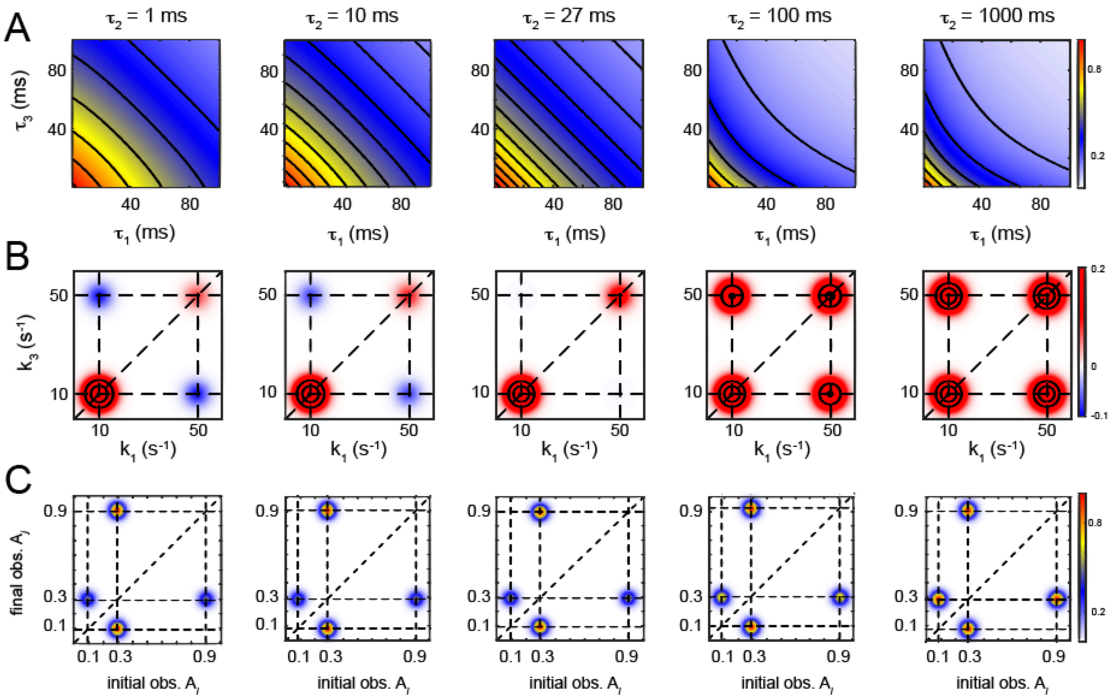


Figure 2.6. Simulations of N=3 4th-order TCFs, 2D rate spectra and TDPs
Model calculations for (*A*) the 4th-order TCFs, (*B*) the 2D rate spectra, and (*C*) the transition density plots (TDPs) as a function of time. For these calculations, we have used the linear $N = 3$ kinetic scheme diagrammed in Fig 2.1B, with values A_0 = 0.9, A_1 = 0.3, and A_2 = 0.1, and the rate constants k_{01} = k_{21} = 10 s^{-1}, and k_{10} = k_{12} = 20 s^{-1}.

Figures for Chapter III

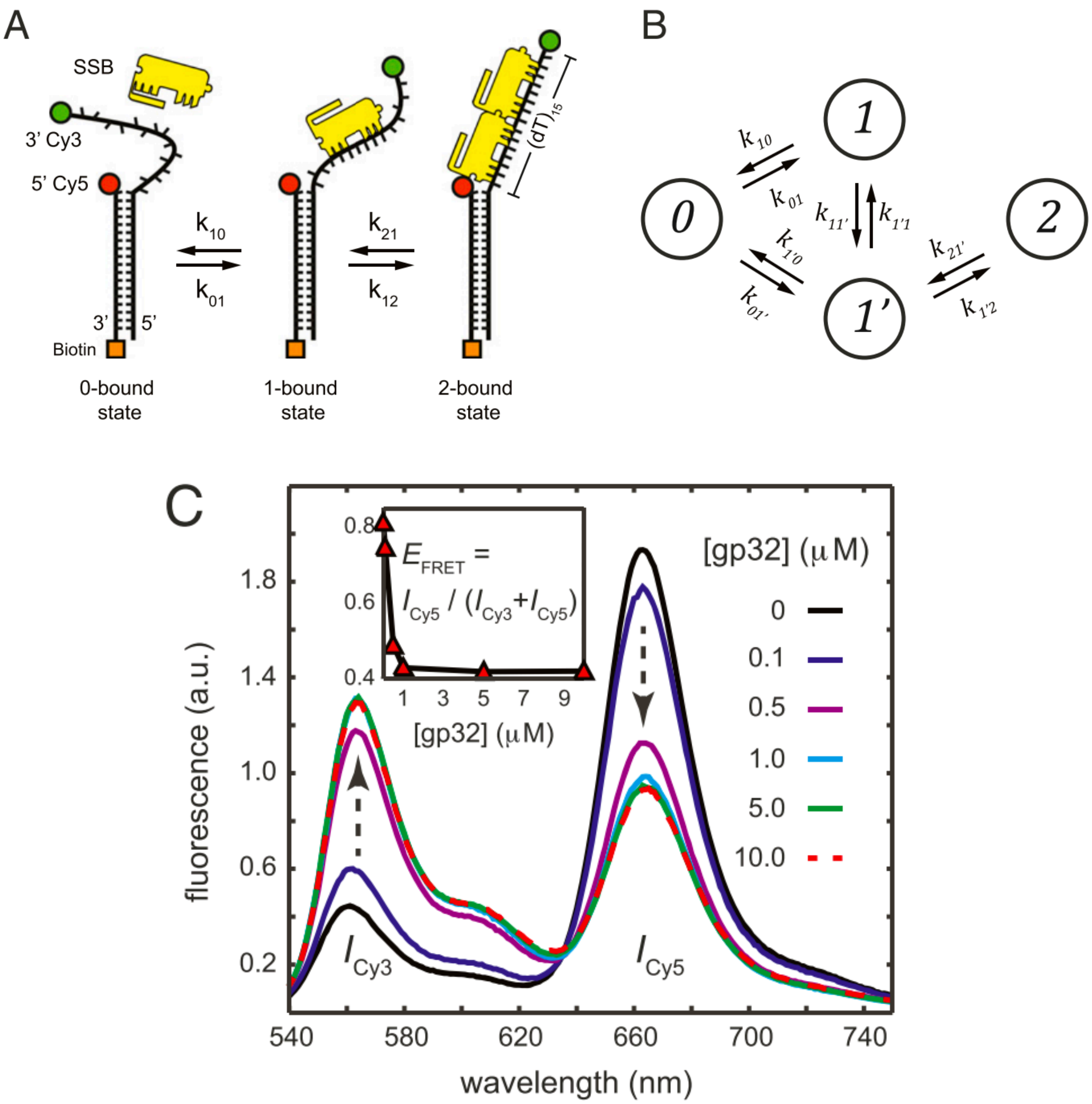


Figure 3.1. Models of Dimer Formation and Ensemble Fluorescence gp32 Titration (*A*) Schematic of the unbound and bound states of binding for the ssDNA-$(gp32)_2$ system. The $p(dT)_{15}$ 'tail' of a p/t DNA construct can bind up to two ssb proteins (gp32, in yellow). FRET donor Cy3 and acceptor Cy5 chromophores (green and red circles) label the ends of the ssDNA region. (*B*) Reaction scheme indicating 'non-productive' (1-bound) and 'productive' (1'-bound) intermediate states. (*C*) Bulk fluorescence measurements at 100 nM p/t DNA concentrations exhibited changes in the FRET efficiency upon titration with gp32. The inset shows the values of E_{FRET} determined from peak Cy3/Cy5 fluorescence intensities.

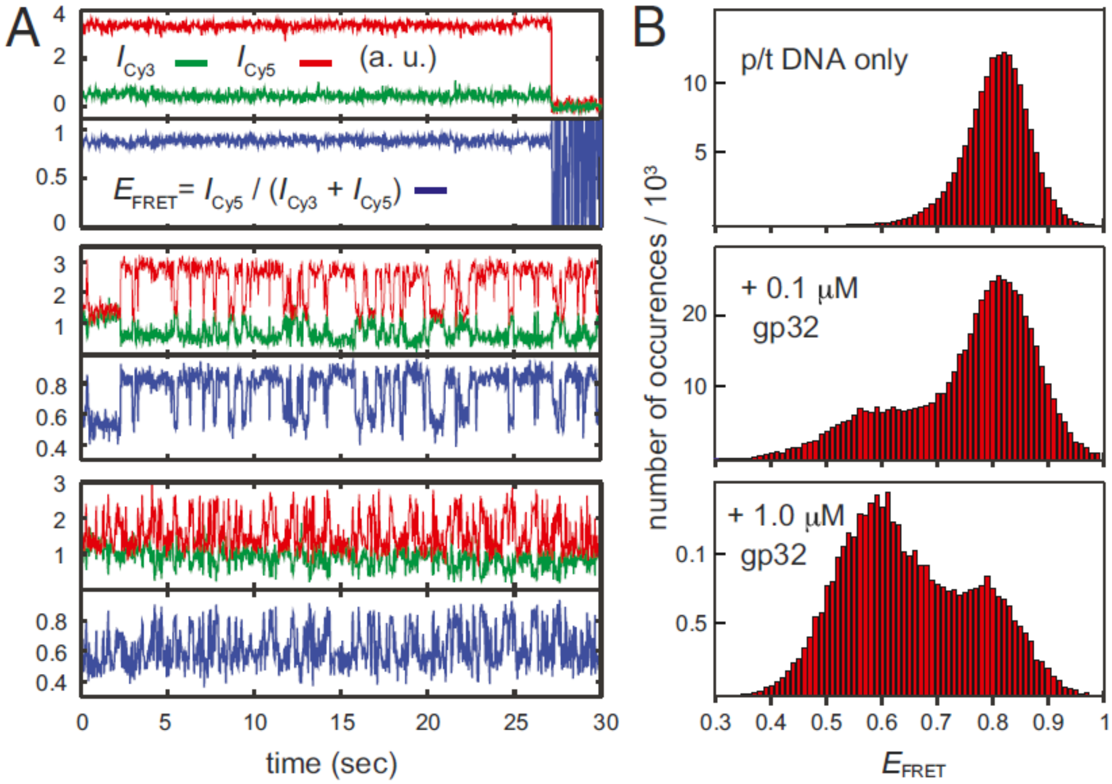


Figure 3.2. Representative Single-Molecule Trajectories and smFRET Distributions (*A*) Representative single-molecule donor Cy3 (green), acceptor Cy5 (red) and smFRET trajectories (blue) taken from the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct in the presence of 0, 0.1, and 1.0 μM gp32, respectively. The smFRET efficiency is calculated according to $E_{FRET} = I_{Cy5}/\left(I_{Cy3} + I_{Cy5}\right)$. (*B*) Histograms of the *E*FRET efficiency were obtained from several hundred smFRET trajectories in the presence of 0, 0.1, and 1.0 μM gp32, respectively.

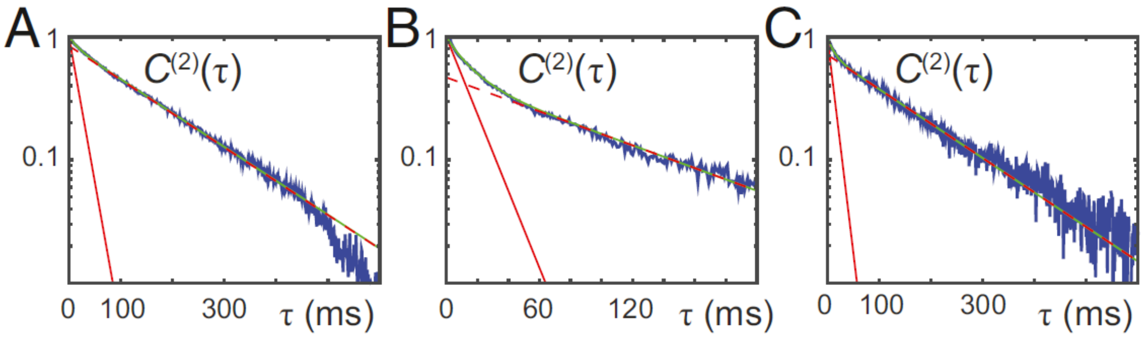


Figure 3.3 Normalized 2nd-order TCFs of microsecond-resolved smFRET trajectories Normalized 2nd-order TCFs of microsecond-resolved smFRET trajectories of the 3'-Cy3/Cy5-p(dT)15-p/t DNA construct in the presence of (*A*) 0.1 µM gp32, (*B*) 1.0 µM gp32 and (*C*) 0.1 µM gp32 + 1.0 µM LAST peptide. Green curves are linear regression best fits to the data. Solid and dashed red lines indicate the fast and slow decay components, respectively. The above decays were constructed from hundreds-of-thousands of data points. The results of the fitting analysis are reported in Table S3.2, in Appendix C, Chapter III.

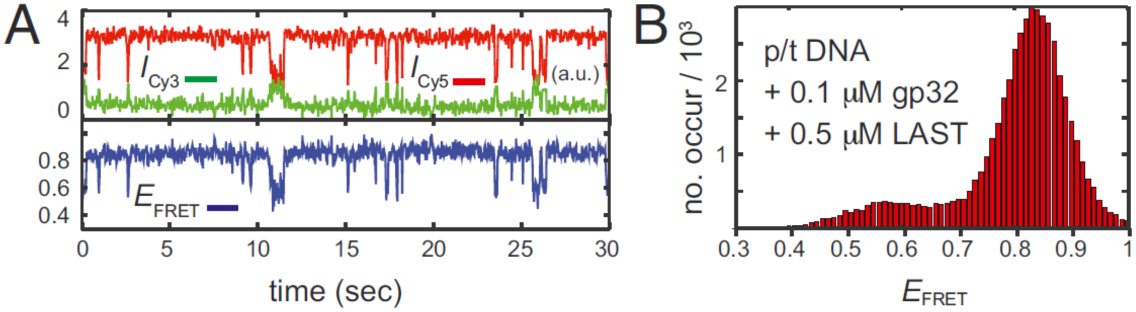


Figure 3.4. Single-Molecule Trajectory and FRET Distribution with LAST Peptide (*A*) Representative single-molecule donor Cy3 (green), acceptor Cy5 (red) and smFRET trajectories (blue) taken from the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct in the presence of 0.1 µM gp32 and 0.5 µM LAST. The smFRET efficiency is calculated according to $E_{FRET} = I_{Cy5}/(I_{Cy3} + I_{Cy5})$. (*B*) Histogram of the E_{FRET} efficiency obtained from several hundred smFRET trajectories in the presence of 0.1 µM gp32 and 1.0 µM LAST.

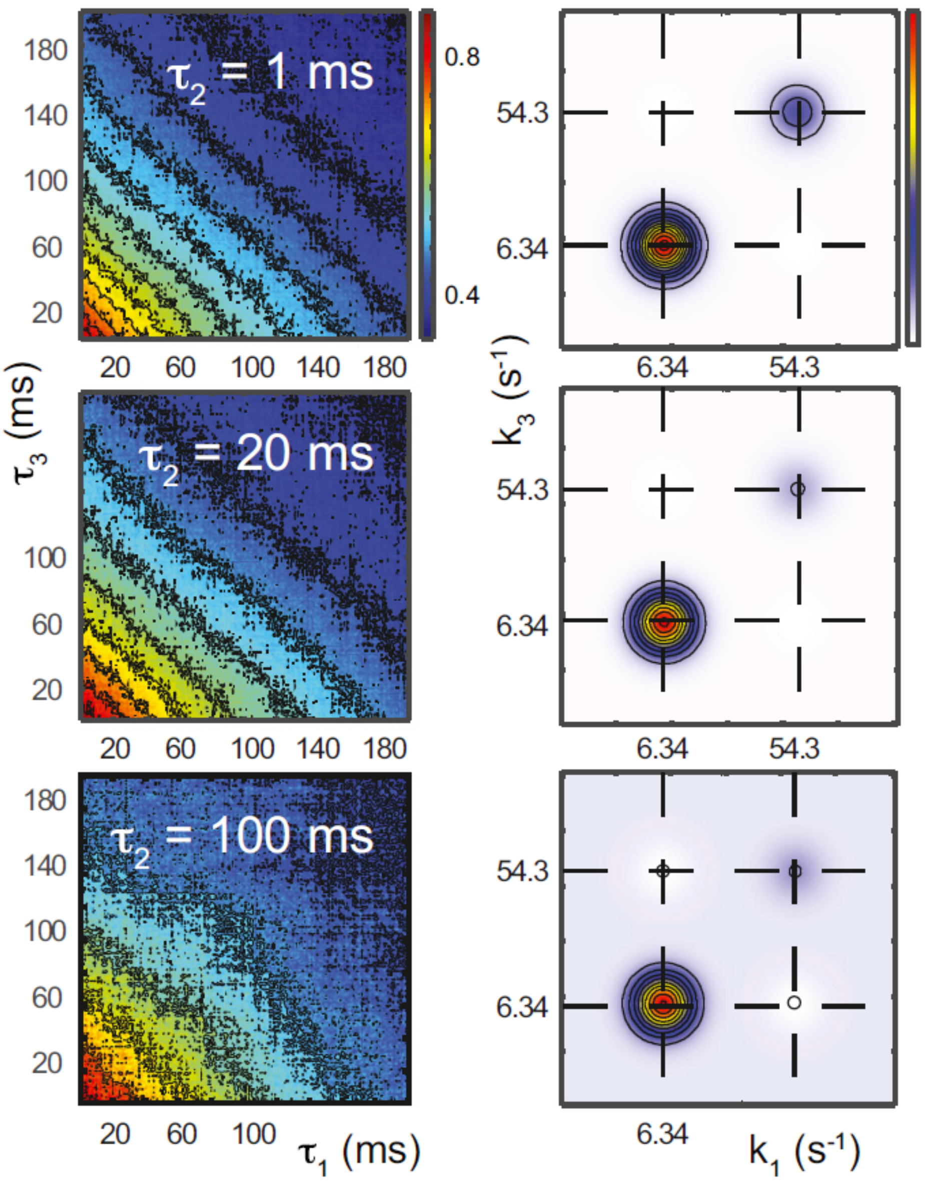


Figure 3.5. 4th-order TCFs and associated two-dimensional (2D) rate spectra
4th-order TCFs (left columns) and associated two-dimensional (2D) rate spectra (right columns) calculated from smFRET trajectories of the 3’-Cy3/Cy5-p(dT)15-p/t DNA construct in the presence of 0.1 μM gp32. The 2D rate spectra were calculated assuming a three-state ($N = 3$) scheme, and the characteristic rate constants determined from the corresponding 2nd-order TCFs (see Table S3.2). Each data set was computed from hundreds-of-thousands of points.

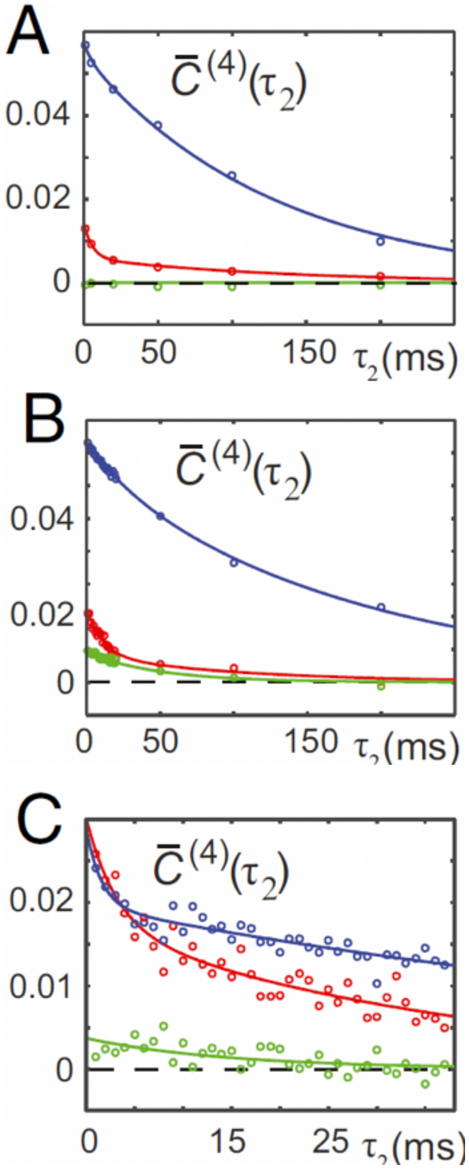


Figure 3.6. Diagonal and off-diagonal amplitudes of the 2D rate spectra
Diagonal and off-diagonal amplitudes of the 2D rate spectra determined from smFRET trajectories of the 3’-Cy3/Cy5-p(dT)15-p/t DNA construct as a function of the interval τ_2, and in the presence of (*A*) 0.1 μM gp32, (*B*) 0.1 μM gp32 + 1.0 μM LAST peptide and (*C*) 1.0 μM gp32. Fast and slow diagonal amplitudes are shown in red and blue, respectively, and the cross-peak amplitude is shown in green. The corresponding eigenvalues obtained in each case are: (*A*) $\lambda_{slow} = 6.34$ s^{-1}, $\lambda_{fast} = 54.3$ s^{-1}; (*B*) $\lambda_{slow} = 6.45$ s^{-1}, $\lambda_{fast} = 80.0$ s^{-1}; and (*C*) $\lambda_{slow} = 10.6$ s^{-1}, $\lambda_{fast} = 71.9$ s^{-1}.

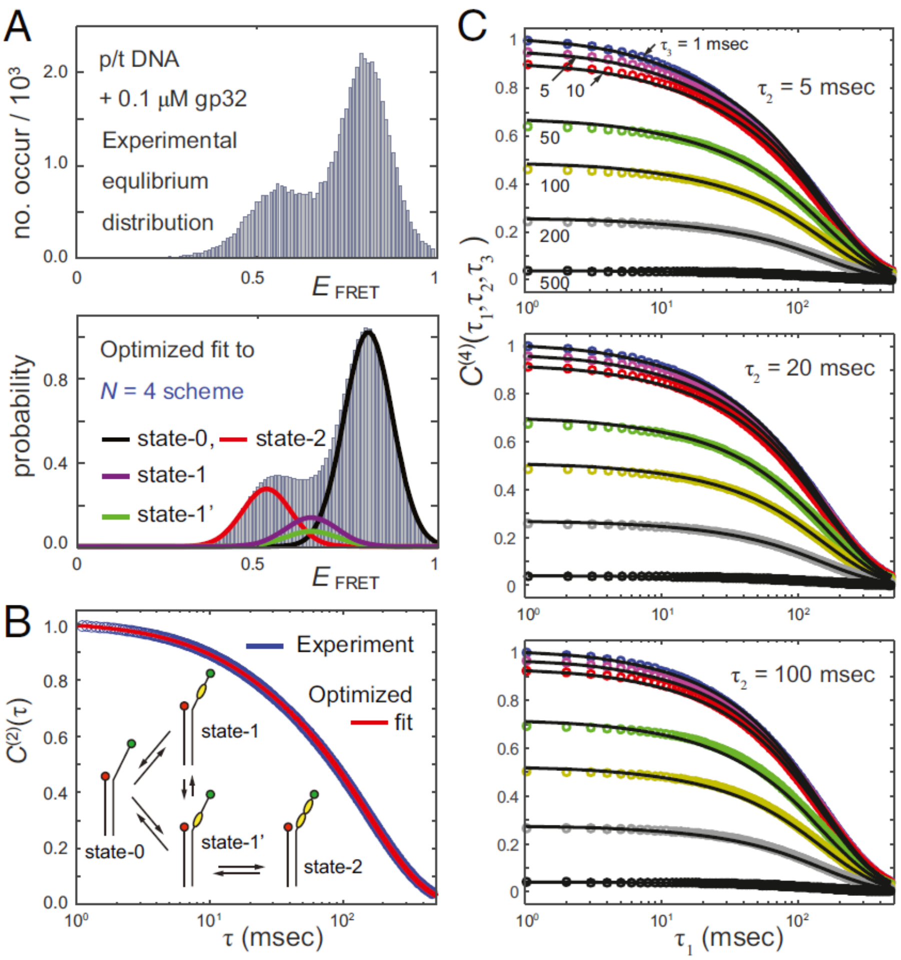


Figure 3.7. Optimized fits of the *N* = 4 scheme

Optimized fits of the *N* = 4 scheme to experimental functions calculated from smFRET trajectories of the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct in the presence of 0.1 μM gp32. (*A*) Top panel: experimental equilibrium distribution of states. Bottom panel: Optimized fits show the component populations of the 0-bound state (black curve), the 2-bound state (red curve), and the 1- and 1'-bound intermediates (purple and green curves, respectively). (*B*) The experimental 2nd-order TCF (blue) is shown overlaid with the optimized fit (red). (*C*) Experimental 4th-order TCFs for various waiting times (colored circles) are shown in comparison to the corresponding optimized theoretical fits (solid black curves). The cumulative fitness of the optimized solutions for the *N* = 3 and *N* = 4 schemes were determined using Eq. (S3.9) and are reported in Table S3.4.

Figures for Chapter III Supporting Information

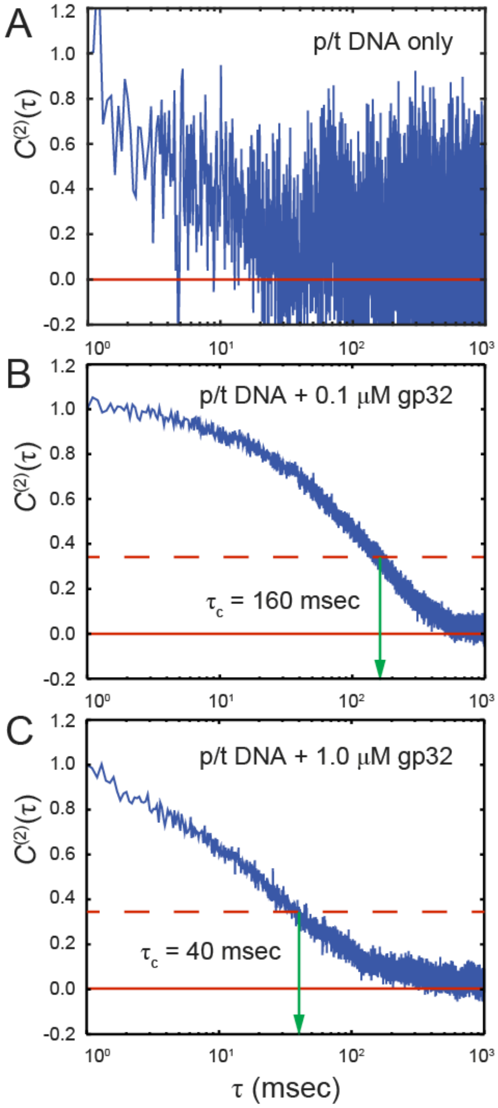


Figure S3.1. Fits to 2nd order TCFs
2nd-order TCFs of microsecond-resolved smFRET trajectories obtained from the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct (at 100 nM) in the presence of (*A*) 0, (*B*) 0.1 and (*C*) 1.0 μM gp32. The decays are plotted using a linear-log axes scale. All decays were constructed from hundreds-of-thousands of data points. Note that for the p/t DNA construct alone, the decay of the TCF is dominated by stochastic noise.

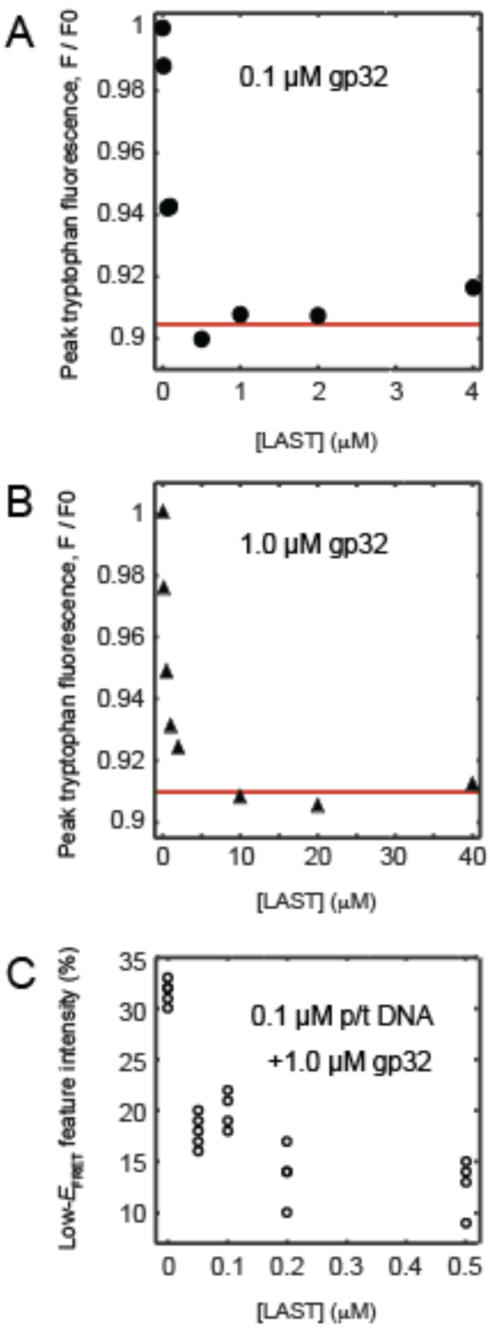


Figure S3.2. Intrinsic Tryptophan Fluorescence Vs. Concentration of LAST Peptide Intrinsic tryptophan fluorescence of (*A*) 0.1 µM gp32 and (*B*) 1.0 µM gp32 protein is plotted versus LAST concentration. (*C*) The relative intensity of the low-E_{FRET} histogram feature (at ~0.56) for the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct in the presence of 0.1 µM gp32 (see Fig. 3.4B), is plotted versus LAST concentration. Individual data points represent the results of separate experiments. The binding affinity of the LAST peptide for gp32 molecules is too strong to determine an accurate estimate of the binding constant. However, under these conditions the 1:1 LAST:gp32 binding interaction must be tighter than ~ 0.1 µM.

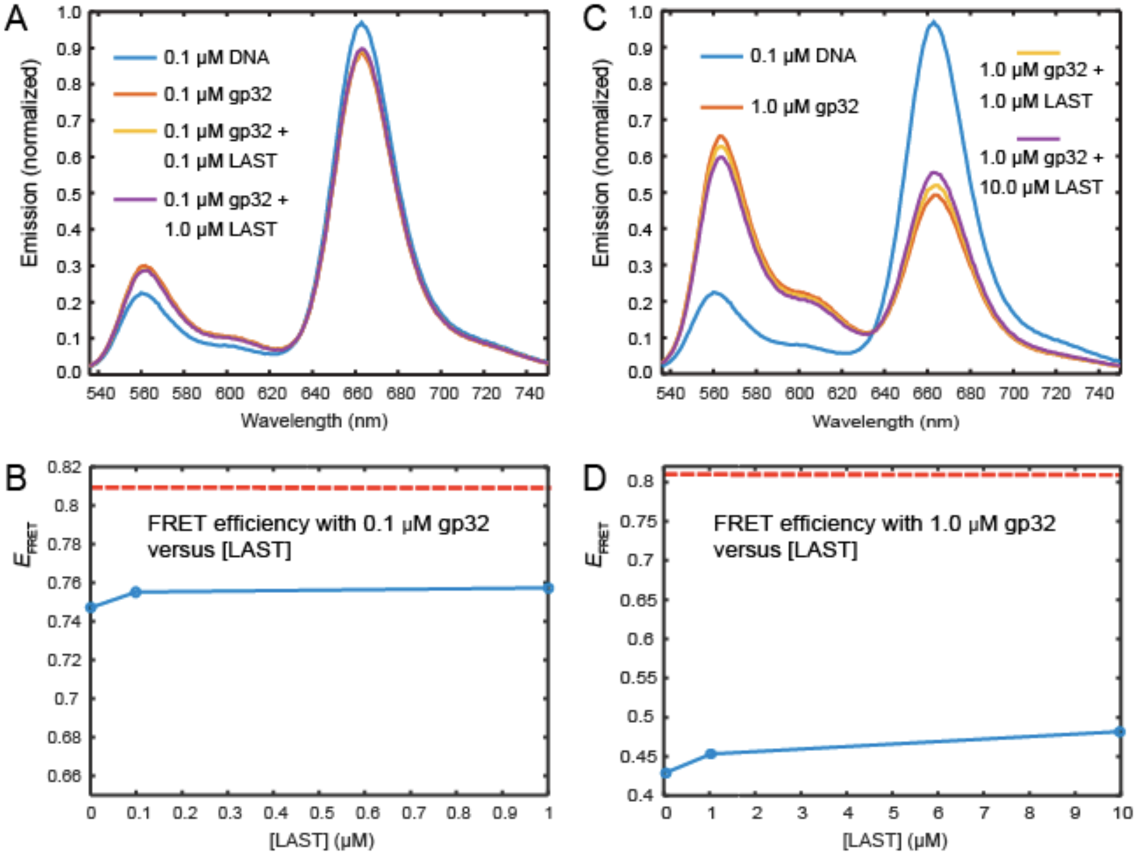


Figure S3.3. Ensemble FRET efficiency vs. Concentration Last Peptide
Bulk FRET efficiency, E_{FRET}, obtained from the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct (at 100 nM) in the presence of 0.1 μM gp32 (panels *A* and *B*) and 1.0 μM gp32 (panels *C* and *D*), and as a function of LAST concentration. The bulk FRET efficiency is calculated from the peak Cy3 and Cy5 intensities according to $E_{FRET} = I_{Cy5}/\left(I_{Cy3} + I_{Cy5}\right)$.

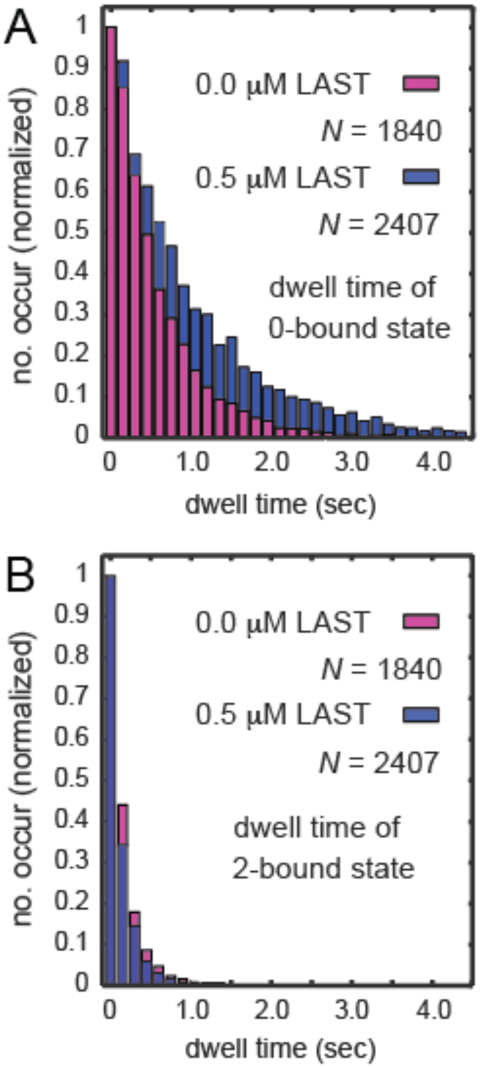


Figure S3.4. Dwell Time Distributions from HMM Analysis
Dwell time histograms of the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA and 0.1 μM gp32 in the presence (blue) and absence (red) of 0.5 μM LAST. (*A*) Dwell times of high-E_{FRET} (0-bound) state, and (*B*) low-E_{FRET} (2-bound) state. Histograms were constructed from the results of HMM analysis of smFRET trajectories, as described in the text. The number *N* of state-to-state transitions in each case are indicated.

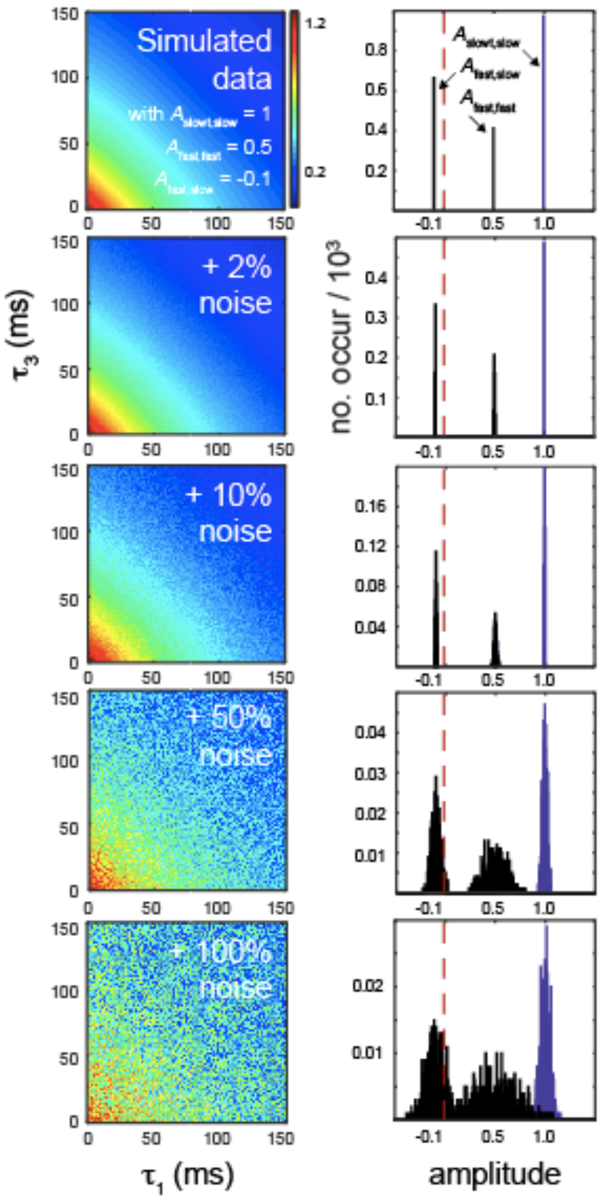


Figure S3.5. Simulated 4th-Order TCFs and Its Inverse Laplace Transform
Simulated 4th-order TCFs (left column) and associated spectral peak amplitudes (right column) obtained from our inverse Laplace transform (ILT) algorithm for an $N = 3$ system, as described in the SI text. The 4th-order TCF is given by Eq. (S3.7), and its ILT by Eq. (S3.8). The effect of noise on the precision of peak amplitude determination is shown for various levels of noise (0 – 100% error). Robust determination of the peak amplitudes was possible for 4th-order TCFs with noise levels up to 50%, which we found to be sufficient for the analysis of our smFRET data.

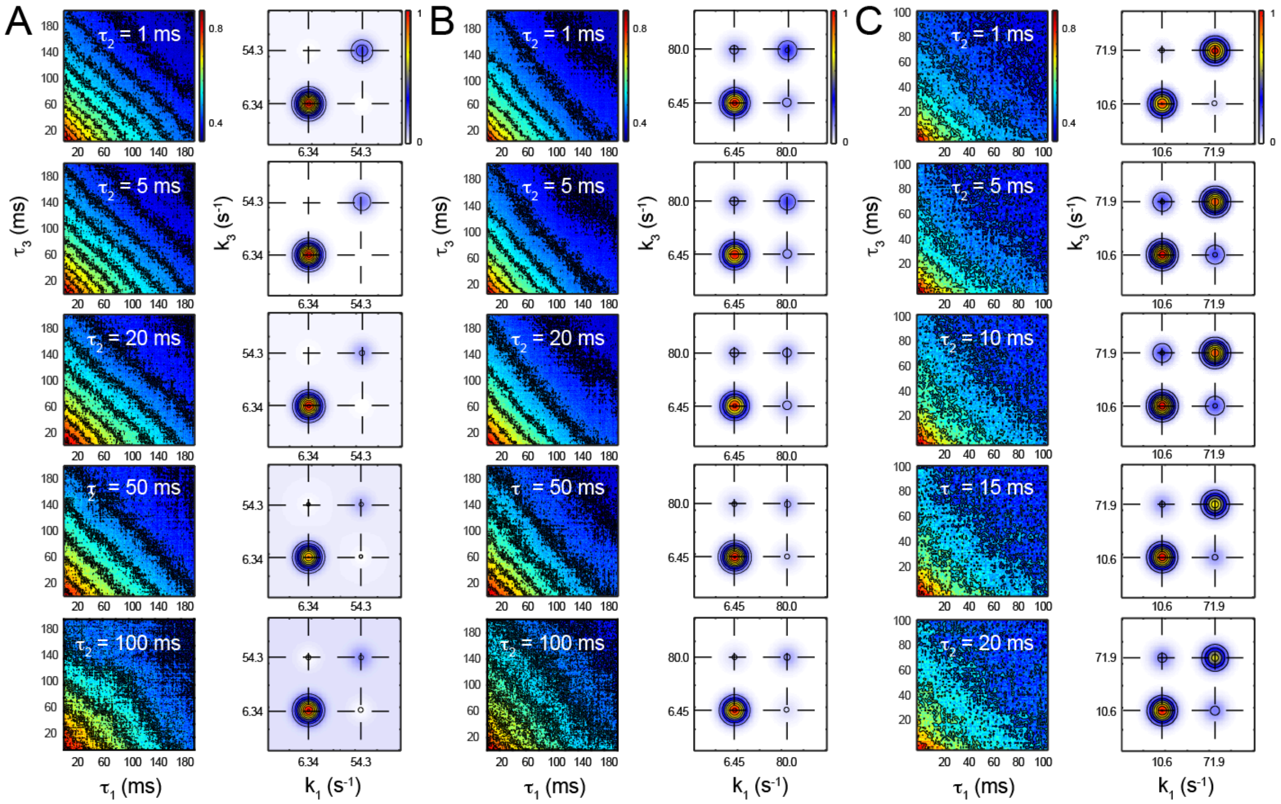


Figure S3.6. Experimental 4th Order TCFs And Associated Rate Spectra
(*A*) and (*B*) 4th-order TCFs (left columns) and associated two-dimensional (2D) rate spectra (right columns) calculated from smFRET trajectories of the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct in the presence of (*A*) 0.1 µM gp32; (*B*) 0.1 µM gp32 and 1.0 µM LAST; and (*C*) 1.0 µM gp32. The 2D rate spectra were calculated assuming a three-state ($N = 3$) scheme, and the characteristic rate constants were determined from the corresponding 2nd-order TCFs (see Table S3.2). Each data set was computed from hundreds-of-thousands of points.

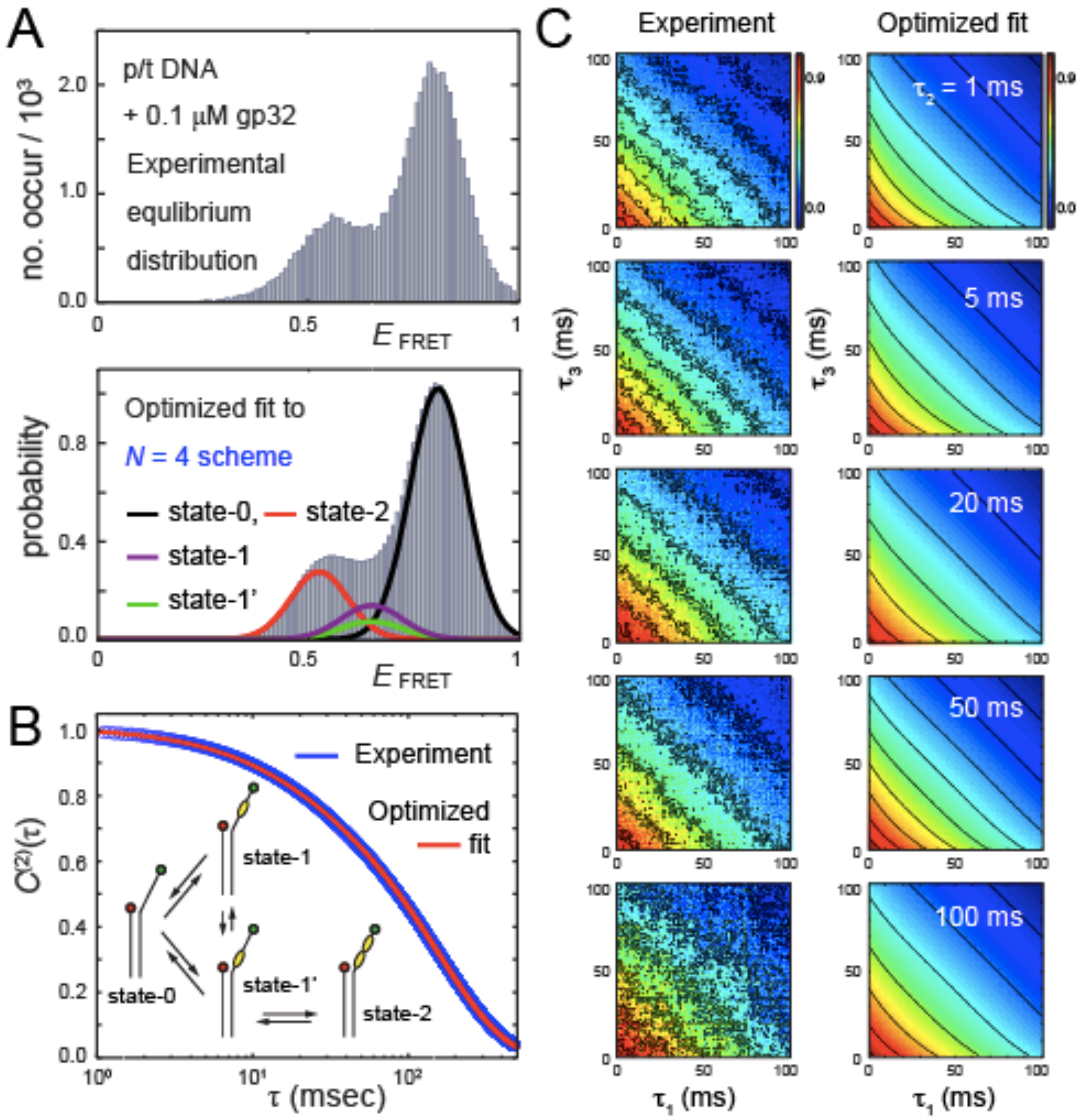


Figure S3.7. Optimized fits of the N = 4 scheme (0.1 μM gp32)

Optimized fits of the $N = 4$ scheme to experimental functions calculated from smFRET trajectories of the 3'-Cy3/Cy5-p(dT)₁₅-p/t DNA construct in the presence of 0.1 μM gp32. The data and fits are the same as shown in Fig. 3.7, except that the 4th-order TCFs are presented as contour diagrams. (*A*) Top panel: experimental equilibrium distribution of states. Bottom panel: Optimized fits show the component populations of the 0-bound state (black curve), the 2-bound state (green curve), and the 1-bound intermediate (red curve). (*B*) The experimental 2nd-order TCF (blue) is shown overlaid with the optimized fit (red). (*C*) Experimental 4th-order TCFs (left column) for various waiting times are shown in comparison to the corresponding optimized theoretical fits (right column). The cumulative fitness of optimized solutions for the $N = 3$ and 4 schemes were determined using Eq. (S3.9) and are given in Table S3.4 in Appendix C: Chapter III.

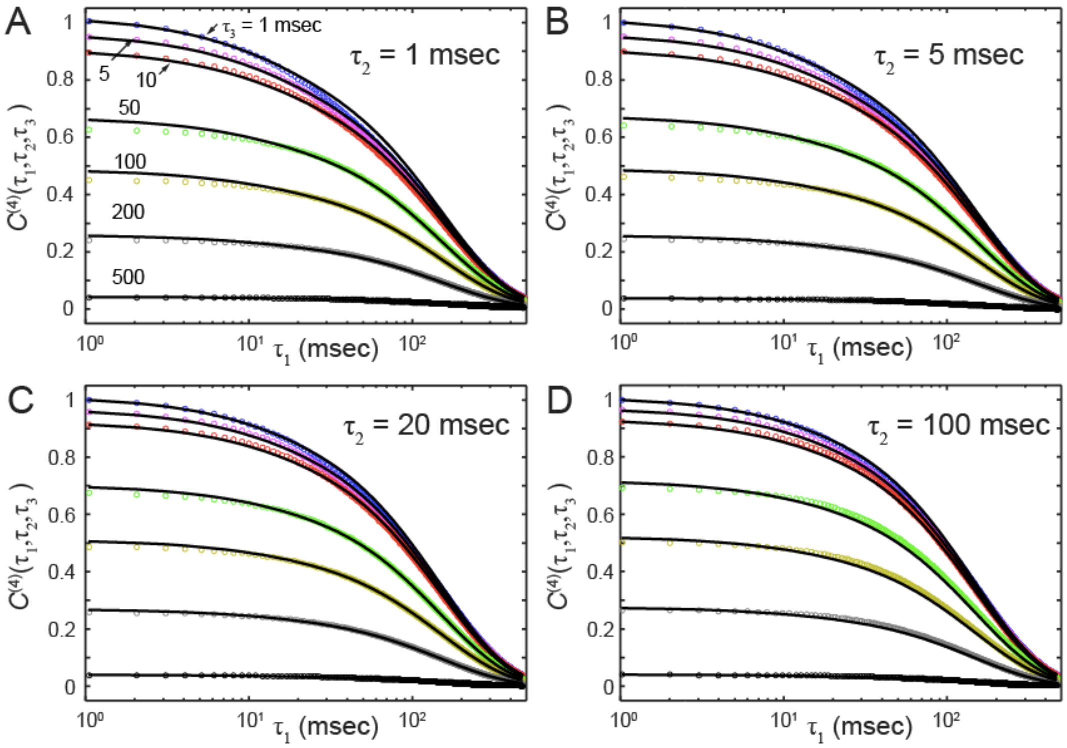


Figure S3.8. Cross-sections Of 4th Order TCF Fits N = 4 (0.1 μM gp32)

Direct comparison between optimized fits of the *N* = 4 scheme to experimental 4th-order TCFs calculated from smFRET trajectories of the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct in the presence of 0.1 μM gp32, re-plotted from panel C of Fig. S3.7. Each plot shows a set of horizontal slices of the 4th-order TCFs for various values of τ_3, and as a function of τ_1. The value of the waiting time interval τ_2 is varied for each plot: (*A*) 1 msec, (*B*) 5 msec, (*C*) 20 msec, and (*D*) 100 msec. The cumulative fitness of optimized solutions for the *N* = 3 and 4 schemes are were determined using Eq. (S3.9) and are given in Table S3.4 in Appendix C: Chapter III.

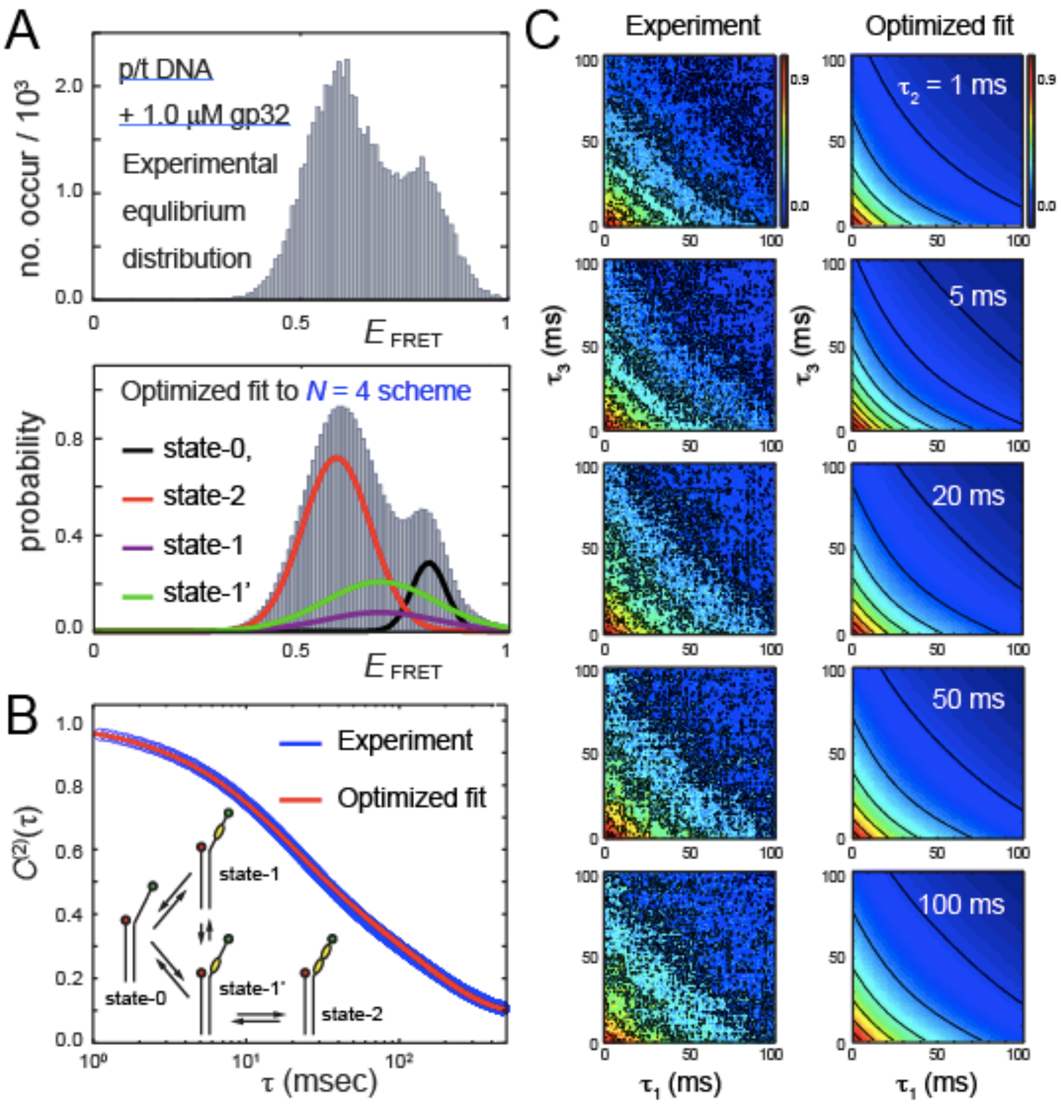


Figure S3.9. Optimized fits of the $N = 4$ scheme (1.0 μM gp32)

Optimized fits of the $N = 4$ scheme to experimental functions calculated from smFRET trajectories of the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct in the presence of 1.0 μM gp32. (*A*) Top panel: experimental equilibrium distribution of states. Bottom panel: Optimized fits show the component populations of the 0-bound state (black curve), the 2-bound state (green curve), and the 1-bound intermediate (red curve). (*B*) The experimental 2nd-order TCF (blue) is shown overlaid with the optimized fit (red). (*C*) Experimental 4th-order TCFs (left column) for various waiting times are shown in comparison to the corresponding optimized theoretical fits (right column). The cumulative fitness of optimized solutions for the $N = 3$ and 4 schemes were determined using Eq. (S3.9) and are given in Table S3.4 in Appendix C: Chapter III.

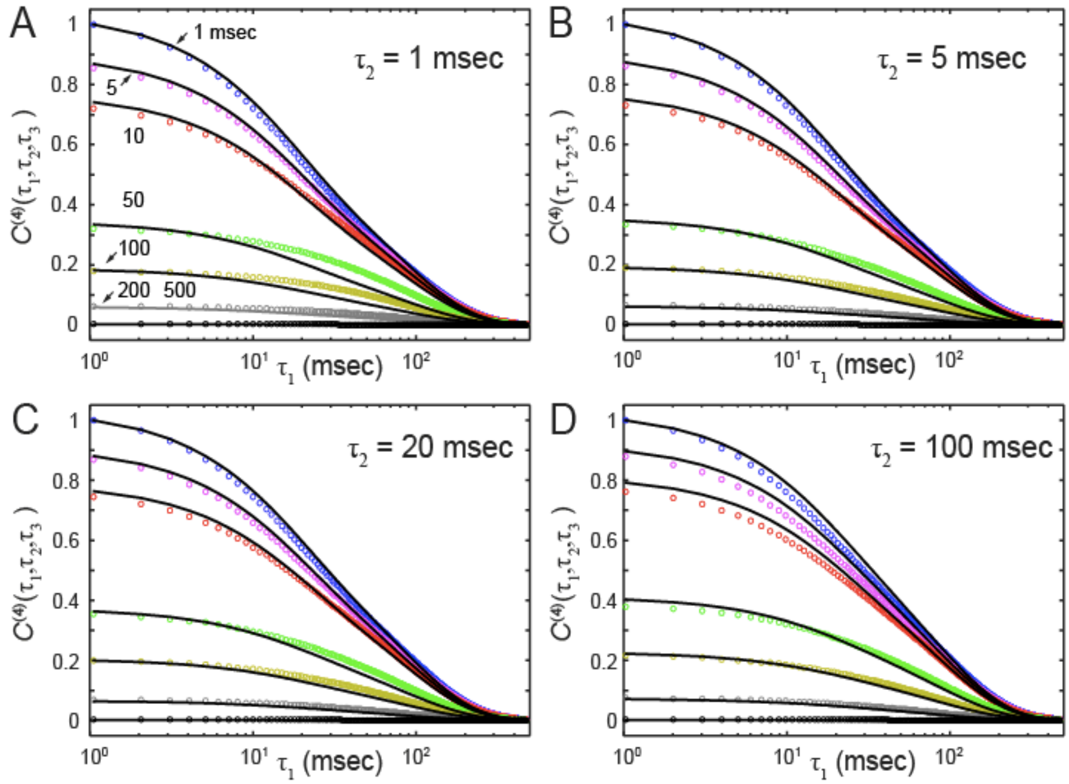


Figure S3.10. Cross-sections Of 4th Order TCF Fits N = 4 (1.0 µM gp32)
Direct comparison between optimized fits of the $N = 3$ scheme to experimental 4th-order TCFs calculated from smFRET trajectories of the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct in the presence of 1.0 µM gp32, re-plotted from panel C of Fig. S3.9. Each plot shows a set of horizontal slices of the 4th-order TCFs for various values of τ_3, and as a function of τ_1. The value of the waiting time interval τ_2 is varied for each plot: (*A*) 1 msec, (*B*) 5 msec, (*C*) 20 msec, and (*D*) 100 msec. The cumulative fitness of optimized solutions for the $N = 3$ and 4 schemes were determined using Eq. (S3.9) and are given in Table S3.4.

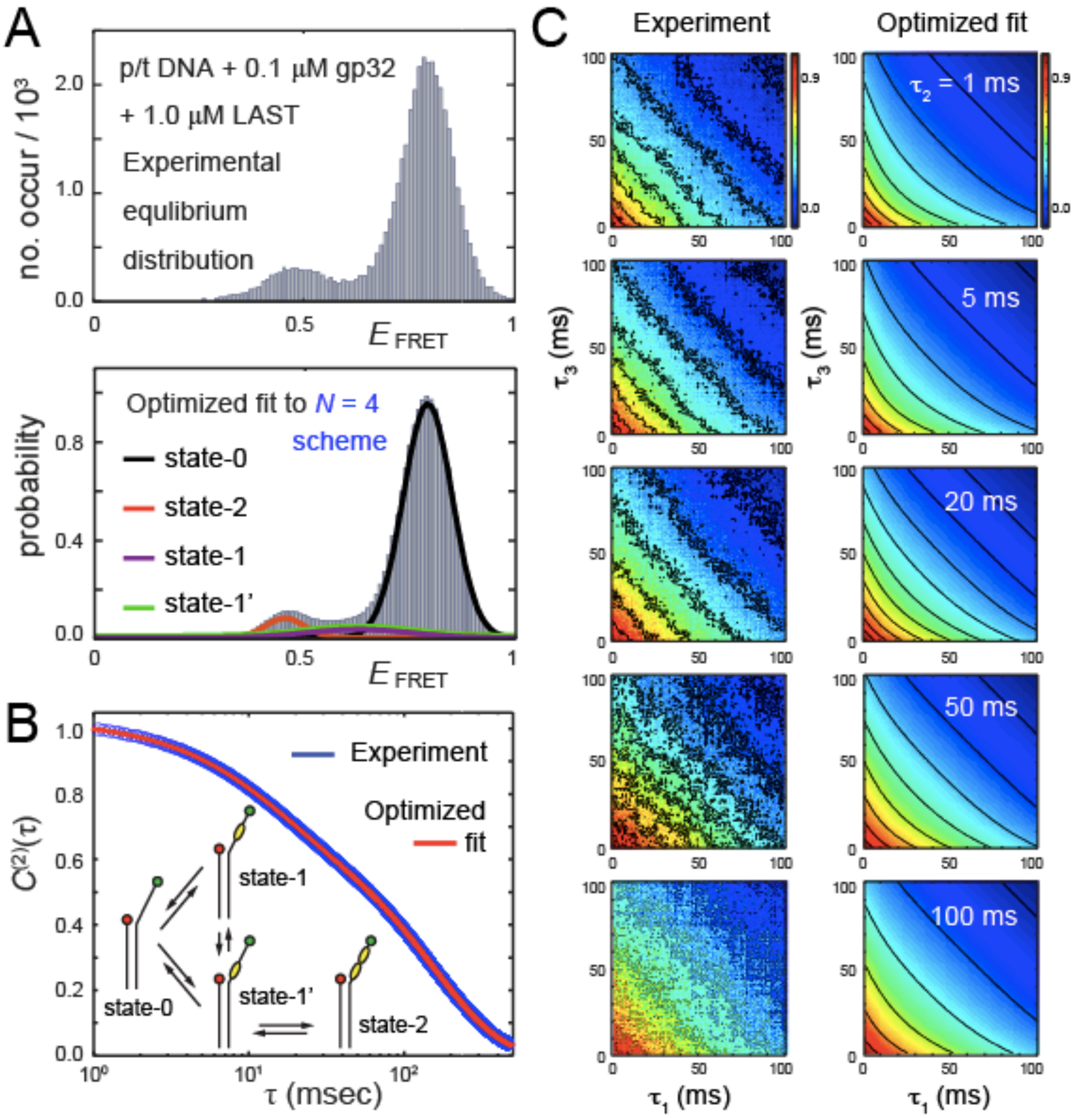


Figure S3.11. Optimized fits of the *N* = 4 scheme (With LAST)

Optimized fits of the *N* = 4 scheme to experimental functions calculated from smFRET trajectories of the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct in the presence of 0.1 µM gp32 plus 1.0 µM LAST. (*A*) Top panel: experimental equilibrium distribution of states. Bottom panel: Optimized fits show the component populations of the 0-bound state (black curve), the 2-bound state (green curve), and the 1-bound intermediate (red curve). (*B*) The experimental 2nd-order TCF (blue) is shown overlaid with the optimized fit (red). (*C*) Experimental 4th-order TCFs (left column) for various waiting times are shown in comparison to the corresponding optimized theoretical fits (right column). The cumulative fitness of optimized solutions for the *N* = 3 and 4 schemes were determined using Eq. (S3.9) and are given in Table S3.4.

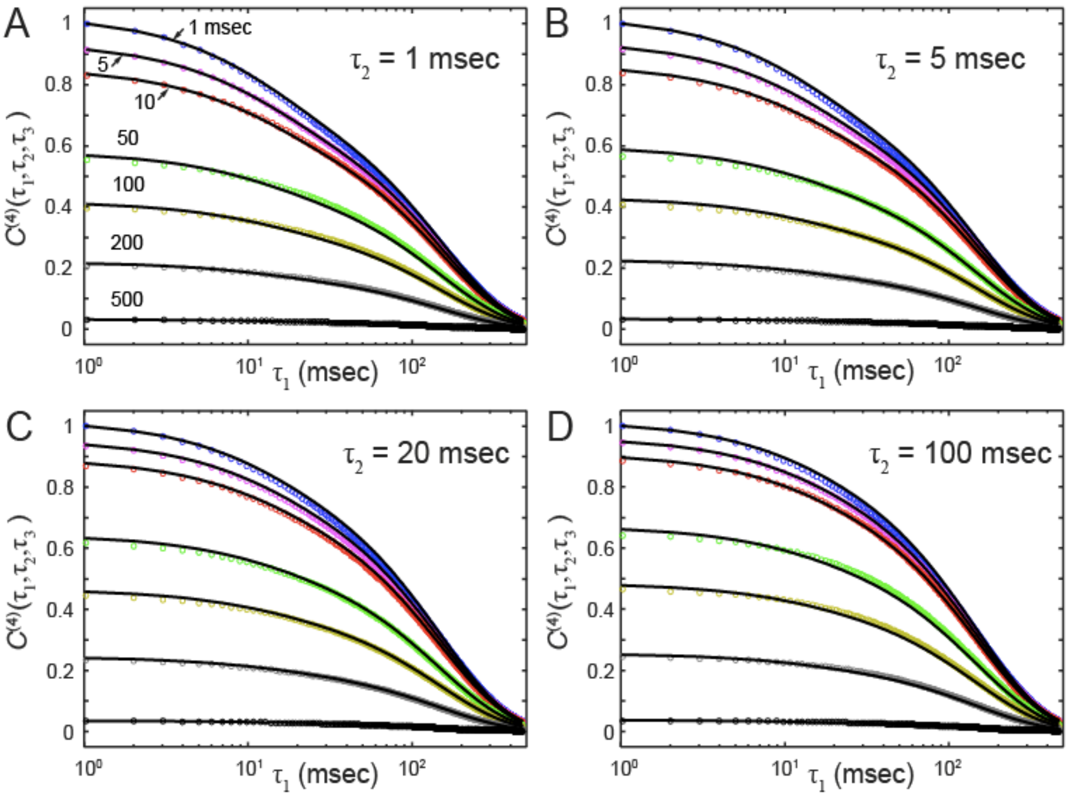


Figure S3.12. Cross Sections Of 4th Order TCF Fits N = 4

Direct comparison between optimized fits of the *N* = 4 scheme to experimental 4th-order TCFs calculated from smFRET trajectories of the 3’-Cy3/Cy5-p(dT)$_{15}$-p/t DNA construct in the presence of 0.1 µM gp32 plus 1.0 µM LAST, re-plotted from panel C of Fig. S3.11. Each plot shows a set of horizontal slices of the 4th-order TCFs for various values of τ_3, and as a function of τ_1. The value of the waiting time interval τ_2 is varied for each plot: (*A*) 1 msec, (*B*) 5 msec, (*C*) 20 msec, and (*D*) 100 msec. The cumulative fitness of optimized solutions for the *N* = 3 and 4 schemes were determined using Eq. (S3.9) and are given in Table S3.4.

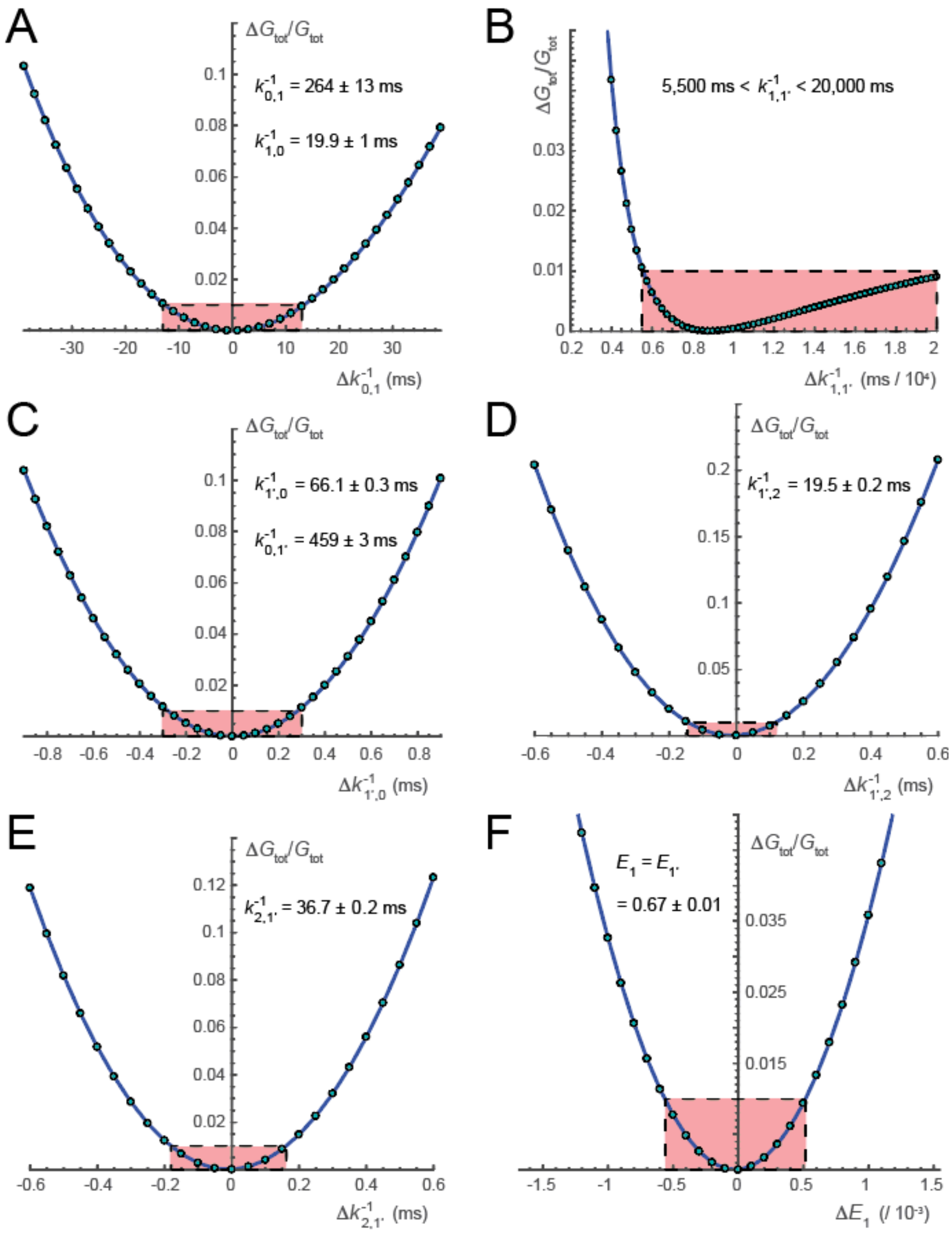


Figure S3.13. Error Analysis

Relative deviation of the target function, $\Delta G_{tot}/G_{tot}$, from the optimized value, $G_{tot,ref}$, as a function of the rate constant and FRET efficiency parameter uncertainties. Cross-sections of the target function given by Eq. (S3.9) are shown for the optimization to the *N* = 4 scheme under the 0.1 µM gp32 condition, and for the uncertainties (*A*) $\Delta k_{0,1}^{-1}$, (*B*) $\Delta k_{1,1'}^{-1}$, (*C*) $\Delta k_{1',0}^{-1}$, (*D*) $\Delta k_{1',2}^{-1}$, (*E*) $\Delta k_{2,1'}^{-1}$, and (*F*) ΔE_1. Here $\Delta x = x - x_{ref}$, and x_{ref} is the value corresponding to the optimized set of parameters, which corresponds to a minimum of the multi-dimensional parameter surface. The red-shaded rectangles indicate the 'trust-interval' based on a 1% relative deviation, which we associate with the experimental data quality. A similar analysis was performed for the other experimental conditions, and the cross sections are similar to those shown above. The associated error bars reported in Table 3.1 are rounded up to two significant figures. We note that the rate constants associated with Panels (*A*) – (*C*) are related by the detailed balance condition, and the error bars of forward and backward elementary steps were determined accordingly.

Figures for Chapter IV

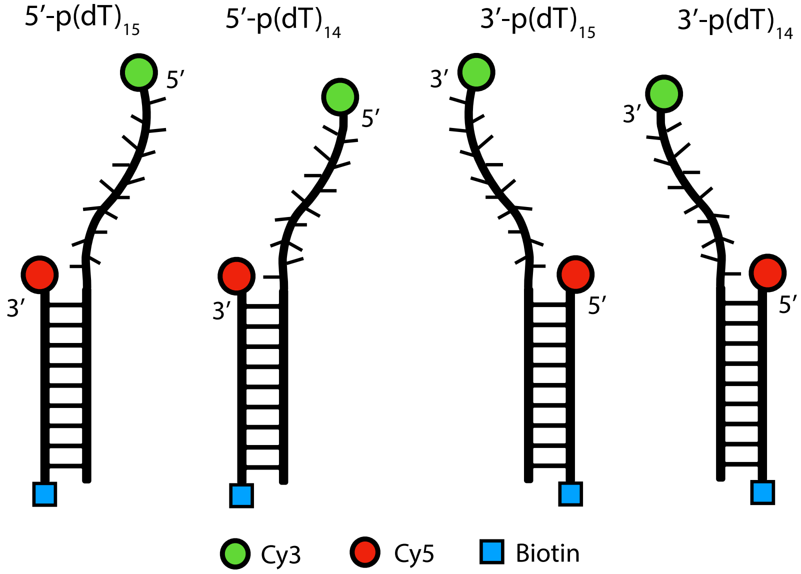


Figure 4.1 Primer/Template DNA Constructs
The four p/t DNA constructs feature an 18-nucleotide dsDNA region attached to a single-stranded (ss) DNA overhang that differs in length and polarity. Prior to the FRET donor chromophore Cy3, two of the constructs terminate in a 5`-phosphate group, and the other two terminate in a 3`-hydroxyl group. The complementary strand is labeled by the FRET acceptor Cy5 at the ss-dsDNA junction (shown in red), and a biotin residue at the opposite end (shown in blue), where the construct attaches to the slide. For both 3′ and 5′ polarities, there is a construct which consist of a ssDNA region of 14-thymine nucleotides and a construct with slightly longer stretch of 15-thymine nucleotides...

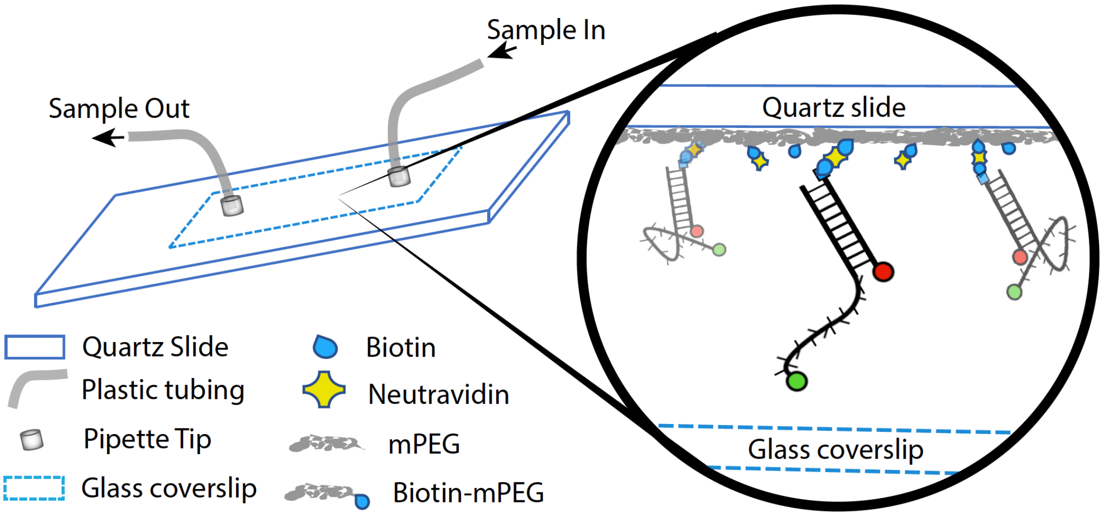


Figure 4.2. Microfluidic Sample Chamber
The sample chamber is a combination of a quartz slide epoxied to a 0.2 mm thick glass coverslip. The quartz slide has plastic tubing inserted into drilled holes that enable the experimenter to flow in solutions. The quartz slide has been chemically modified to have a layer of mPEG and Biotin-mPEG, supporting the DNA constructs to bind. The figure shows three DNA molecules at various distances and conformations.

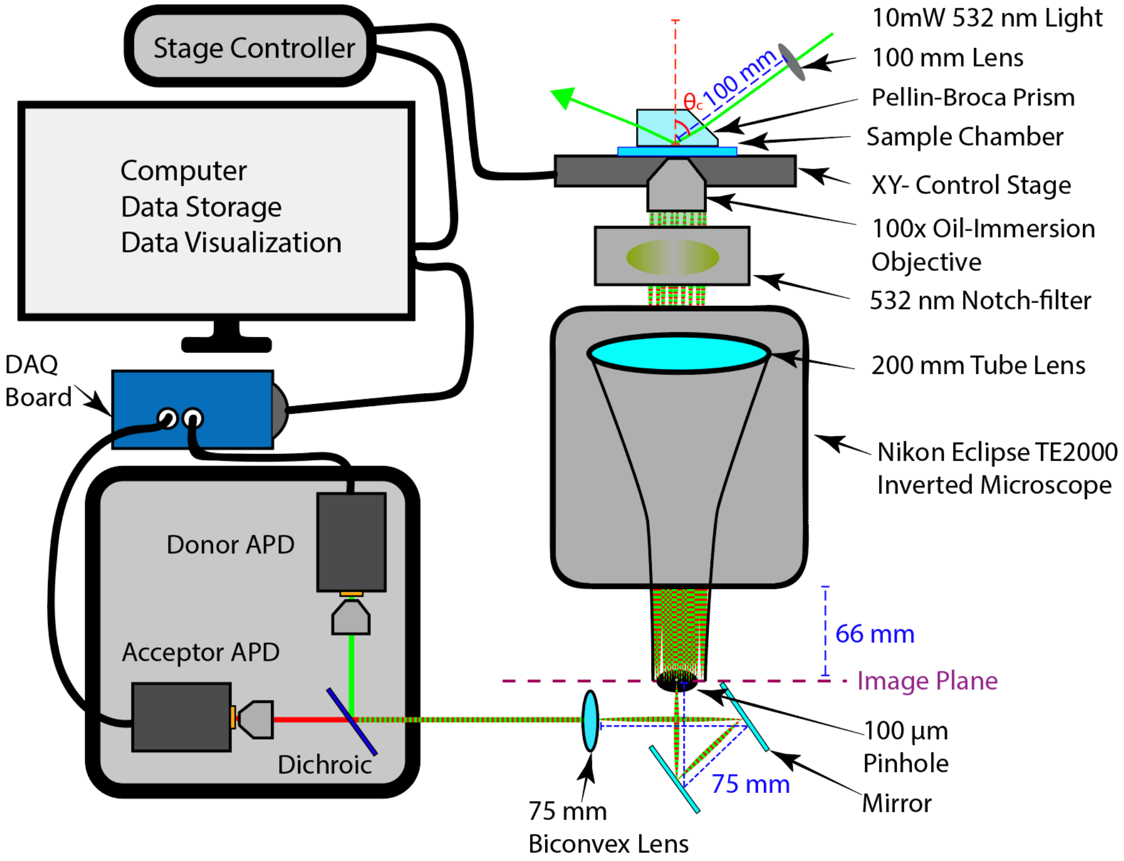
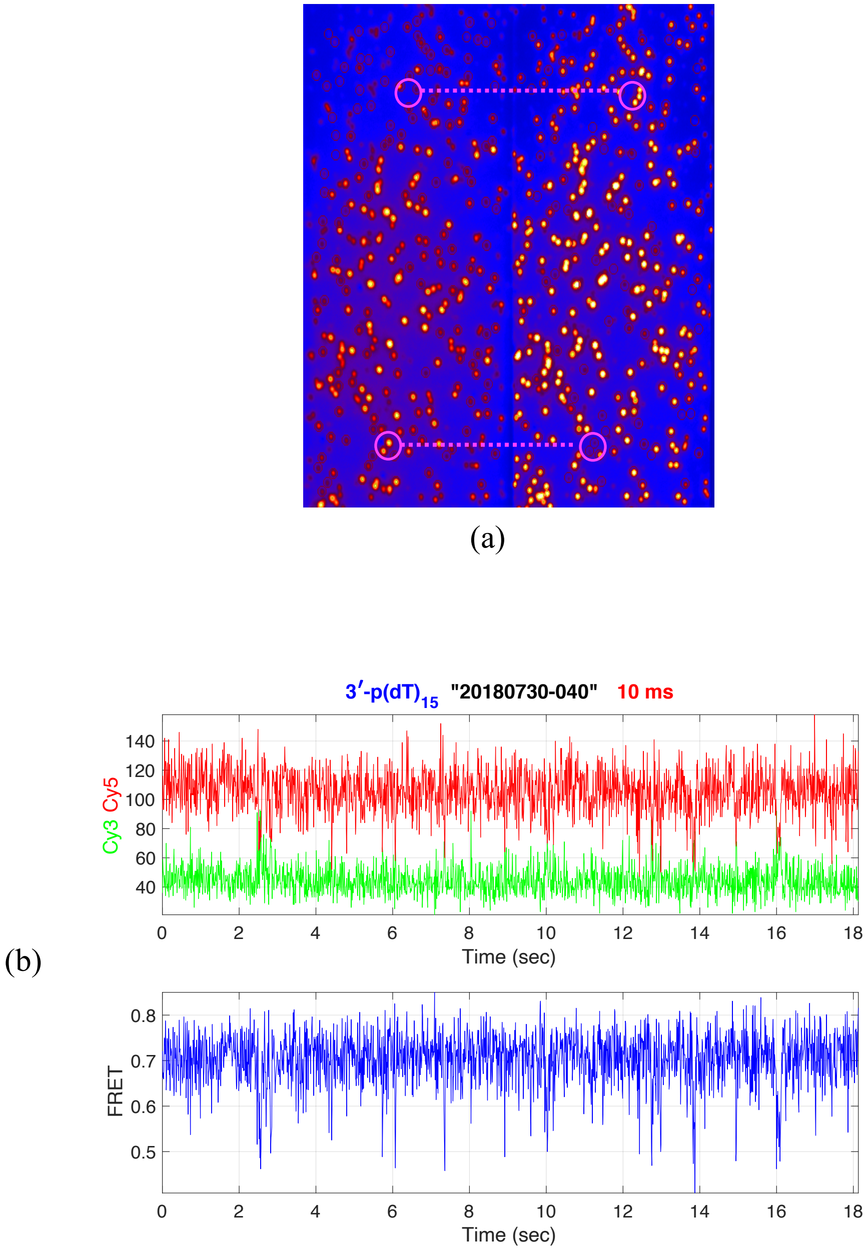


Figure 4.3 Single-Molecule Instrumentation
The collection geometry used to perform microsecond-resolved single-molecule FRET. Starting at the top-right of the figure is the excitation geometry with the sample. The fluorescence is magnified by the inverted microscope and focused to the image plane. At the image plane, we block the fluorescence of all but one molecule with a 100 µM pinhole. The collimating-lens is placed its focal length from the image plane such that the collimated fluorescence is translated without loss to a covered box with the instrumentation necessary for microsecond-resolved smFRET. A longpass dichroic beamsplitter separates the donor-acceptor fluorescence at 635 nm to 60x objectives that focus the collimated fluorescence into two Avalanche Photo Diodes (APDs). From there, the signal is digitalized by the Data Acquisition Board (DAQ) where it is relayed to the computer for analysis.


(a)

(b)

Figure 4.4 EMCCD Image of Single-Molecule DNA Constructs
(a) EMCCD Image: 512 × 512 pixel, 8-bit tiff image taken with 30 ms exposure. The left-hand side of panel a) is the fluorescence from the donor channel and the right-hand side is the fluorescence of the acceptor channel. The colored circles mark two groups of molecules for illustrative purposes. In the top, the two pairs of molecules both coincidentally have a high FRET state (donor dimmer, acceptor brighter); the trio at the bottom feature two molecules in a low FRET state (donor brighter, acceptor dim) and one in a high FRET state. (b) Example single-molecule trajectory binned to 10 ms resolution. The trajectory ends at 18 seconds because of an irreversible photobleaching event.

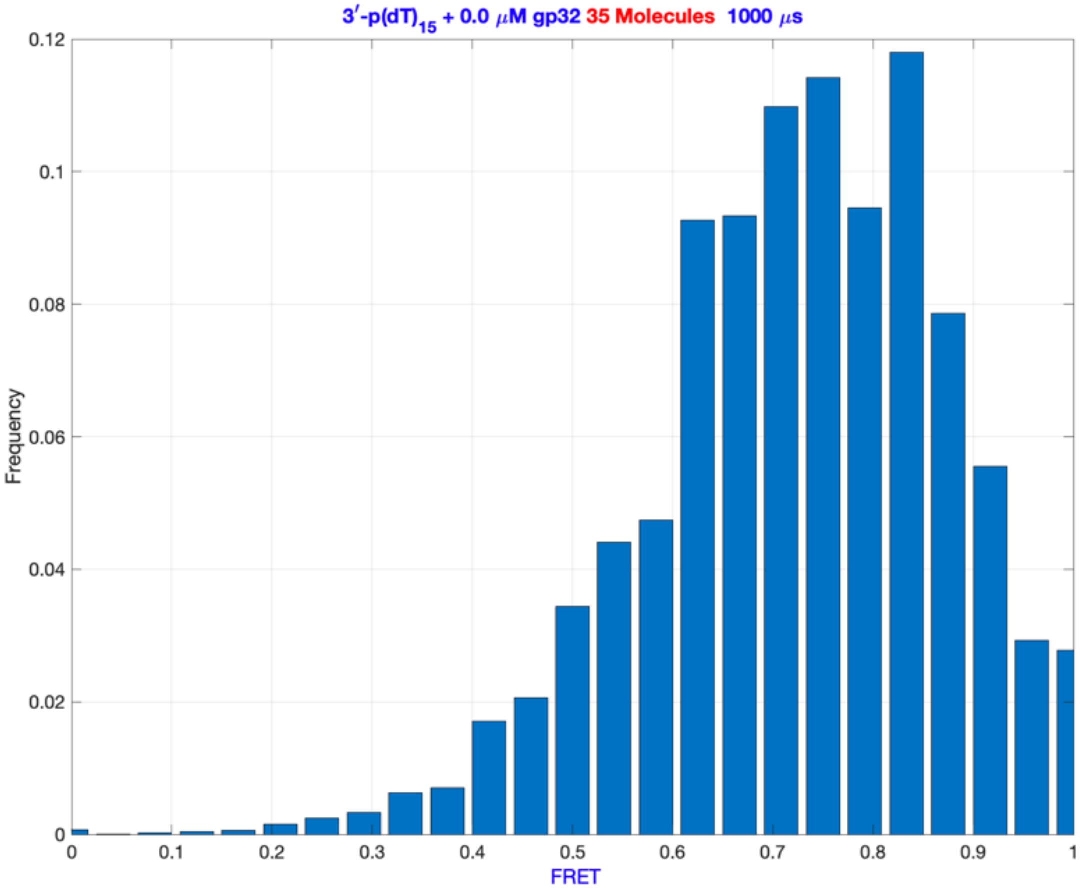


Figure 4.5: Probability Distribution of Conformational Macrostates
The normalized probability distribution of FRET states suggests that the distribution is described by a single feature, but rather by a combination of underlying states. This distribution is often called a FRET histogram. Each histogram is composed of hundreds of thousands of FRET values that were visited by dozens of single-molecules during the 30 second scans. This histogram is the result of binning the microsecond-detected fluorescence into 1000 µs chunks before calculating the FRET with the formula $F = \frac{I_A}{I_A+I_D}$, where $I_D$is the intensity of the donor chromophore and I_A is the intensity of the acceptor chromophore.

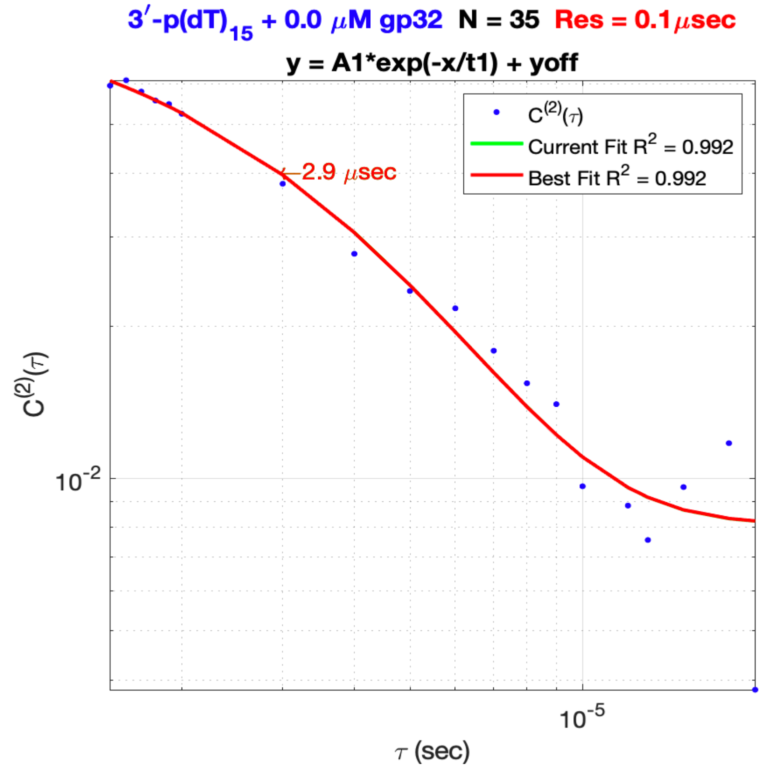


Figure 4.6 Fits to the first 20 µs of the Experimental 2nd Order TCF

The first 20 µs of the autocorrelation function for each construct was fit to a single exponential using the equation $C^{(2)} = Ae^{-kt} + y_{offset}$. The blue dots are the experimentally derived 2nd order TCF calculated using Eq. 4.7, and the red line is the best fit to the 2nd order TCF data, calculated with MATLAB's *fmincon*, a least-square minimization function. The value reported is the inverse rate $\tau = k^{-1}$. Each DNA construct exhibited a single exponential decay around $\tau = 3$ µs. We believe this initial decay to be a result of photophysical processes of the chromophores.

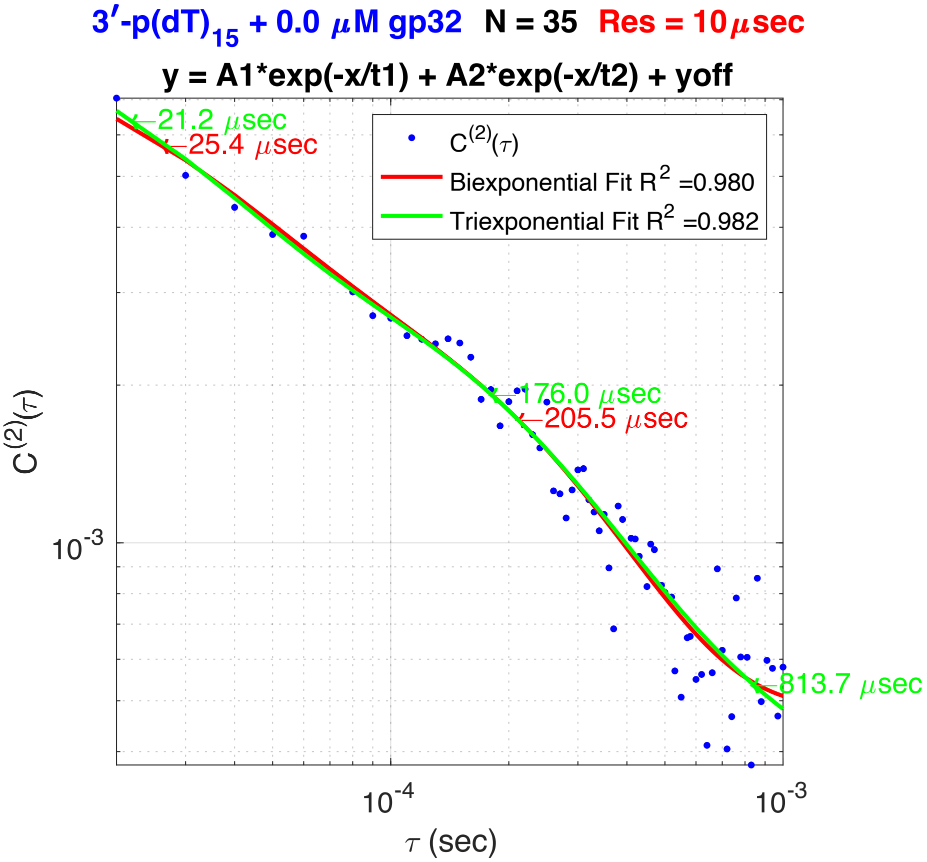


Figure 4.7: Exponential Fits to 2nd Order TCF

The autocorrelation function of single-molecule FRET data from DNA primer-template junctions past 20 µs has at least two non-trivial eigenmodes. A bi-exponential fit was able to capture the majority of the information present in the 2nd order TCFs; using an extra exponential only improves the fit marginally. This implies that our single-molecule measurements are sensitive to at least three interconverting conformational states in the ssDNA constructs.

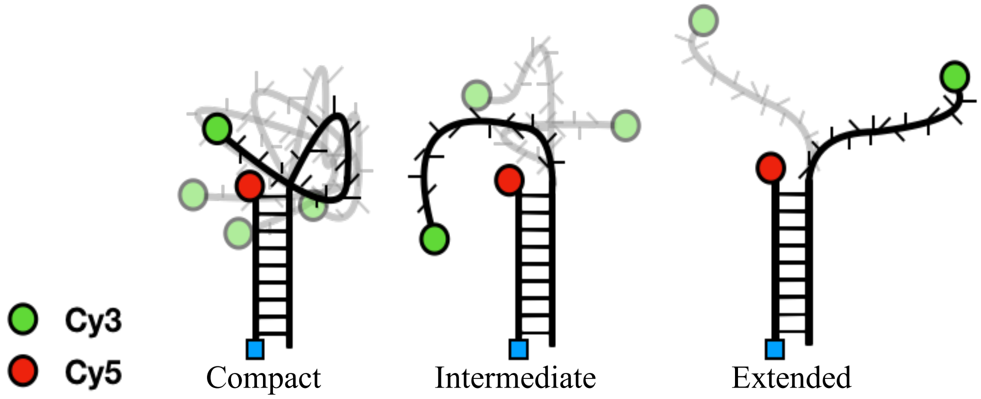


Figure 4.8 ssDNA Conformational Macrostates
The three conformational macrostates our analysis suggests are present in the ssDNA backbones near a p/t DNA junction. The light grey lines are other possible microstates within each macrostate. The compact macrostate gives rise to the common high-FRET values, the extended conformation is responsible for the low FRET values, and the intermediate-state yields FRET values in between.

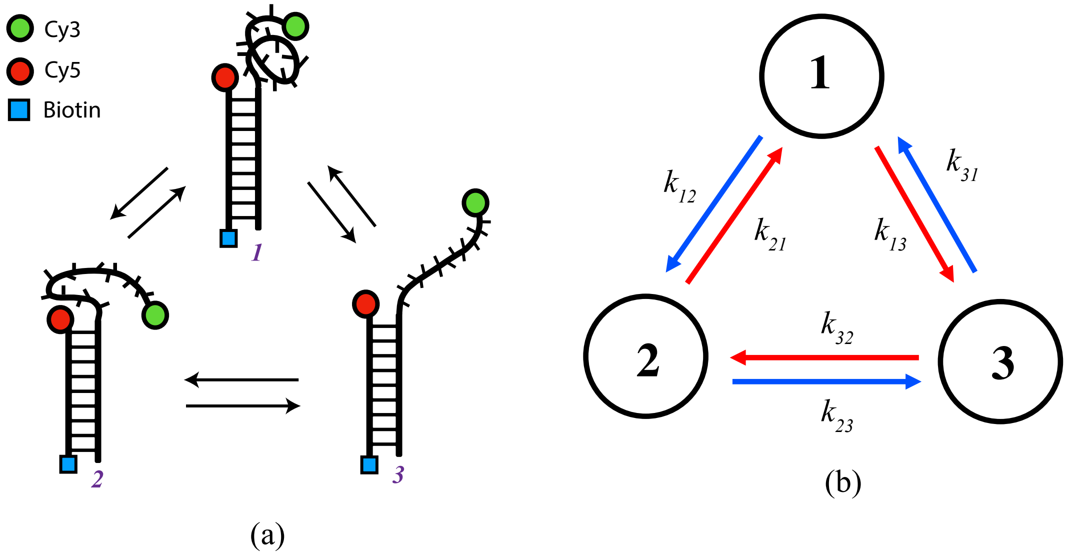


Figure 4.9. Cyclical Three-State Model
(a) The three-state model of ssDNA conformational fluctuations with representative structures for each macrostate. State-1 is the compact macrostate, state-2 is the intermediate macrostate, and State-3 is the extended conformation. (b) A schematic model with rate constants k_{ij} between states-i and -j.

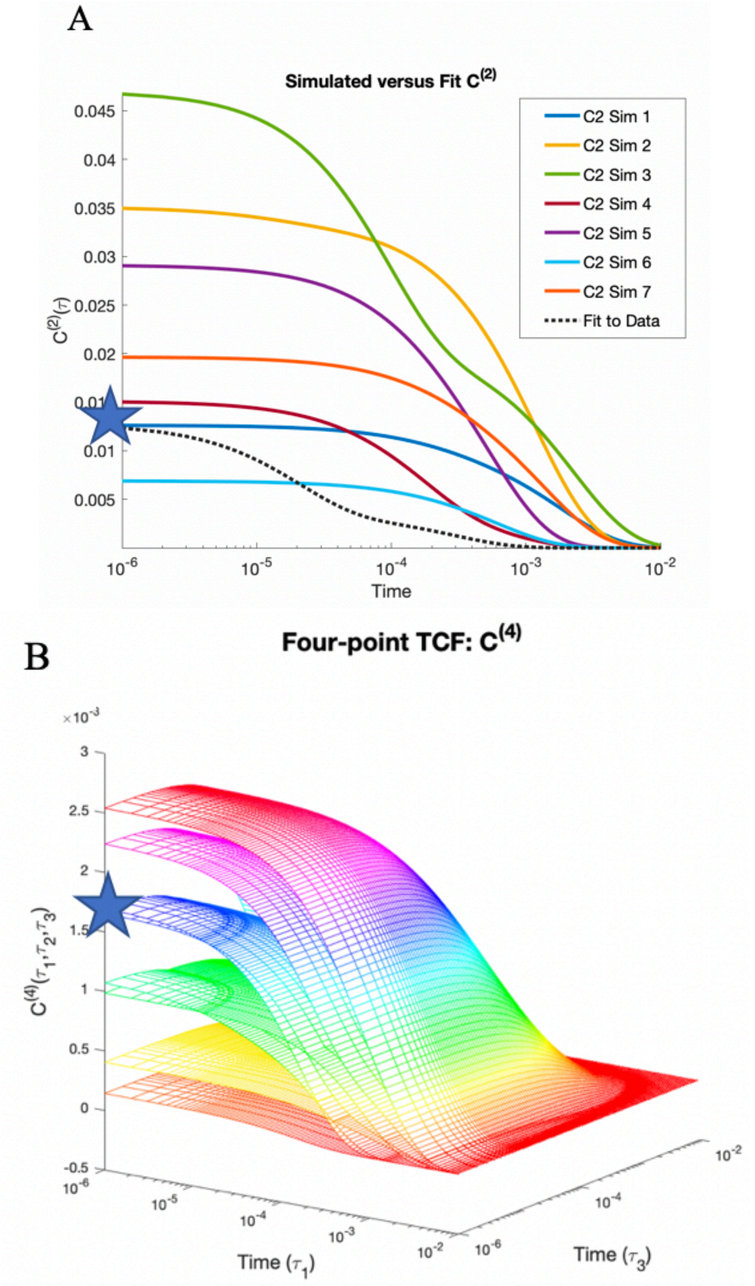


Figure 4.10: Simulations of 2nd and 4th Order Time Correlation Functions
(A) The two-point correlation functions simulated with randomly selected input parameters for the rate constants and FRET values using Eq. (4.7). The starred and dashed line represents the TCF which best fits the data. Note that the various autocorrelation functions have different initial values and decay rates. (B) Corresponding simulated 4th order TCFs using Eq. (4.10). The surface which best simulates the data is again marked with a star. Both panels (A) and (B) demonstrate the dependence of the correlation functions on the values of the input parameters.

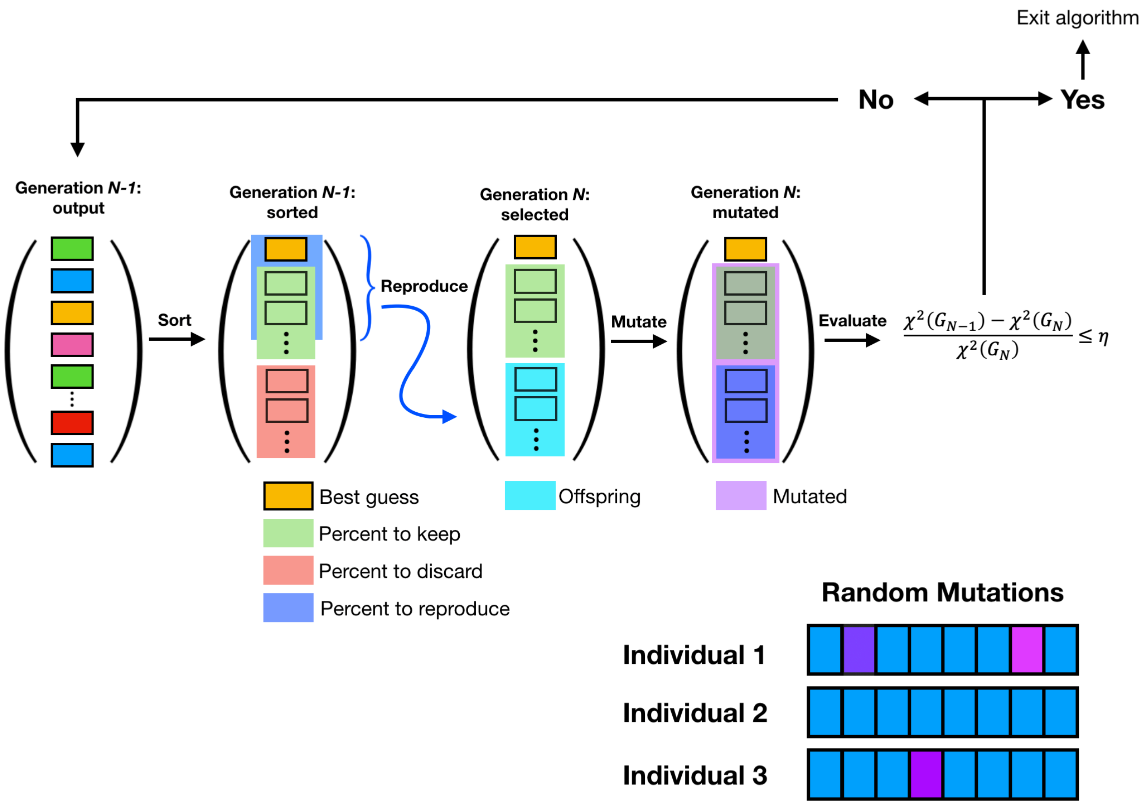


Figure 4.11 Genetic Algorithm

This schematic explains the various steps of the genetic algorithm (GA), which is an optimization strategy analogous to evolution by natural selection. The GA consists of four basic parts which occur each generation: initialization, selection, mutation, and evaluation. 1) initialization: The first generation of the optimization consist of a randomly generated population of *guesses* with mixed levels of fitness. The unsorted guesses are represented by the first column. The "guesses" are the numerical arrays containing the values of the model parameters (FRET values and state-to-state rates). 2) selection: the population is sorted by fitness. The fittest individuals are those which yield the lowest valued difference function using Eq. (4.20). The best guesses are carried on to the next stage of the GA unmolested, the top tiers of guesses are allowed to mix (reproduce), and the bottom tier of guesses are removed from the population. 3) mutation: all but the best guess is randomly mutated to varying degrees as a means to ensure continued genetic diversity. 4) evaluation: the final population (each generation) is evaluated and the "best guess" is selected as the current optimization solution. The population is unsorted due to the random mutagenesis in step-3. The entire population goes though the GA again starting at step-1, as if it were the newly initialized population. However, the new population at generation N has the worst guesses from generation $N - 1$ removed and new guesses that are formed from the better of the random guesses. The result is that the overall fitness of the population has increased. If the "best guess" from the current generation is less than $\eta\%$ improved over the previous generation, the algorithm is terminated and the current best guess is chosen as the optimization solution.

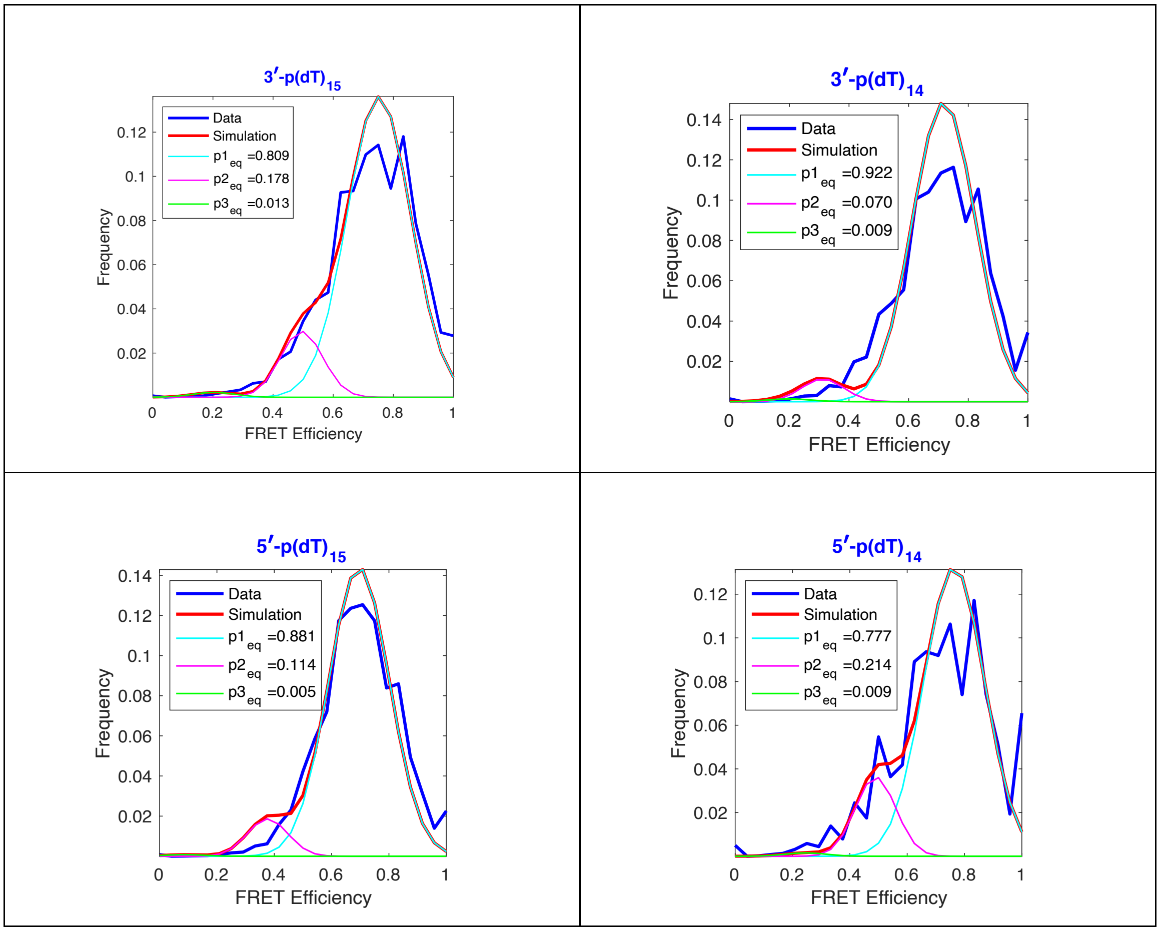


Figure 4.12 Optimized Fits of the FRET Probability Distribution Functions
Plots of the FRET Histograms for the four constructs 3`-p(dT)$_{15}$, 3`-p(dT)$_{14}$, 5`-p(dT)$_{15}$, and 5`-p(dT)$_{14}$. The blue lines represent values taken from the single molecule experiment. The thinner lines in the histogram are the individual contributions from each macrostate that sum to the cumulative simulated histogram (red): the compact macrostate (cyan), the intermediate macrostate (magenta), and the extended conformation (green). The cumulative simulated FRET distribution was calculated using Eq. (4.3).

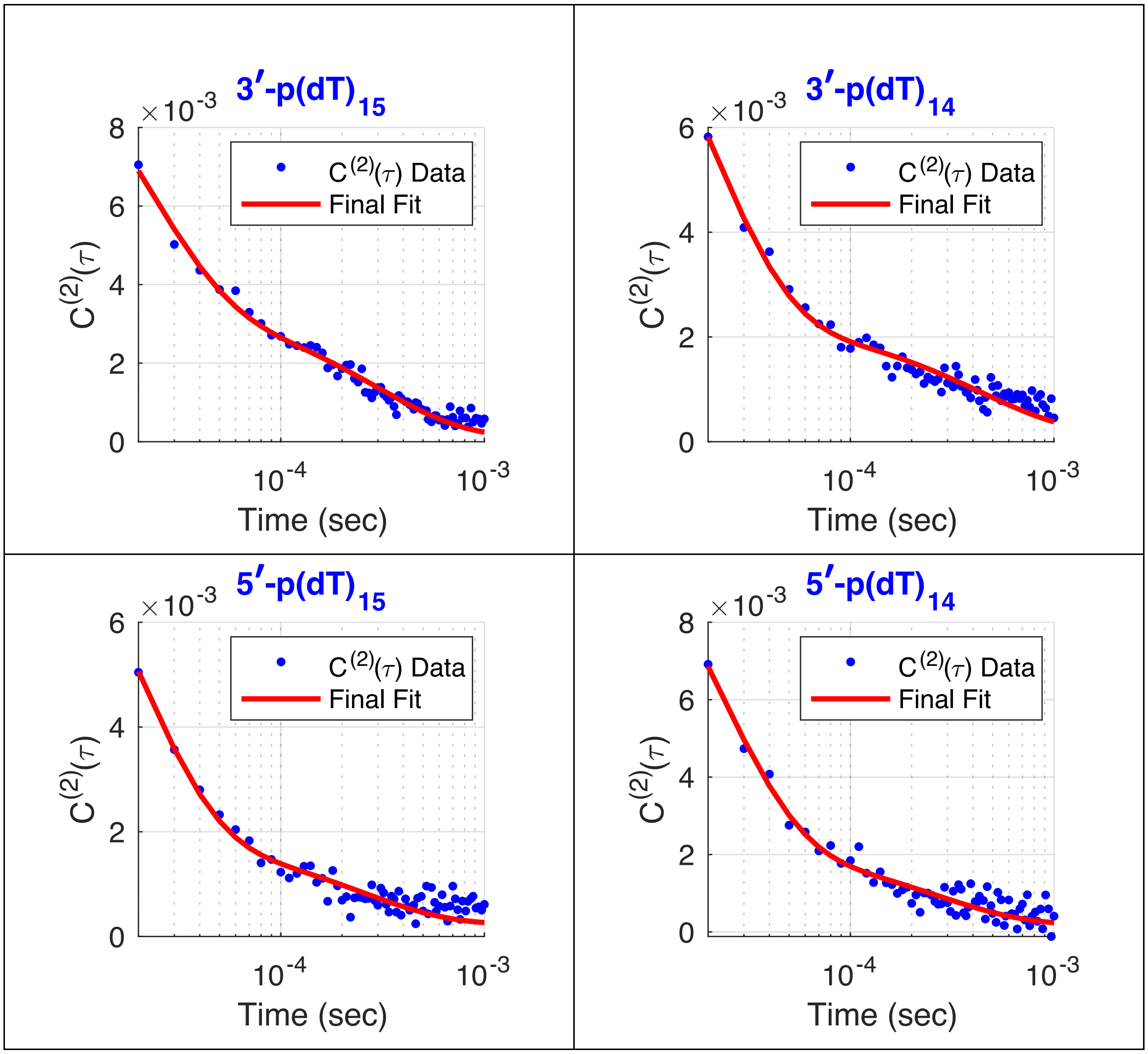


Figure 4.13 Optimized Fits of the 2nd Order TCFs

Plots of the two-point TCF for the four constructs 3`-p(dT)$_{15}$, 3`-p(dT)$_{14}$, 5`-p(dT)$_{15}$, and 5`-p(dT)$_{14}$. The blue data points represent values derived from the single molecule experiments calculated using Eq. (4.6). The red lines are the simulated 2nd order time correlation functions calculated using Eq. (4.7).

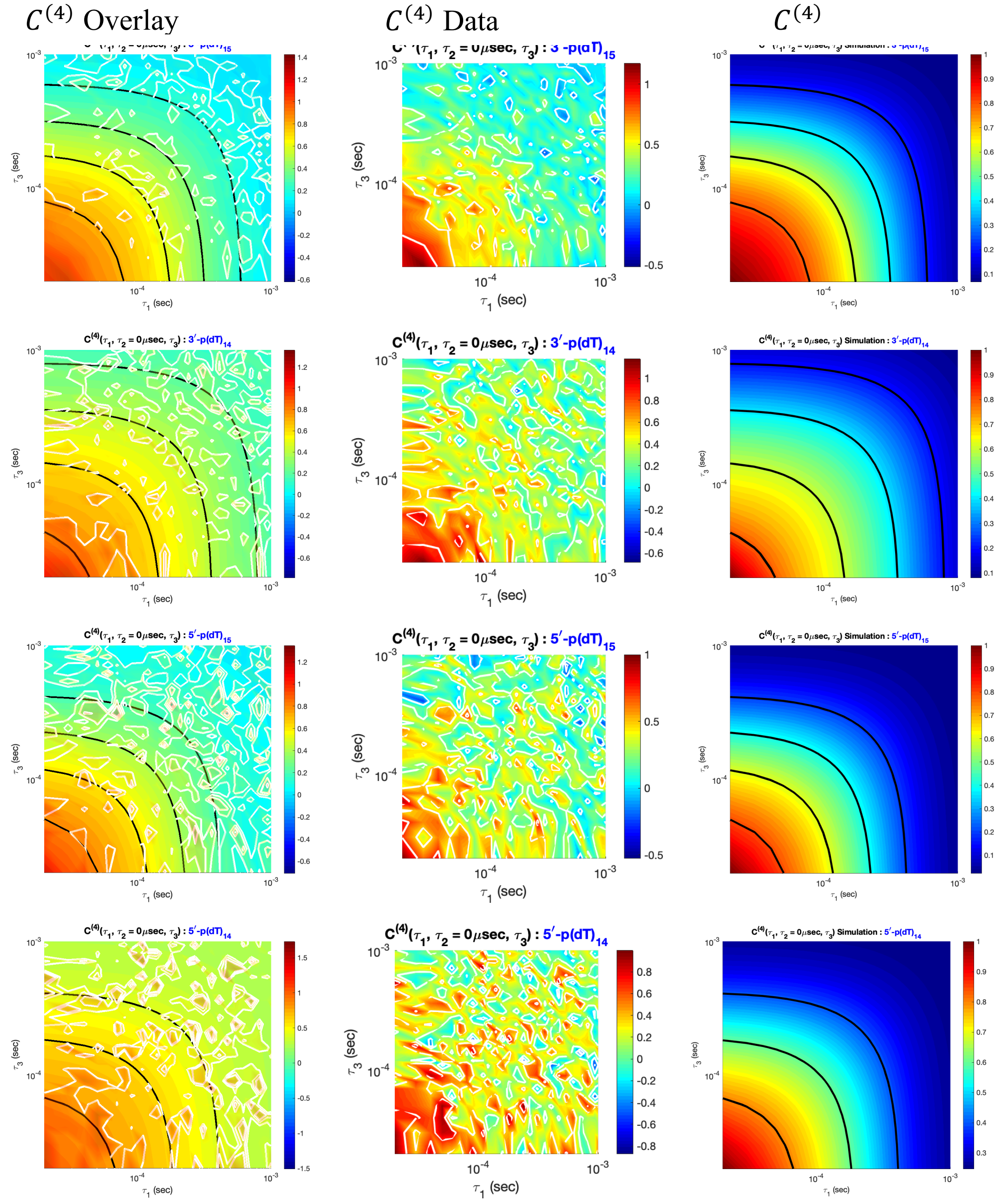


Figure 4.14 Optimized Fits of the 4th Order TCFs

Plots of the simulated 4th order TCF fit overlaid with the experimental 4th order TCF (column 1), the experimental 4th order TCF (column 2), and the simulated 4th order TCF (column 3) corresponding to the four p/t DNA constructs 3`-p(dT)$_{15}$, 3`-p(dT)$_{14}$, 5`-p(dT)$_{15}$, and 5`-p(dT)$_{14}$ (rows 1-4, respectively). The experimental 4th order TCF was calculated using Eq. (4.9), and the simulated 4th order TCF was calculated using Eq. (4.10).

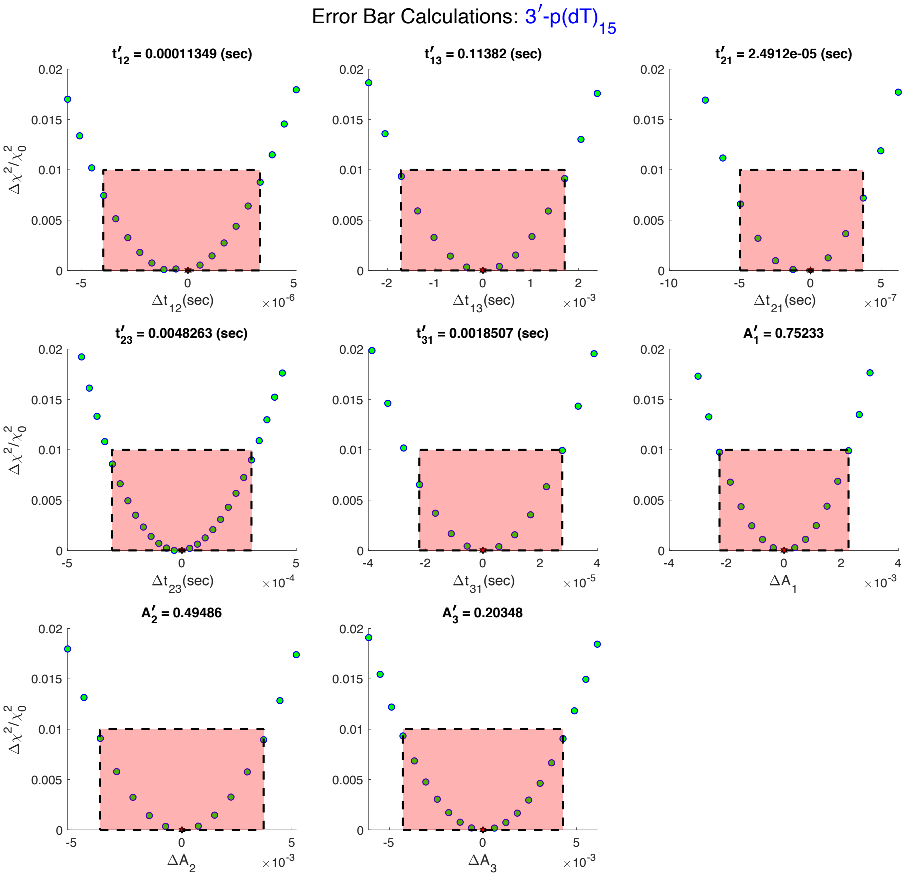


Figure 4.15 Error Analysis

In order to calculate the uncertainty of each parameter in the model, we vary the value of the parameter by small amounts until the fitness function Eq. (2.20) changes in a 1% change. The x-axis is the deviation from the optimized value, and the y-axis is the % change in χ^2.

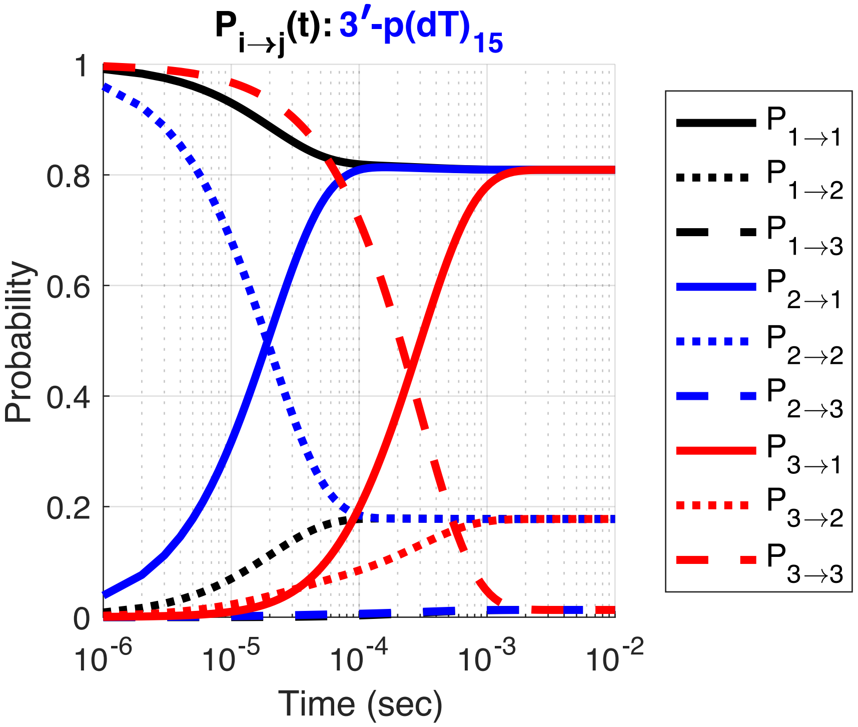


Figure 4.16 Conditional Probabilities: Solutions of the Transport Master Equation
The y-axis is the probability that the transition occurs, and the x-axis is the time in seconds. The domain is shown logarithmically, so the probability at time $t = 0$ cannot be shown on this plot. The three conformational states (compact, intermediate, and extended) are numbered states 1-3, in that order. The color of the line indicates the final state (solid = compact, dotted = intermediate, dashed = extended).

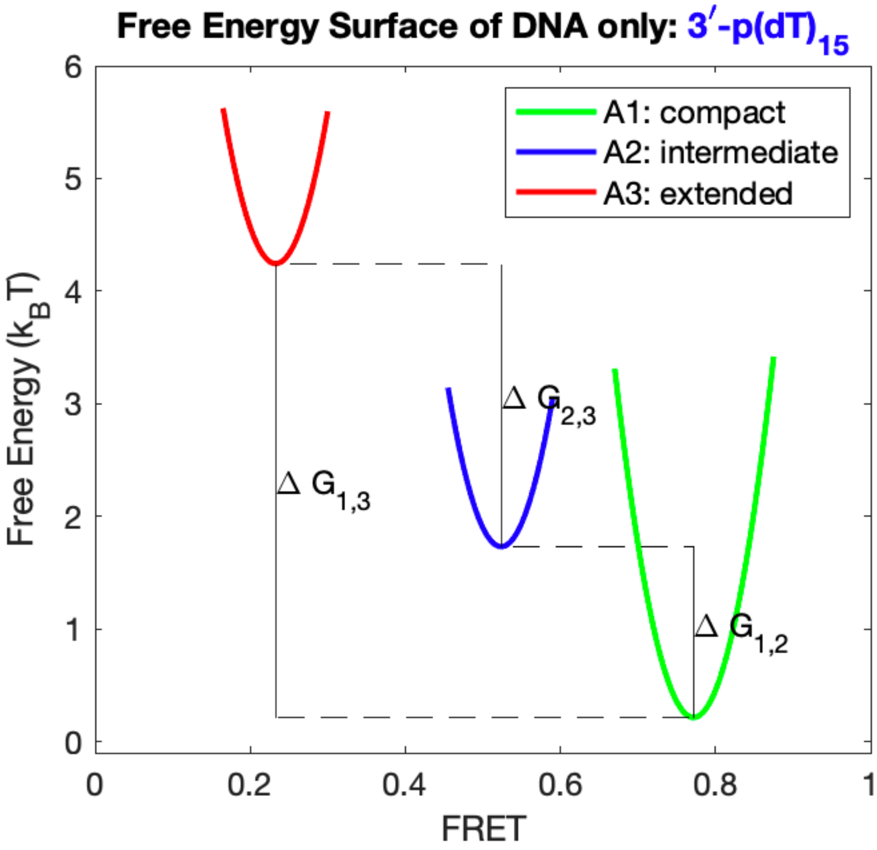


Figure 4.17 ssDNA Free Energy Surface

A parabolic free energy surface is shown corresponding to the 3`-p(dT)$_{15}$ construct results. The three macrostates are centered over their expected FRET values. The minimum of free energy for each macrostate was calculated using Eq. (4.21).

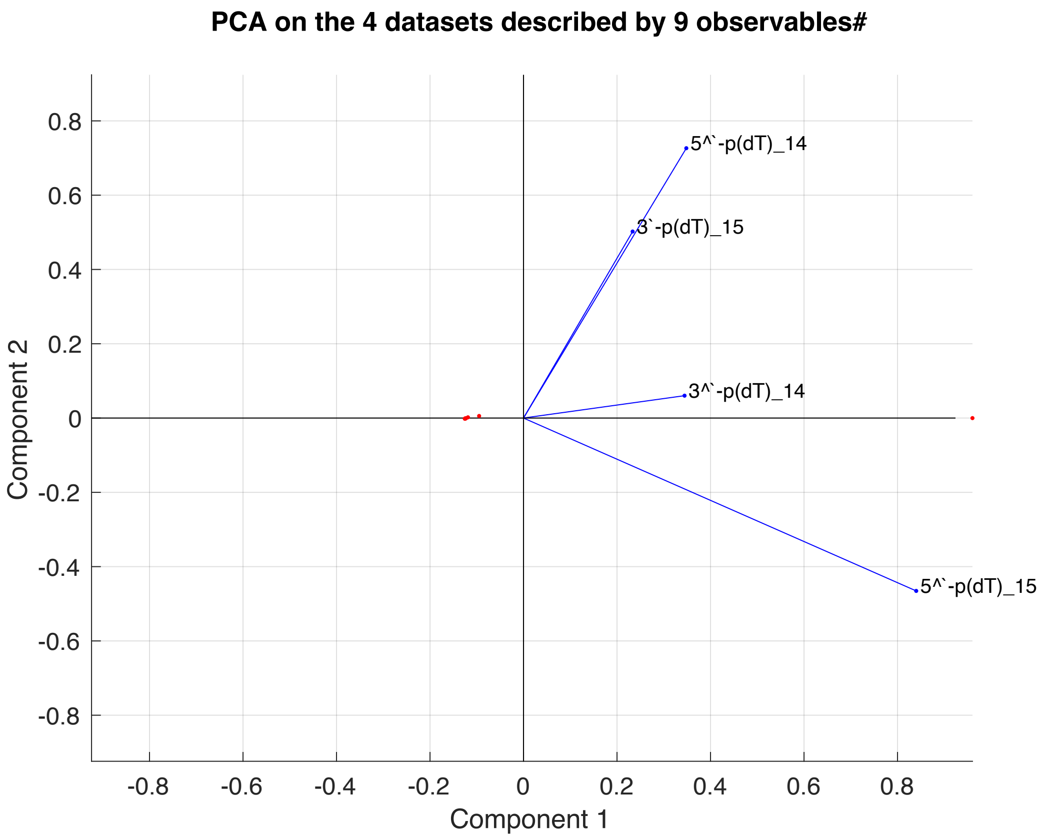


Figure 4.18 Principal Component Analysis
A principal component analysis of the parameters, including FRET values and rate constants, revealed that there is no clear pattern differentiating the constructs from one another as a function of length or polarity. The first principal component explains 99.9925 percent of the variability between data sets.

Figures for Chapter V

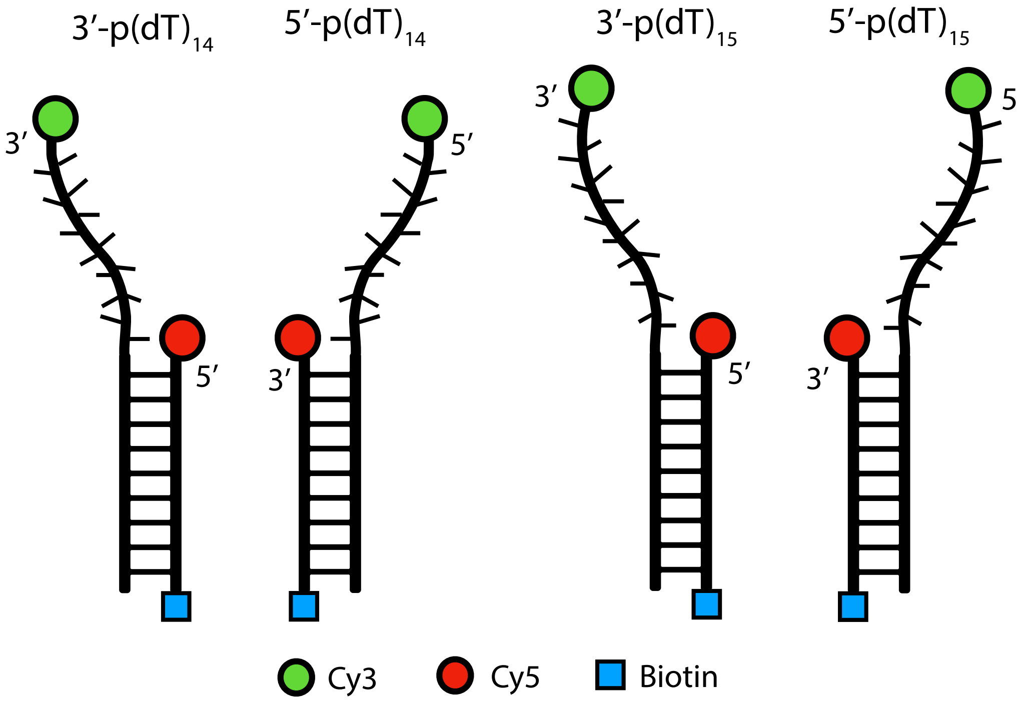


Figure 5.1: Primer/Template DNA Constructs
The four ssDNA-dsDNA junction constructs differ in length and polarity. Prior to the fluorescent label, two of the constructs terminate in a 5`-phosphate group, and the other two terminate in a 3`-hydroxyl group. For each polarity, one of the constructs consist of a shorter sequence of 14-thymine nucleotides and a longer sequence of 15-nucleotides. The acceptor-containing strand is labeled with a biotin molecule to attach to the microscope slide.

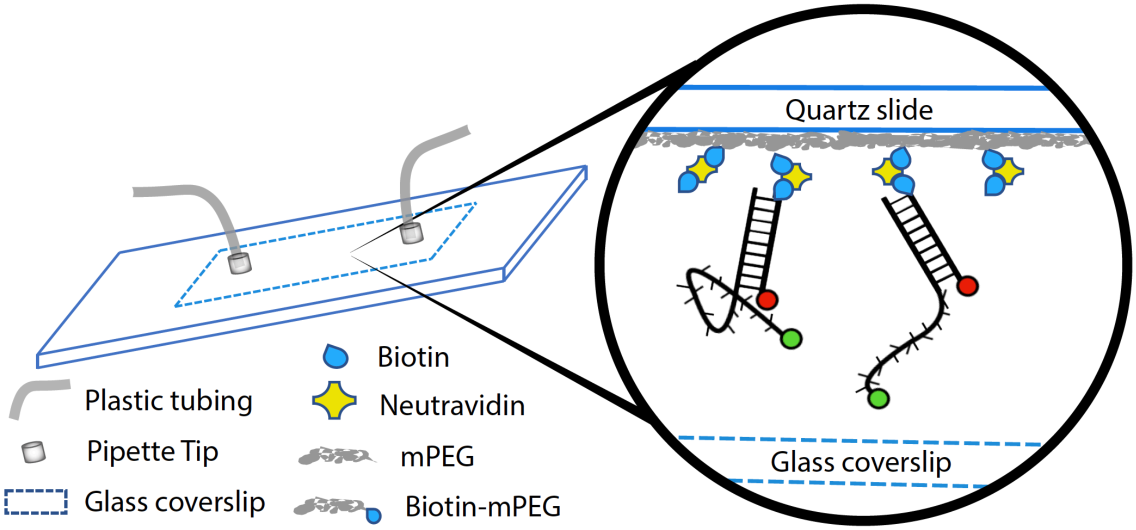


Figure 5.2. Microfluidic Sample Chamber
The ~50 µL microfluidic chamber with two access ports for adding and flushing solution through the imaging chamber. (b) Quartz slides are coated with a biotin-mPEG layer using amino-silane chemistry. The Neutravidin is a tetramer which binds both to the biotin-mPEG on the slide as well as the biotin-labeled DNA constructs. The evanescent field created by total internal reflection at the quartz-water interface will excite the fluorescent molecules but not penetrate through the glass coverslip.

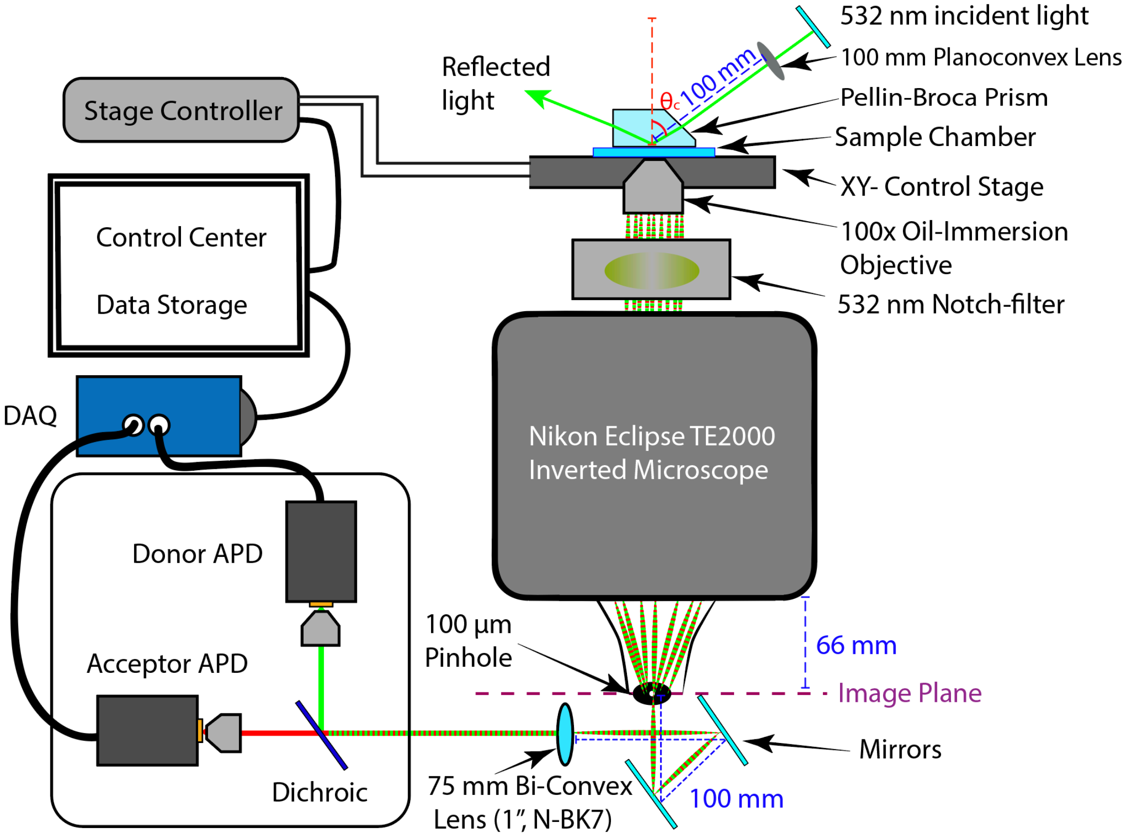


Figure 5.3 Single-Molecule Instrumentation
The microsecond resolved smFRET optical set up can be thought of as four basic parts: 1) the excitation source, 2) the interaction at the sample, 3) the redirection of fluorescence from the molecules to the detectors, and 4) the transfer of a stream of photons into an electronic representation of the signal versus time. Each part has several underlying components shown in this figure.

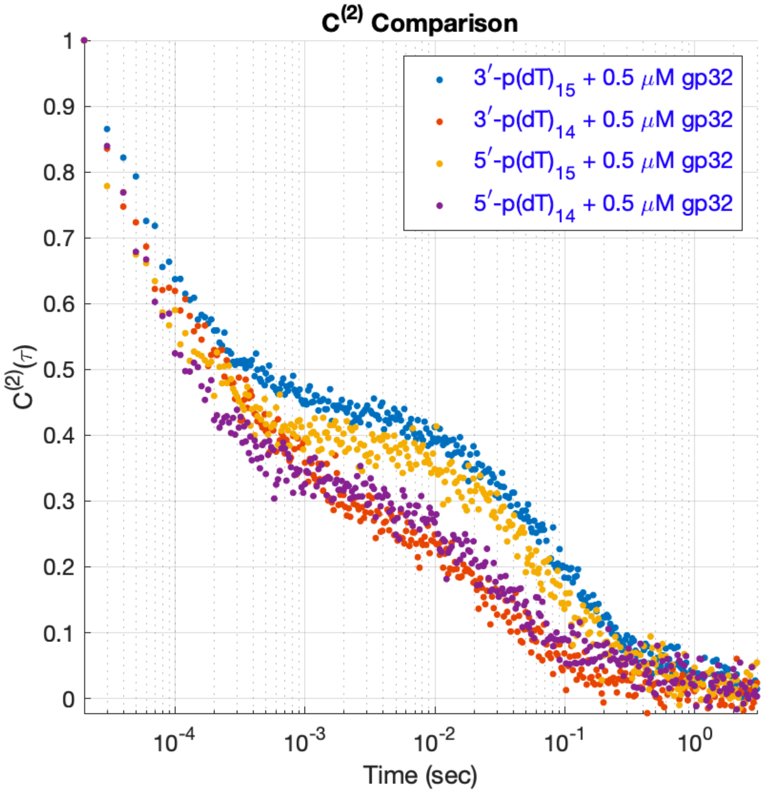
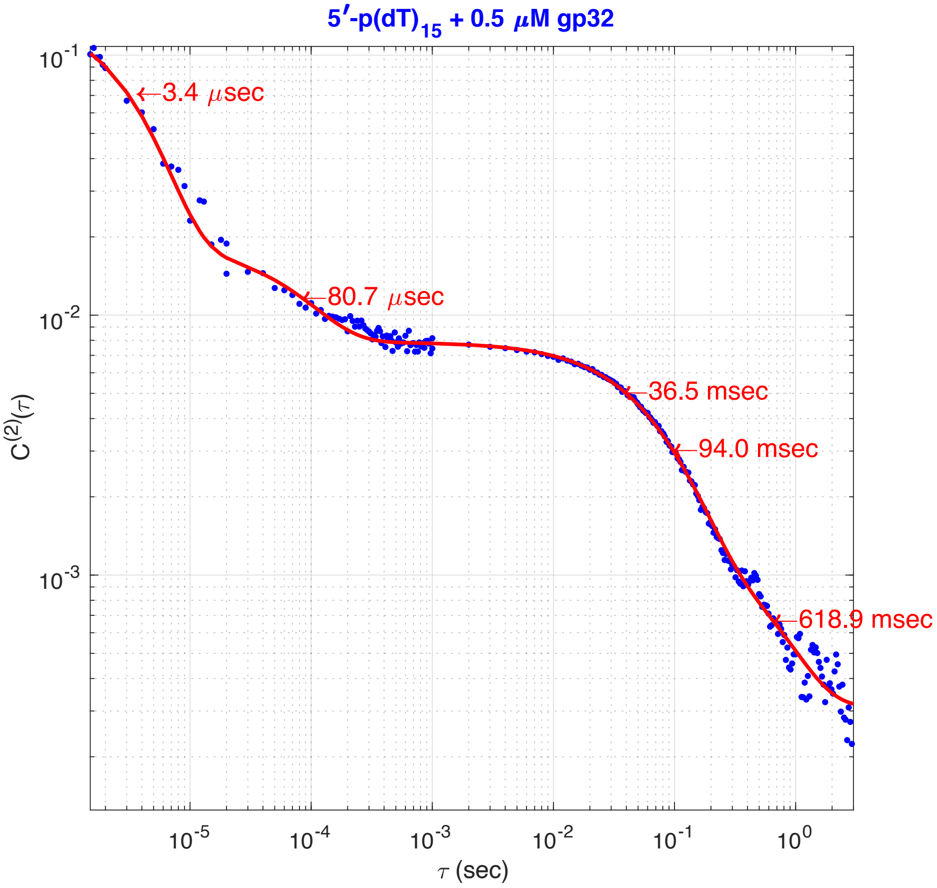


Figure 5.4 Experimental 2nd Order Time Correlation Functions
The time correlation functions are constructed from experimental single molecule FRET data. The y-axis is the normalized 2nd-order TCF, and the x-axis is the time between measurements. The plot is calculated using equation 5.10. Each point is the average of hundreds of thousands of measurements.



.

Figure 5.5 Exponential fits to 2nd Oder Time Correlation Function

An exponential fit to the $C^{(2)}$ revealed that there were processes that effected the conformational fluctuations over several orders of magnitude in time. The red line is the 2nd order time correlation function fit to a sum of five-exponential functions with inverse rate constants indicated also in red on the figure. An example fit is shown for the 5'-p(dT)$_{15}$-p/t DNA construct.

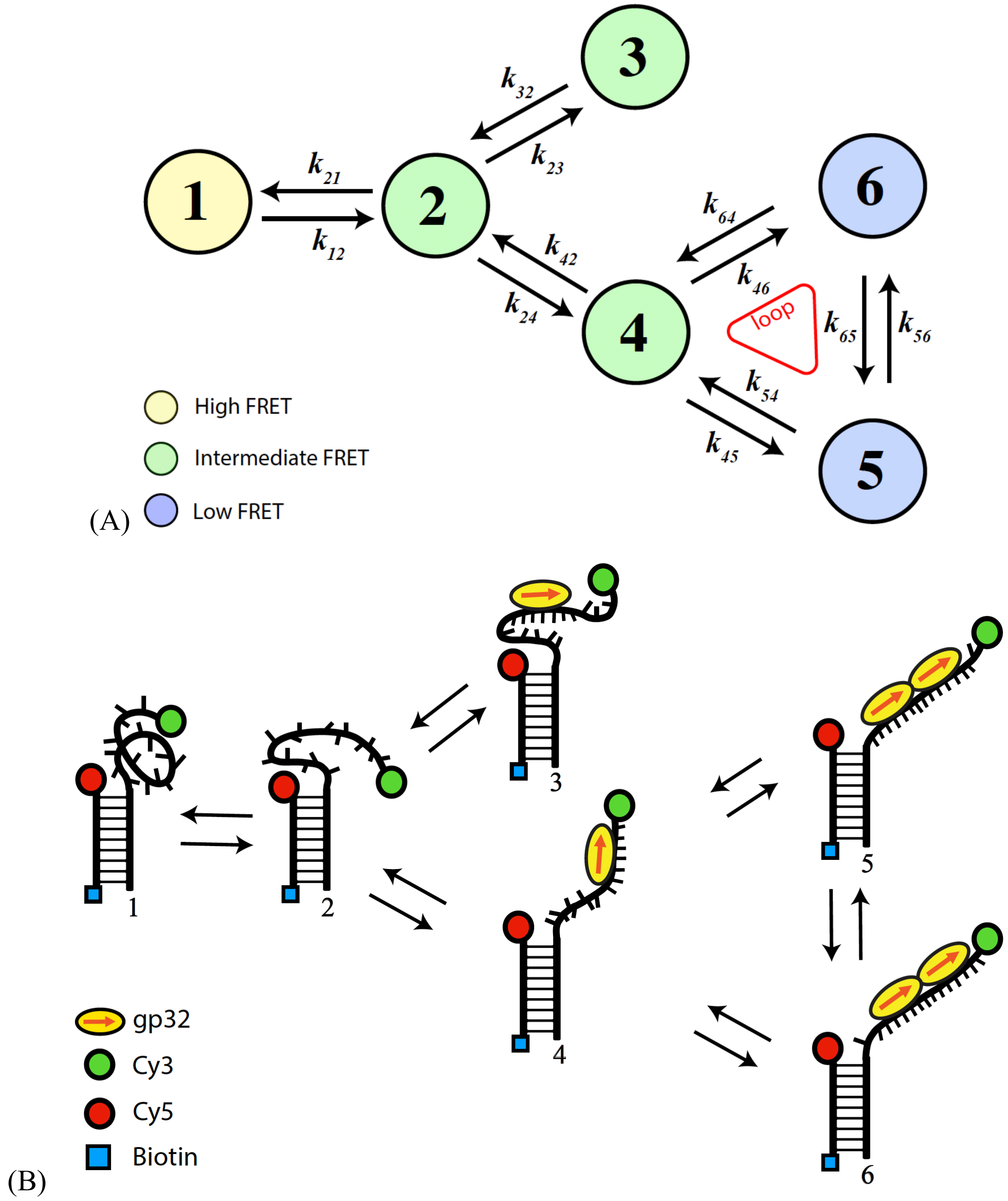


Figure 5.6 Five- and Six-State Models of ssb gp32 Dimer Assembly
All six states are shown. (A) A network schematic showing rates k_{ij} between states including a symbol indicating that there is a loop relationship between states-4, -5, and -6 which fixes one of the rates relative to the others. State-1 and state-2 are the DNA only conformations. States-3 and -4 are the 1:1 DNA:gp32 associations (monomer binding), and states-5 and -6 are the 1:2 DNA:gp32 interactions (dimer binding). (B) Possible microstates within each macrostates.

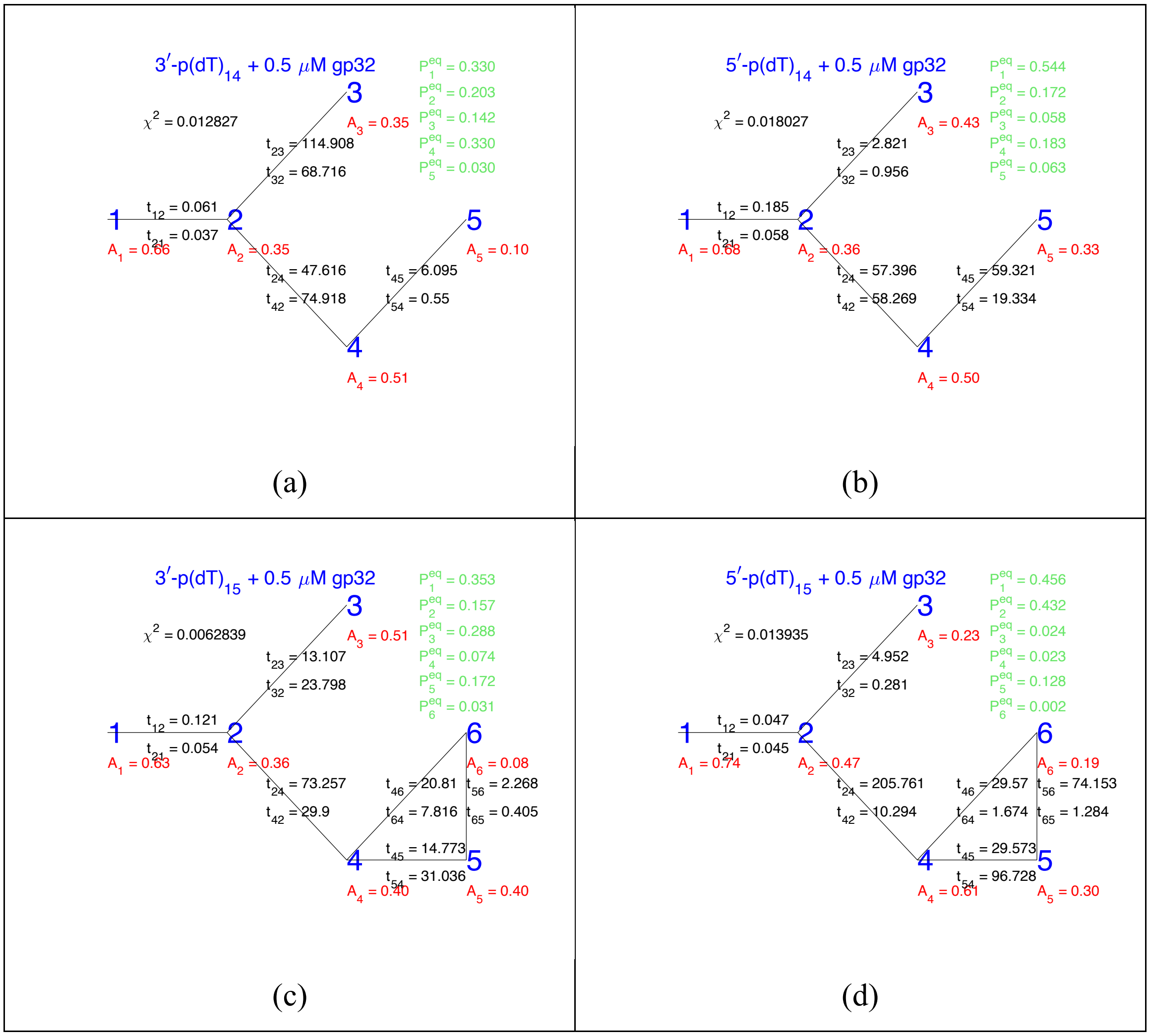


Figure 5.7 Optimized Markov Models for each Construct

Proposed mechanism of gp32 dimer assembly onto a DNA p/t junction for each of the four constructs. Each model produced the best cumulative fit to the FRET probability distribution, and the 2nd- and 4th order TCFs of the microsecond resolved FRET trajectories. Each state (blue) has a FRET value (red) is reported underneath. All inverse rates (black) are given in ms. Panel (a) features the 3'- p(dT)$_{14}$-p/t DNA construct, (b) the 5'- p(dT)$_{14}$-p/t DNA construct, (c) the 3'- p(dT)$_{15}$-p/t DNA construct, and (d) the 5'- p(dT)$_{15}$-p/t DNA construct.

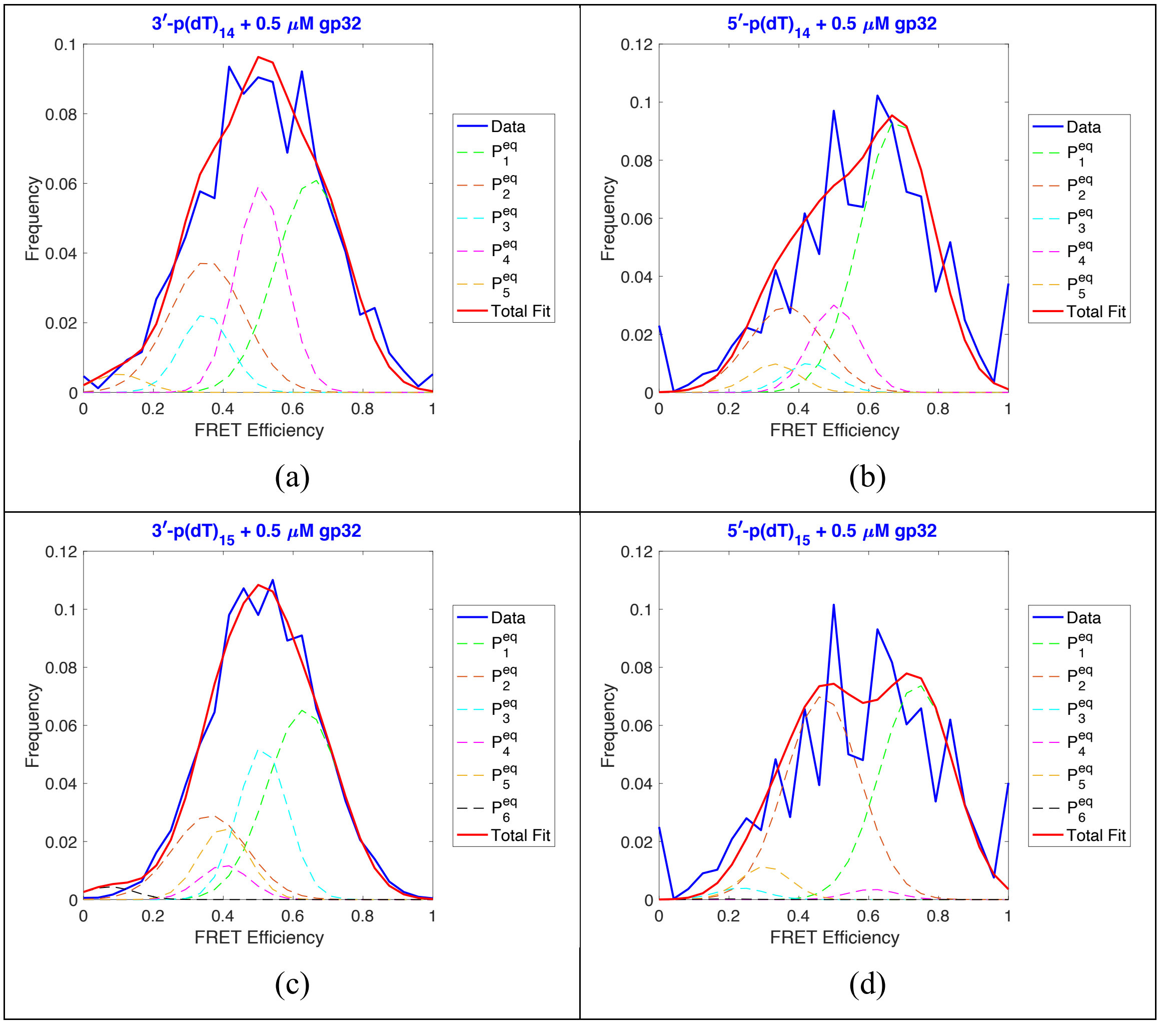
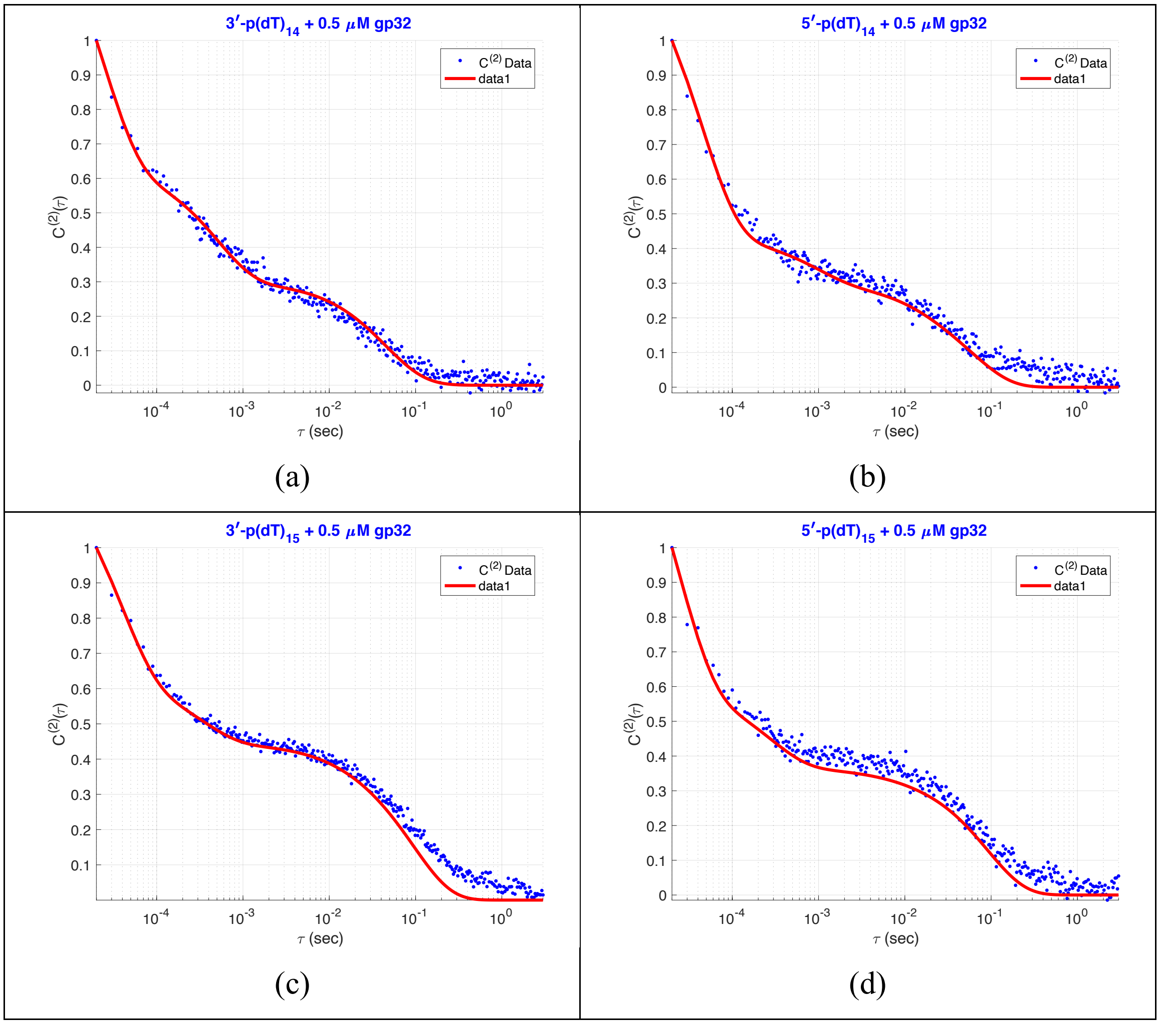


Figure 5.8 Optimized Fits to the FRET Probability Distributions

The equilibrium probability distribution is calculated using Eq. (5.3). (a) features the 3'-p(dT)$_{14}$-p/t DNA construct, (b) the 5'- p(dT)$_{14}$-p/t DNA construct, (c) the 3'- p(dT)$_{15}$-p/t DNA construct, and (d) the 5'- p(dT)$_{15}$-p/t DNA construct.



(a)

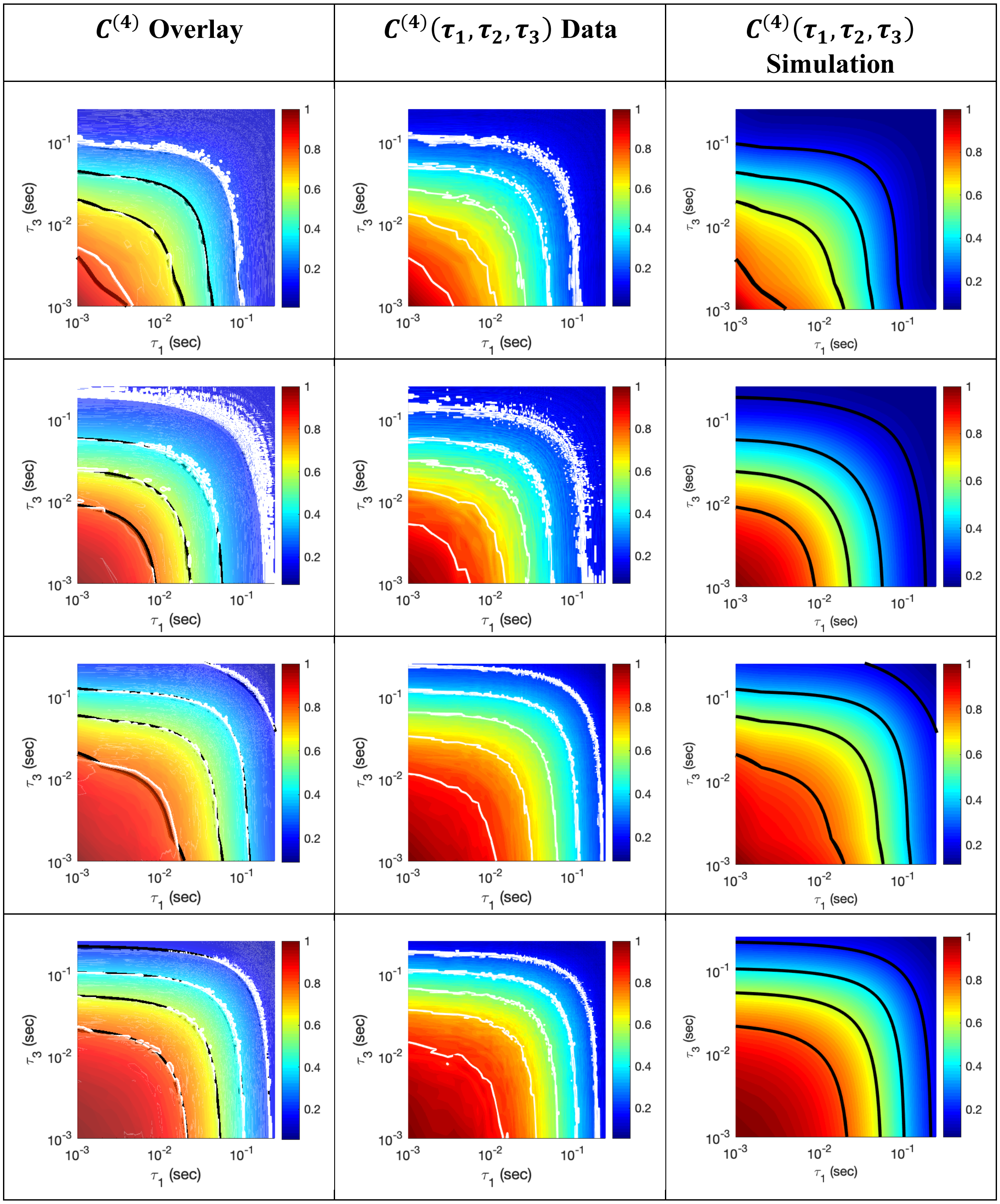
(b)

(c)

(d)

Figure 5.9 Optimized Fits to the 2nd Order Time Correlation Functions

The red line is the simulated 2nd order TCF calculated using Eq. (5.8) with the values from Table 5.2. Panel (a) features the 3'- p(dT)$_{14}$-p/t DNA construct, (b) the 5'- p(dT)$_{14}$-p/t DNA construct, (c) the 3'- p(dT)$_{15}$-p/t DNA construct, and (d) the 5'- p(dT)$_{14}$-p/t DNA construct.



$C^{(4)}$ Overlay

$C^{(4)}(\tau_1, \tau_2, \tau_3)$ Data

$C^{(4)}(\tau_1, \tau_2, \tau_3)$ Simulation

Figure 5.10 Optimized Fits to 4th Oder Time Correlation Functions

Rows (a-d) feature the 3'- p(dT)$_{14}$-p/t DNA construct, the 5'- p(dT)$_{14}$-p/t DNA construct, the 3'- p(dT)$_{15}$-p/t DNA construct, and the 5'- p(dT)$_{14}$-p/t DNA construct. Column 1 is the overlay of the simulated $C^{(4)}(\tau_1, \tau_2, \tau_3)$ plotted on top of the $C^{(4)}$ constructed from measured data. Column-3 is the simulated $C^{(4)}$.

Figures for Chapter V Supplemental Material

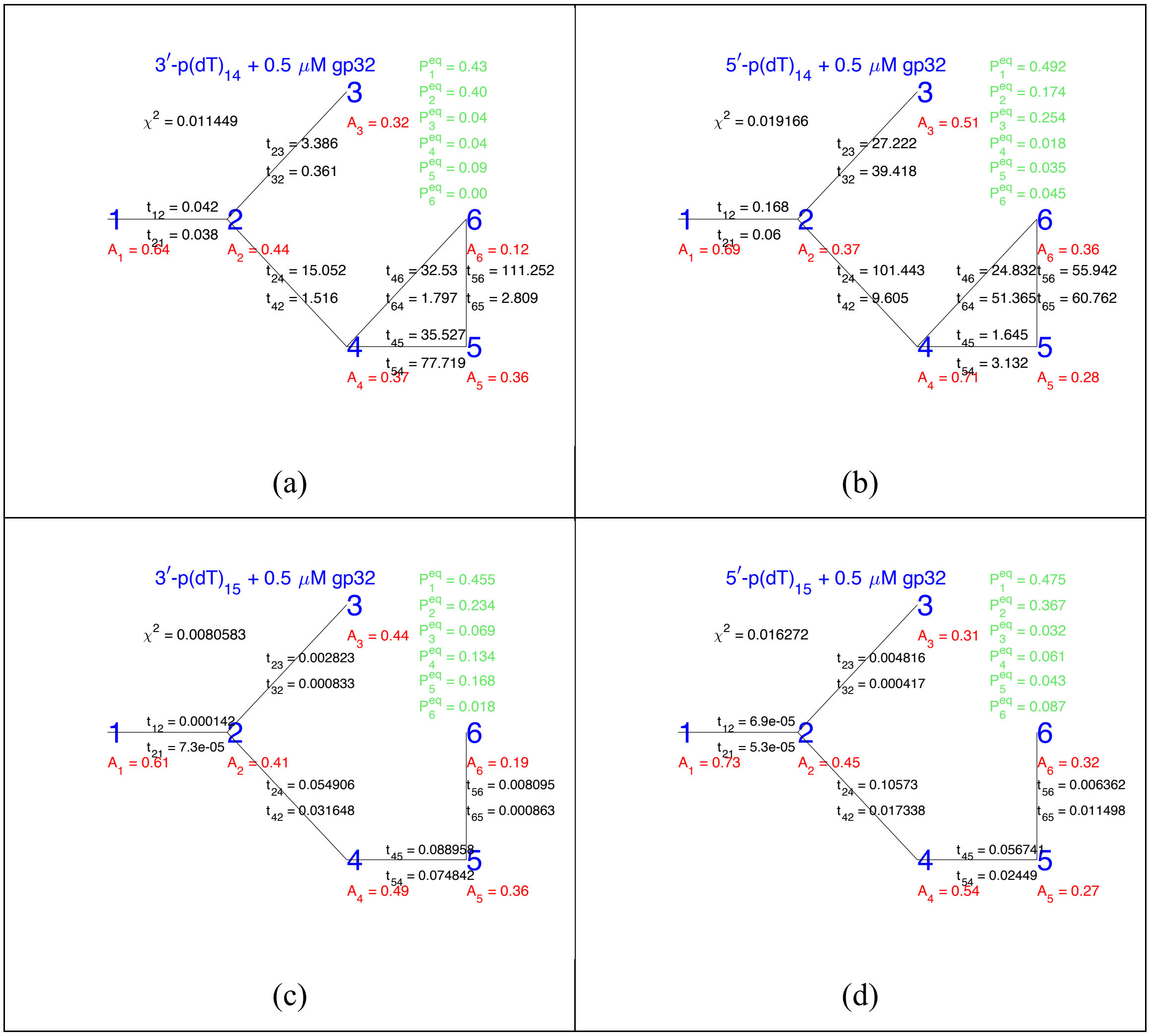


Figure S5.1 Second Best Fits to Each DNA Construct
As shown the 14-nucleotide constructs had poor agreement with the 6-state loop models, and the 15-nucleotide constructs had 2nd best agreement with the 6-state linear models. (a) features the 3'- p(dT)14-p/t DNA construct, (b) the 5'- p(dT)14-p/t DNA construct, (c) contains the 3'- p(dT)15-p/t DNA construct, and (d) the 5'- p(dT)15-p/t DNA construct.

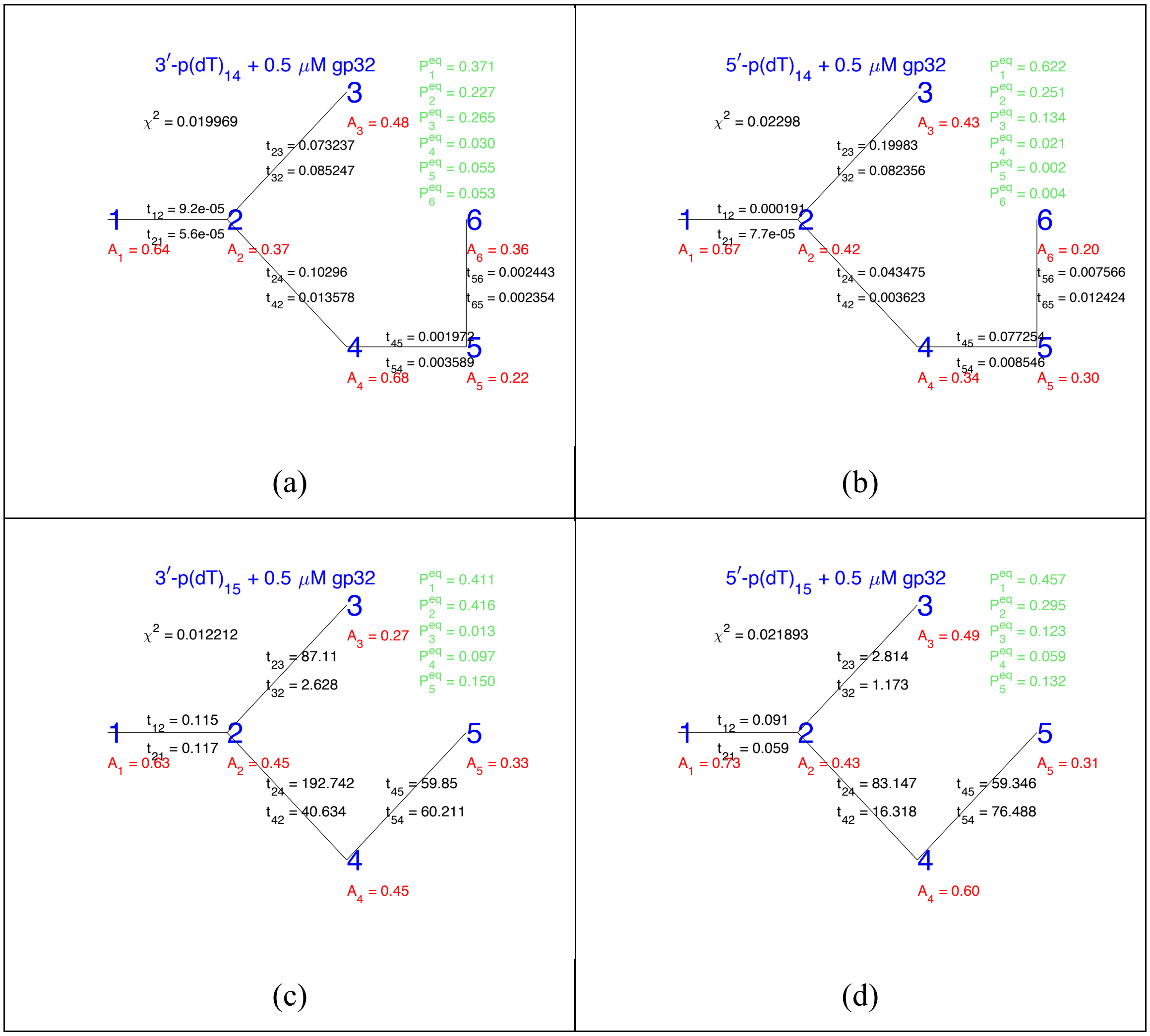


Figure S5.2 Third Best Fits to Each DNA Construct

As shown, the worst of the best fits for the 15-nucleotide constructs were the 5-state models. The 6-state linear model was the worst of the best model for the 14-nucleotide constructs. (a) the 3'- p(dT)$_{14}$-p/t DNA construct, (b) the 5'- p(dT)$_{14}$-p/t DNA construct, (c) the 3'- p(dT)$_{15}$-p/t DNA construct, and (d) the 5'- p(dT)$_{15}$-p/t DNA construct.

APPENDIX B

TABLES

Tables for Chapter III

Table 3.1. Optimized inverse rate constants for the N = 4 scheme depicted in Fig. 3.1B for the 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA under various solution conditions.
All samples contained 100 mM NaCl, 6 mM MgCl$_2$, and 10 mM Tris at pH 8.0. Fits are visualized in Fig. 3.7 of the main text, and Figs. S3.8 – S3.13 of the SI section. All inverse rate constants are given in milliseconds. Error bars are based on a 1% deviation from the optimized target function, as described in the SI section (see Fig. S3.14).

Sample condition	k_{01}^{-1}	k_{10}^{-1}	$k_{01'}^{-1}$	$k_{1'0}^{-1}$	$k_{11'}^{-1}$	$k_{1'1}^{-1}$	$k_{1'2}^{-1}$	$k_{21'}^{-1}$	$E_1 = E_{1'}$
0.1 µM gp32	264 ± 13	19.9 ± 1.0	459 ± 3.0	66.1 ± 0.3	> 5,500	> 5,500	19.5 ± 0.2	36.7 ± 0.2	0.67 ± 0.01
1.0 µM gp32	30.3 ± 1.5	26.8 ± 1.4	47.7 ± 0.4	111 ± 1.0	> 1,000	> 1,000	16.7 ± 0.6	36.0 ± 0.5	0.68 ± 0.01
0.1 µM gp32 + 1.0 µM LAST	160 ± 7.0	14.0 ± 0.7	1,200 ± 14	92.3 ± 1.0	> 1,000	> 1,000	26.6 ± 0.4	23.2 ± 0.3	0.69 ± 0.01

Tables for Chapter III Supporting Information

Table S3.1. Nucleotide base sequence and nomenclature for the p/t DNA construct used in these studies.

DNA construct	Nucleotide base sequence
3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA	5'-GTCGCCAGCCTCGCAGCCTTTTTTTTTTTTTTT/**Cy3**/-3 3'-/**biotin**/CAGCGGTCGGAGCGTCGG-**Cy5**/-5'

Table S3.2. Time constants and amplitudes from bi-exponential fits to experimental 2nd-order TCFs for 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA under various solution conditions. All samples contained 100 mM NaCl, 6 mM MgCl$_2$, and 10 mM Tris at pH 8.0. Fits are visualized in Fig. 3.3. Time constants are accurate to within ± 0.1 msec.

Sample condition	τ_{fast} (msec)	τ_{slow} (msec)	A_{fast}	A_{slow}	R^2
0.1 μM gp32	18.4	158	0.146	0.854	0.999
1.0 μM gp32	13.9	94.2	0.531	0.469	0.998
μM gp32 + 1.0 μM LAST	24.3	184	0.646	0.354	0.995

Table S3.3. Optimized inverse rate constants for the N = 3 scheme depicted in Fig. 3.1A for the 3'-Cy3/Cy5-p(dT)15-p/t DNA under various solution conditions. All samples contained 100 mM NaCl, 6 mM MgCl$_2$, and 10 mM Tris at pH 8.0. All inverse rate constants are given in milliseconds.

Sample condition	k_{01}^{-1}	k_{10}^{-1}	k_{12}^{-1}	k_{21}^{-1}	k_{02}^{-1}	k_{20}^{-1}	E_1
0.1 μM gp32	479	49.1	80.4	95.6	10,000	10,000	0.56
1.0 μM gp32	635	10,000	22.4	206	97.6	1,414	0.80
0.1 μM gp32 + 1.0 μM LAST	10,000	267	254	85.9	1,500	13.5	0.59

Table S3.4. Minimized target function values corresponding to the N = 3 and N = 4 fit results reported in Table S3.3 and Table 3.1 (main text), respectively. G_2 is the first term in Eq. (S3.9), G_4 is the second term, G_{eq} is the third term, and G_{total} is the sum value. The fits are uniformly better for the $N = 4$ model than for the $N = 3$ model.

Sample condition	Scheme	G_2	G_4	G_{eq}	G_{total}
0.1 μM gp32					
	$N = 3$	6.1	71.1	4.1	81.3
	$N = 4$	0.3	18.3	0.1	18.7
1.0 μM gp32					
	$N = 3$	49	328	30	407
	$N = 4$	1.5	125	7.8	134

Tables for Chapter IV

Table 4.1. Nucleotide Base Sequence for the p/t DNA Constructs

DNA construct	Nucleotide base sequence
3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA	5'-GTCGCCAGCCTCGCAGCCTTTTTTTTTTTTTTT/**Cy3**/-3’ 3’-/**biotin**/CAGCGGTCGGAGCGTCGG-**Cy5/**-5’
3'-Cy3/Cy5-p(dT)$_{14}$-p/t DNA	5'-GTCGCCAGCCTCGCAGCCTTTTTTTTTTTTTT/**Cy3**/-3’ 3’-/**biotin**/CAGCGGTCGGAGCGTCGG-**Cy5/**-5’
5'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA	3'-GTCGCCAGCCTCGCAGCCTTTTTTTTTTTTTTT/**Cy3**/-5’ 5’-/**biotin**/CAGCGGTCGGAGCGTCGG-**Cy5/**-3’
5'-Cy3/Cy5-p(dT)$_{14}$-p/t DNA	3'-GTCGCCAGCCTCGCAGCCTTTTTTTTTTTTTT/**Cy3**/-5’ 5’-/**biotin**/CAGCGGTCGGAGCGTCGG-**Cy5/**-3’

Table 4.2 Time Constants of Photophysical Decays in 2nd Order TCF
The first 20 μs of the autocorrelation function for each construct exhibited a single exponential decay all around 3 μs which we interpret to be a result of photophysical processes of the chromophores. The blue dots are the data points calculated with Eq. 4.9, and the fit is calculated with $\bar{C}^{(2)}(\tau) = \sum_i A_i e^{-\lambda_i \tau_1}$.

Construct	t_1 (μs)
$3'\mathrm{p(dT)}_{15}$	2.9 ± 0.45
$3'\mathrm{p(dT)}_{14}$	2.7 ± 0.35
$5'\mathrm{p(dT)}_{15}$	3.2 ± 0.51
$5'\mathrm{p(dT)}_{14}$	3.0 ± 0.50

Table 4.3 Optimized Inverse Rate Constants and FRET values
The first four rows are the values of the best-fit parameters from the four constructs using the cyclical model. The values in the last two rows are the mean and standard deviation of the values across all four constructs. The values displayed in columns 2-7 are the inverse rate constants t_{ij} = k_{ij}^{-1}, i.e. the time to go from state-i to state-j. The values in the final three columns, E_i, are the FRET value associated with state-i.

DNA Sample	$\boldsymbol{t_{12}}$ (ms)	$\boldsymbol{t_{21}}$ (ms)	$\boldsymbol{t_{23}}$ (ms)	$\boldsymbol{t_{32}}$ (ms)	$\boldsymbol{t_{13}}$ (ms)	$\boldsymbol{t_{31}}$ (ms)	$\boldsymbol{E_1}$	$\boldsymbol{E_2}$	$\boldsymbol{E_3}$
$3'p(dT)_{15}$	0.113	0.025	4.826	0.358	113.816	1.851	0.752	0.495	0.203
$3'p(dT)_{14}$	0.264	0.020	4.973	0.609	167.516	1.552	0.718	0.312	0.199
$5'p(dT)_{15}$	0.154	0.020	6.807	0.289	326.802	1.798	0.696	0.377	0.088
$5'p(dT)_{14}$	0.097	0.027	7.231	0.310	169.599	2.000	0.764	0.490	0.246
Mean	0.157	0.023	5.959	0.391	194.433	1.800	0.733	0.418	0.184
σ	0.075	0.003	1.237	0.148	91.945	0.186	0.031	0.090	0.067

Tables for Chapter V

Table 5.1. Nucleotide Base Sequence of The Four DNA Constructs

DNA construct	Nucleotide base sequence
3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA	5'-GTCGCCAGCCTCGCAGCCTTTTTTTTTTTTTTT/**Cy3**/-3 3’-/**biotin**/CAGCGGTCGGAGCGTCGG-**Cy5/**-5’
3'-Cy3/Cy5-p(dT)$_{14}$-p/t DNA	5'-GTCGCCAGCCTCGCAGCCTTTTTTTTTTTTTT/**Cy3**/-3’ 3’-/**biotin**/CAGCGGTCGGAGCGTCGG-**Cy5/**-5’
5'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA	3'-GTCGCCAGCCTCGCAGCCTTTTTTTTTTTTTTT/**Cy3**/-5 5’-/**biotin**/CAGCGGTCGGAGCGTCGG-**Cy5/**-3’
5'-Cy3/Cy5-p(dT)$_{14}$-p/t DNA	3'-GTCGCCAGCCTCGCAGCCTTTTTTTTTTTTTT/**Cy3**/-5’ 5’-/**biotin**/CAGCGGTCGGAGCGTCGG-**Cy5/**-3’

Table 5.2 Optimized inverse rate constants for the N = 5 and N= 6 scheme depicted in Fig. 5.4 and Fig. 5.5 for the 4 p/t DNA constructs.
All samples contained 100 mM NaCl, 6 mM $MgCl_2$, and 10 mM Tris at pH 8.0. Fits are visualized in Figure 5.6 – 5.10. All inverse rate constants are given in milliseconds.

Sample Condition	3'-p(dT)$_{15}$-p/t DNA + 0.5 µM gp32	5'-p(dT)$_{15}$-p/t DNA + 0.5 µM gp32	3'-p(dT)$_{14}$-p/t DNA + 0.5 µM gp32	5'-p(dT)$_{14}$-p/t DNA + 0.5 µM gp32
τ_{12}	0.121	0.047	0.061	0.185
τ_{21}	0.054	0.045	0.037	0.058
τ_{23}	13.107	4.952	114.908	2.821
τ_{24}	73.257	205.761	47.616	57.396
τ_{32}	23.798	0.281	68.716	0.956
τ_{42}	29.900	10.294	74.918	58.269
τ_{45}	14.773	29.573	6.095	59.321
τ_{46}	20.810	29.570	—	—
τ_{54}	31.036	96.728	0.550	19.334
τ_{56}	2.268	74.153	—	—
τ_{64}	7.816	1.674	—	—
τ_{65}	0.405	1.284	—	—
A1	0.632	0.739	0.657	0.682
A2	0.362	0.469	0.353	0.364
A3	0.513	0.234	0.349	0.430
A4	0.403	0.610	0.507	0.503
A5	0.402	0.303	0.101	0.330
A6	0.075	0.192	—	—

APPENDIX C

Supplemental Material

Chapter II

Appendix 2.1 Analytical Expression for $N = 3$ System

The general solution to Eq. (13) is

$$\boldsymbol{p}(t) = \boldsymbol{p}^{eq} + c_1 \boldsymbol{v}_1 e^{-\lambda_1 t} + c_2 \boldsymbol{v}_2 e^{-\lambda_2 t}, \qquad N = 3 \tag{A2.1}$$

where the eigenvalues are given by $\lambda_1 = a + b$ and $\lambda_2 = a - b$ with $a = \frac{1}{2}(k_{01} + k_{10} + k_{12} + k_{21} + k_{02} + k_{20})$ and $b = \frac{1}{2}[(k_{01} - k_{12})^2 + (k_{02} - k_{21})^2 + (k_{10} - k_{20})^2 + 2k_{01}(k_{02} + k_{10} - k_{20} - k_{21}) + 2k_{02}(k_{20} - k_{10} - k_{12}) + 2k_{12}(k_{10} - k_{20} + k_{21}) - 2k_{10}k_{21} + 2k_{20}k_{21}]^{\frac{1}{2}}$, and the eigenvectors are given by $\boldsymbol{v}_1 = [v_1^0, v_1^1, v_1^2]$ and $\boldsymbol{v}_2 = [v_2^0, v_2^1, v_2^2]$ with $v_1^0 = (k_{12} + k_{20} + k_{21} - a - b)/(k_{02} - k_{12})$, $v_1^1 = (a + b - k_{02} - k_{20} - k_{21})/(k_{02} - k_{12})$, $v_2^0 = (k_{12} + k_{20} + k_{21} - a + b)/(k_{02} - k_{12})$, $v_2^1 = (a - b - k_{02} - k_{20} - k_{21})/(k_{02} - k_{12})$, and $v_1^2 = v_2^2 = 1$.

To satisfy detailed balance, one rate constant must depend on the others, such that $k_{20} = k_{02}k_{21}k_{10}/k_{12}k_{01}$. The equilibrium populations $\boldsymbol{p}^{eq} = \left[p_0^{eq}, p_1^{eq}, p_2^{eq}\right]$ are found by solving Eq. (13) with the boundary condition $\dot{\boldsymbol{p}}(t) = 0$. These solutions must also satisfy completeness: $\sum_{i=0}^{2} p_i\,(t) = 1$. This gives $p_0^{eq} = \{1 + [(k_{01} + k_{02})/k_{10}] + [(1 - k_{20})/k_{10}] \cdot (k_{10} + k_{12})(k_{01} + k_{02}) - k_{01}k_{10}/[(k_{10} + k_{12})k_{20} + k_{21}k_{10}]\}^{-1}$, $p_2^{eq} = p_0^{eq} \cdot [(k_{10} + k_{12})(k_{01} + k_{02}) - k_{01}k_{10}] \cdot [(k_{10} + k_{12})k_{20} + k_{21}k_{10}]^{-1}$, and $p_1^{eq} = k_{10}^{-1} \cdot \left[p_0^{eq}(k_{01} + k_{02}) - p_2^{eq}k_{20}\right]$.

To determine the nine conditional probabilities $p_{ji}(\tau)$ with $i, j \in \{0,1,2\}$, we solve Eq. (A2.1) for the expansion coefficients c_1 and c_2, while assuming the appropriate boundary conditions. We label each expansion coefficient with a superscript to indicate the boundary condition. For example, the expansion coefficient c_1^0 corresponds to the case when all population resides in state-0 at time zero, i.e. $p_0(0) = 1$ and $p_1(0) = p_2(0) = 0$. This leads to the following expressions for the expansion coefficients: $c_2^0 = (v_1^0 v_2^1 - v_1^1 v_2^0)^{-1}\left[v_1^1 p_0^{eq} - v_1^0 p_1^{eq} - v_1^1\right]$, $c_2^1 = (v_1^0 v_2^1 - v_1^1 v_2^0)^{-1}\left[v_1^0 - v_1^0 p_1^{eq} + v_1^1 p_0^{eq}\right]$, $c_2^2 =$

$(v_1^0 v_2^1 - v_1^1 v_2^0)^{-1}[v_1^1 p_0^{eq} - v_1^0 p_1^{eq}]$, $c_1^0 = (v_1^0)^{-1}[1 - p_0^{eq} - c_2^0 v_2^0]$, $c_1^1 = (v_1^0)^{-1}[-p_0^{eq} - c_2^1 v_2^0]$, and $c_1^2 = (v_1^0)^{-1}[-p_0^{eq} - c_2^2 v_2^0]$. Upon substitution of these into Eq (A2.1), we obtain the conditional probabilities described by Eq. (15) of the text.

Appendix 2.2 Analytical Description of Time-Dependent Transition Density Plots (TDPs)

Consider an *N*-state Markov system at equilibrium for which stochastic transitions may occur from state-i to state-j. At any instant in time, the probability to observe the system in state-i is given by the rate expression

$$\dot{p}_i = -k_{ij} p_i, \tag{A2.2}$$

which decays according to the general solution

$$p_i(t) = A e^{-k_{ij} t}, \tag{A2.3}$$

where A is an integration constant. The elements of the time-dependent TDP are described by the probabilities that a transition occurs from state-i to state-j within a time interval τ. By integrating Eq. (A3) over this time interval, we obtain:

$$p_{ij}(\tau) = A \int_0^\tau e^{-k_{ij} t} dt = \frac{A}{k_{ij}}\left(1 - e^{-k_{ij}\tau}\right). \tag{A2.4}$$

In the limit of very long times ($\tau \to \infty$), we expect the transition probability $p_{ij}(\tau \to \infty)$ to depend on the equilibrium probability that the system resides in state-i, according to $p_{ij}(\tau \to \infty) = k_{ij}\, p_i^{eq}$. Taking the long-time limit of Eq. (A2.4), we obtain $p_{ij}(\tau \to \infty) = A/k_{ij}$. Solving for A, and substitution into Eq. (A2.4) gives the expression for the elements of the time-dependent TDP:

$$p_{ij}(\tau) = k_{ij}\, p_i^{eq}\left(1 - e^{-k_{ij}\tau}\right). \tag{A2.5}$$

Chapter III

I. Calculation of 2nd- and 4th-order TCFs using the Theory of Markov Chains

Both 2nd- and 4th-order TCFs can be modeled using the theory of Markov chains[1]. These expressions are given by

$$\bar{C}^{(2)}(\tau) = \sum_{i,j=0}^{N-1} \delta E_j p_{ji}(\tau) \delta E_i p_i^{eq} \tag{S3.1}$$

and

$$\begin{aligned} &\langle \delta E_{FRET}(0) \delta E_{FRET}(\tau_1) \delta E_{FRET}(\tau_2) \delta E_{FRET}(\tau_3) \rangle \\ &\quad = \sum_{i,j,k,l=0}^{N-1} \delta E_l p_{lk}(\tau_3) \delta E_k p_{kj}(\tau_2) \delta E_j p_{ji}(\tau_1) \delta E_i p_i^{eq} \end{aligned} \tag{S3.2}$$

In Eqs. (S3.1) and (S3.2), p_i^{eq} is the equilibrium (time-independent) probability to observe the system in the ith state, δE_i is the value of the fluctuation observable associated with that state, and $p_{ji}(\tau)$ is the conditional probability that the system will be in the jth state at a time interval τ after it was initially observed to be in the ith state. The conditional probabilities are the solutions to the memory-less Master equation that governs the kinetics of a network of N chemical species, which are interconnected by elementary reactions according to a well-defined scheme, such as those depicted in Figs. 3.1A and 3.1B of the main text.

$$\begin{aligned} \dot{\boldsymbol{p}}(t) = \boldsymbol{K}\boldsymbol{p}(t) \equiv \begin{bmatrix} \dot{p}_0 \\ \dot{p}_1 \\ \vdots \\ \dot{p}_{N-1} \end{bmatrix} \\ = \begin{bmatrix} -\sum_{i=0}^{N-1} k_{0,i} & k_{1,0} & \cdots & k_{N-1,0} \\ k_{0,1} & -\sum_{i=0}^{N-1} k_{1,i} & \ddots & \vdots \\ \vdots & \ddots & \ddots & k_{N-1,N-2} \\ k_{0,N-1} & \cdots & k_{N-2,N-1} & -\sum_{i=0}^{N-1} k_{N-1,i} \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ \vdots \\ p_{N-1} \end{bmatrix} \end{aligned} \tag{S3.3}$$

In Eq. (S3.3), $\boldsymbol{p}(t)$ is an N-dimensional vector containing the probabilities to find the system in each of its N states at time t, and $\boldsymbol{K}$ is the N×N rate matrix, with elements k_{ij} associated with the transitions from state-i to state-j. The choice of the rate constants k_{ij} used in Eq. (S3.3) must satisfy continuity and detailed balance conditions, as discussed in[2].

For our calculations based on the $N = 3$ scheme depicted in Fig. 3.1A, Eq. (S3.3) has the form

$$\begin{bmatrix} \dot{p}_0 \\ \dot{p}_1 \\ \dot{p}_2 \end{bmatrix} = \begin{bmatrix} -k_{0,1} - k_{0,2} & k_{1,0} & k_{2,0} \\ k_{0,1} & -k_{1,0} - k_{1,2} & k_{2,1} \\ k_{0,2} & k_{1,2} & -k_{2,0} - k_{2,1} \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ p_2 \end{bmatrix} \tag{S3.4}$$

In Eq. (S3.4), we have allowed for the hypothetical transition between state-0 and state-2, which would require the binding of appropriately preformed gp32 dimers directly from solution. Although this does not occur in the real system, we have included the possibility in our modeling to provide generality.

For our calculations based on the $N = 4$ scheme shown in Fig. 3.1B, Eq. (S3.3) has the form

$$\begin{bmatrix} \dot{p}_0 \\ \dot{p}_1 \\ \dot{p}_{1'} \\ \dot{p}_2 \end{bmatrix} = \begin{bmatrix} -k_{0,1} - k_{0,1'} & k_{1,0} & k_{1',0} & 0 \\ k_{0,1} & -k_{1,0} - k_{1,1'} & k_{1',1} & 0 \\ k_{0,1'} & k_{1,1'} & -k_{1',0} - k_{1',1} - k_{1',2} & k_{2,1'} \\ 0 & 0 & k_{1',2} & -k_{2,1'} \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ p_{1'} \\ p_2 \end{bmatrix} \tag{S3.5}$$

We note that the structure of the rate matrix in Eq. (S3.5) satisfies continuity and detailed balance conditions as required by the connectivity of the states depicted in Fig. 3.1B.

Using Eqs. (S3.1) – (S3.2), it is straightforward to show that the 2nd- and 4th-order TCFs have functional forms given by, respectively

$$\bar{C}^{(2)}(\tau) = \mathcal{A}_1 e^{-\lambda_1 \tau} + \mathcal{A}_2 e^{-\lambda_2 \tau} + \ldots + \mathcal{A}_{N-1} e^{-\lambda_{N-1} \tau} \tag{S3.6}$$

and

$$\begin{aligned} &\bar{C}^{(4)}(\tau_1, \tau_3)]_{\tau_2\, fixed} = \\ &\mathcal{A}_{1,1}(\tau_2) e^{-\lambda_1(\tau_1+\tau_3)} + \mathcal{A}_{1,2}(\tau_2) e^{-\lambda_1\tau_1 - \lambda_2\tau_3} + \cdots + \mathcal{A}_{1,N-1}(\tau_2) e^{-\lambda_1\tau_1 - \lambda_{N-1}\tau_3} + \\ &\mathcal{A}_{2,1}(\tau_2) e^{-\lambda_2\tau_1 - \lambda_1\tau_3} + \mathcal{A}_{2,2}(\tau_2) e^{-\lambda_2(\tau_1+\tau_3)} + \cdots + \mathcal{A}_{2,N-1}(\tau_2) e^{-\lambda_2\tau_1 - \lambda_{N-1}\tau_3} + \\ &\vdots \\ &\mathcal{A}_{N-1,1}(\tau_2) e^{-\lambda_{N-1}\tau_1 - \lambda_1\tau_3} + \mathcal{A}_{N-1,2}(\tau_2) e^{-\lambda_{N-1}\tau_1 - \lambda_2\tau_3} + \cdots + \mathcal{A}_{N-1,N-1}(\tau_2) e^{-\lambda_{N-1}(\tau_1+\tau_3)} \end{aligned} \tag{S3.7}$$

As discussed in the text, the 2nd-order TCF given by Eq. (S3.6) is composed of $N - 1$ exponentially decaying terms, each with characteristic decay rates λ_1, λ_2, …, λ_{N-1}, and amplitudes $\mathcal{A}_1$, $\mathcal{A}_2$, …,$\mathcal{A}_{N-1}$. The decay rates and amplitudes are polynomial functions of the rate constants k_{ij} that connect the elementary chemical steps of the *N*-state reaction scheme.[2] The 4th-order TCF given by Eq. (S3.7) is composed of $(N-1)^2$ terms, each with an amplitude $\mathcal{A}_{n,m}$ $[n, m \in \{1,2,\dots,N-1\}]$ that depends on the waiting time τ_2. For a fixed waiting time, the decay of the 4th-order TCF occurs in two dimensions, corresponding to the time intervals τ_1 and τ_3. The characteristic decay rates of the 4th-order TCF are the same as those of the 2nd-order TCF. The *N* – 1 terms with amplitudes $\mathcal{A}_{n,n}$ are 'diagonal' terms, which each depends on a single decay constant λ_n. The terms with amplitudes $\mathcal{A}_{n,m}$ (with $n \neq m$) designate 'off-diagonal' coupling terms, which each depends on two decay constants, λ_n and λ_m. For an equilibrium system, the principles of detailed balance require that $\mathcal{A}_{n,m} = \mathcal{A}_{m,n}$[2].

II. Calculation of the 2D Rate Spectrum by Inverse Laplace Transform of the 4th-Order TCF.

To perform the ILT, we carried out the following algorithm based on the Tikhonov regularization method[3-5]: (*i*) We created a rate domain *test function* corresponding to the ILT of Eq. (S3.7), which was composed of (*N* – 1)^2 delta functions with fit parameter amplitudes $\mathcal{A}_{n,m}$ $[n, m \in \{1,2,\dots,N-1\}]$.

$$\bar{C}^{(4)}(k_1, k_3)]_{\tau_2\, fixed} = \tag{S3.8}$$
$$\begin{gathered}
\mathcal{A}_{1,1}(\tau_2)(k_1-\lambda_1)\delta(k_3-\lambda_1) + \mathcal{A}_{1,2}(\tau_2)\delta(k_1-\lambda_1)\delta(k_3-\lambda_2) + \cdots \\
+ \mathcal{A}_{1,N-1}(\tau_2)\delta(k_1-\lambda_1)\delta(k_3-\lambda_{N-1}) + \\
\mathcal{A}_{2,1}(\tau_2)\delta(k_1-\lambda_2)\delta(k_3-\lambda_1) + \mathcal{A}_{2,2}(\tau_2)\delta(k_1-\lambda_2)\delta(k_3-\lambda_2) + \cdots \\
+ \mathcal{A}_{2,N-1}(\tau_2)\delta(k_1-\lambda_2)\delta(k_3-\lambda_{N-1}) + \\
\vdots \\
\mathcal{A}_{N-1,1}(\tau_2)\delta(k_1-\lambda_{N-1})\delta(k_3-\lambda_1) + \mathcal{A}_{N-1,2}(\tau_2)\delta(k_1-\lambda_{N-1})\delta(k_3-\lambda_2) \\
+ \cdots + \mathcal{A}_{N-1,N-1}(\tau_2)\delta(k_1-\lambda_{N-1})\delta(k_3-\lambda_{N-1})
\end{gathered}$$

We fixed the characteristic rates λ_1, λ_2, …, λ_{N-1}, by enforcing the values obtained from our fits to the 2nd-order TCFs of the corresponding data sets. (*ii*) We converted this test function to the time domain by multiplying the trial amplitudes by the corresponding exponential factors given by Eq. (S3.7). (*iii*) We computed χ^2-difference function between the experimental data and the test function using MATLAB's optimization routine "fmincon." By re-iterating the above procedure to minimize the χ^2-difference function, we obtained a set of optimized fit parameters $\mathcal{A}_{n,m}$.

Of course, the ILT is an ill-conditioned problem when applied to noisy single-molecule data.[3] To ascertain the reliability of our ILT algorithm, we performed a series of optimizations to a simulated test function that incorporated various levels of noise (see Fig. S3.5). From these control simulations, we concluded that it was possible to achieve robust results for noise levels consistent with our data following the procedure described above.

III. Comment About Single-Molecule FRET Efficiency Values

In a recent paper (6), our group used single-molecule FRET techniques with 100-ms time resolution to investigate gp32 binding to the same 3'-Cy3/Cy5-p(dT)$_{15}$-p/t DNA substrate as that used in the current work. Although the general pattern of high and low E_{FRET} values corresponding, respectively, to unbound, singly-bound, and doubly-bound states are the same for both studies, the specific numerical values of E_{FRET} for the two studies are shifted by ~0.2. The data reported in reference (6) was measured using a different instrument than those reported in the current work. Because the FRET signal $E_{FRET} = I_{Cy5}/\left(I_{Cy3} + I_{Cy5}\right)$ is based on the ratio of donor and acceptor chromophore fluorescence intensities, differences in the donor and acceptor detection efficiencies will result in different E_{FRET} values between the two instruments. In the current work, the acceptor Cy5 detection efficiency is significantly higher than that of the instrument used in reference (6), such that the E_{FRET} values are shifted.

IV. Multi-Dimensional Parameter Optimization Procedure

To efficiently explore the space of input parameters used to calculate histograms of E_{FRET} values, 2nd-order TCFs, and 4th-order TCFs, we implemented an automated computer optimization procedure. This procedure used a genetic algorithm to search for global solutions (7), and commercial software (KNITRO) (8, 9) to refine those solutions. The genetic algorithm implemented a series of random guesses for the input parameters, which were subsequently refined by mixing the parameters from a family of the most successful trials. We repeated the above process until further refinements did not improve the results. We then entered the results of the genetic algorithm into the KNITRO software, which made additional adjustments to the fit parameters to achieve a globally optimized solution. We implemented several hundred genetic algorithm searches to produce ~1,000 initial parameter guesses, which were each optimized using the KNITRO software to obtain the best-fit solutions to the specific model.

To quantify the goodness-of-fit of a given trial, we defined a 'target function' G_{total} as the weighted sum of chi-square differences, χ_2^2 and χ_4^2, between experimentally- and theoretically-derived 2nd- and 4th-order TCFs, respectively.

$$G_{total} = \sum_{i=1}^{m} w_2(\tau_i)\chi_2^2(\tau_i) + \sum_{k}^{p}\left\{\sum_{i,j}^{m,m} w_4\left(\tau_{1,i},\tau_{3,j},\tau_{2,k}\right)\chi_4^2\left(\tau_{1,i},\tau_{3,j}\right)\left[\tau_{2,k}\right]\right\} + w^{eq}\sum_{i=1}^{N} E_i\, p_i^{eq} \qquad \text{(S3.9)}$$

The optimization described above corresponds to the minimization of G_{total} with respect to the input parameters. The first term of Eq. (S3.9) is a weighted sum, with weighting function $w_2(\tau_i)$, of the chi-squared differences $\chi_2^2(\tau_i)$ over the m incrementally sampled time intervals τ_i $(i = 1,2,\dots,m)$. Similarly, the second term of Eq. (S3.9) is a two-dimensional weighted sum of the chi-square differences $\chi_4^2\left(\tau_{1,i},\tau_{3,j}\right)$ over the $m \times m$ incrementally sampled intervals $\tau_{1,i}$ and $\tau_{3,j}$, with weighting function $w_4\left(\tau_{1,i},\tau_{3,j},\tau_{2,k}\right)$. An additional weighted sum of $\chi_4^2\left(\tau_{1,i},\tau_{3,j}\right)\left[\tau_{2,k}\right]$ was carried out over the p sampled time intervals $\tau_{2,k}$. Finally, the third term of Eq. (S3.9) is the weighted contribution, with weighting factor w^{eq}, due to the equilibrium distribution of states. We found that consistent and robust results could be achieved using weighting functions w_2 and w_4 inversely proportional to time such that the short timescale components of the TCFs received nearly equal weight as the longer timescale components. The scaling factor was adjusted so that each contributing function influenced the fitting algorithm to a similar extent. For the calculations presented in this work, we therefore used the weighting functions $w_2(\tau_i) = 30{,}000/\tau_i$, $w_4\left(\tau_{1,i},\tau_{3,j},\tau_{2,k}\right) = 1{,}000/\left(\tau_{1,i}\cdot\tau_{2,k}\right)$, and $w^{eq} = 10$.

Chapter IV

1. Derivation of the Solution to the Transport Master Equation

To solve the transport master equation (Eq. 4.11) we first construct a square, invertible matrix, which we call the modal matrix, U, by combining the N, $N \times 1$ eigenvector of K into a $N \times N$ matrix of column vectors

$$U = (\,\vec{\boldsymbol{v}}_{\boldsymbol{1}}|\,\vec{\boldsymbol{v}}_{\boldsymbol{2}}|\,\dots\,|\,\vec{\boldsymbol{v}}_{\boldsymbol{N}})$$

$$= \begin{pmatrix} v_1^1 & v_2^1 & \cdots & v_N^1 \\ v_1^2 & v_2^2 & & v_N^2 \\ \vdots & & \ddots & \vdots \\ v_1^N & v_2^N & \cdots & v_N^N \end{pmatrix}. \tag{A4.1}$$

The modal matrix is the similarity transform that diagonalizes the rate matrix,

$$\Lambda = U^{-1}\,K\,U \tag{A4.2}$$

where Λ is the diagonal matrix of the eigenvalues of K.

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{\mathrm{N}} \end{pmatrix} \tag{A4.3}$$

The opposite similarity transform allows us to write the rate matrix in the eigen-basis. To show this is true, we left-multiply both sides of Eq. A4.2 by U, then right-multiply both sides by U^{-1}, giving $U\Lambda U^{-1} = UU^{-1}\,K\,UU^{-1}$. Using the identity

$$U^{-1}\,U \;= U\,U^{-1} = 1 \tag{A4.4}$$

we arrive at Eq. A4.5

$$K = U\;\Lambda\,U^{-1}. \tag{A4.5}$$

Since we are solving a first order ordinary differential equation, the solution to the transport master equation should have exponential time dependence, as shown in Eq. 4.15. Next, we will build a matrix with the proper time dependence using similarity transforms. Note that the relations to transform between K and Λ, would be unchanged by multiplication by a variable t, giving

$$\mathit{\Lambda}\,t = U^{-1}\,(\,K\,t\,)\,U \tag{A4.6a}$$

$$K\,t = U^{-1}\,(\,\mathit{\Lambda}\,t\,)\,U \tag{A.6b}$$

Using the identity $U^{-1}\,U \;= U\,U^{-1} = 1$ and Eq. A4.2 we also find the following relation for transforming powers of these matrices

$$\Lambda^n = U \; K^n \, U^{-1}. \tag{A4.7}$$

Then if we take the exponential of the K-matrix multiplied by t, we get $e^{K\,t}$. Now we calculate the similarity transform of this expression by expanding the exponential as a sum.

$$U \, e^{K\,t} \, U^{-1} = U \, \textstyle\sum_{n=0}^{\infty} \frac{(K\,t)^n}{n!} \, U^{-1} \tag{A4.8a}$$

$$= \textstyle\sum_{n=0}^{\infty} \frac{U\,(K\,t)^n\,U^{-1}}{n!} \tag{A4.8b}$$

Using Eqs. A.6 and A.7, we simplify this similarity transform, giving

$$= \sum_{n=0}^{\infty} \frac{(\Lambda\,t)^n}{n!} = \; e^{\Lambda\,\mathrm{t}}$$

which is the diagonal matrix of exponential time dependence containing all the eigenvalues of our rate matrix

$$e^{\,\Lambda\,t} = \begin{pmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_N t} \end{pmatrix}. \tag{A4.9}$$

Now, we use this expression to calculate the matrix of conditional probabilities by

$$P(t) = U \, e^{\Lambda\,t} \, U^{-1} \, P_0 \tag{A4.10}$$

where P_0 is the identity matrix accounting for all possible initial conditions, and U is the modal matrix. When expanded this becomes

$$P(t) = \begin{pmatrix} v_1^1 & v_2^1 & \cdots & v_N^1 \\ v_1^2 & v_2^2 & & v_N^2 \\ \vdots & & \ddots & \vdots \\ v_1^N & v_2^N & \cdots & v_N^N \end{pmatrix} \begin{pmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_N t} \end{pmatrix} \begin{pmatrix} v_1^1 & v_2^1 & \cdots & v_N^1 \\ v_1^2 & v_2^2 & & v_N^2 \\ \vdots & & \ddots & \vdots \\ v_1^N & v_2^N & \cdots & v_N^N \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$. \tag{A4.11}$$

This product (Eq. A4.10) gives a matrix of conditional probabilities, $P(t)$, in which of the corresponds to a conditional probability, $p_{ji}(\tau)$ (Eq. 4.15): $p_{ji}(t) = p_j^{eq} + c_2^i v_2^j e^{\lambda_2 t} + \cdots + c_N^i v_N^j e^{\lambda_N t}$. The expansion coefficients, c_j^i, are equal to the product $U^{-1}\,P_0$. They are the terms which correspond to the boundary conditions of the transport master equation, Eq. 4.11. In Markov state formalism, the matrix of conditional probabilities evaluated at a certain time lag is often referred to as the stochastic matrix or transition matrix.

2. Exponential Form of the Time Correlation Function

For our three-state system, we can use the $P(t)$ solution to Eq. 4.15 to write the two-point correlation function in a form equivalent to Eq. 4.2 but grouping terms by common exponential factors, giving

$$\bar{C}^{(2)}(\tau) = \mathcal{A}_1 e^{\lambda_1 \tau} + \mathcal{A}_2 e^{\lambda_2 \tau}, \tag{A.12}$$

or expanded more generally for N states, as

$$\bar{C}^{(2)}(\tau) = \sum_{n=1}^{N-1} \mathcal{A}_n e^{\lambda_n \tau}. \tag{A.13}$$

where $\mathcal{A}_n = \sum_{i,j=1}^{N-1} \delta E_j v_n^j c_n^i \delta E_i p_i^{eq}$. This form explicitly shows that the structure of $\bar{C}^{(2)}$ is a sum of exponentials. A similar analysis can be done for the four-point time correlation functions. Since we are using the difference time correlation functions, subtracting off the average value of the observable cancels with the overall constant term that results from one eigenvalue being zero. Thus leaving $N-1$ exponential decays for a system with N states. From this, we expect the two-point correlation function to have two exponential decays for a three-state model.

REFERENCES CITED

References from Chapter I

1. B. Alberts - Molecular Biology Of The Cell 4th Ed. *J Chem Inf Model*. 1989. doi:10.1017/CBO9781107415324.004

2. Von Hippel PH, Printz MP. Dynamic aspects of DNA structure as studies by hydrogen exchange. *Fed Proc*. 1965.

3. Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Rüger W. Bacteriophage T4 Genome. *Microbiol Mol Biol Rev*. 2003. doi:10.1128/mmbr.67.1.86-156.2003

4. Karam JD, Miller ES. Bacteriophage T4 and its relatives. *Virol J*. 2010. doi:10.1186/1743-422X-7-293

5. Mueser TC, Hinerman JM, Devos JM, Boyer RA, Williams KJ. Structural analysis of bacteriophage T4 DNA replication: A review in the Virology Journal series on bacteriophage T4 and its relatives. Presented at the: 2010. doi:10.1186/1743-422X-7-359

6. Ha T, Enderle T, Ogletree DF, Chemla DS, Selvin PR, Weiss S. Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc Natl Acad Sci U S A*. 1996;93(13):6264-6268. doi:10.1073/pnas.93.13.6264

References from Chapter II

1. Lee, W.; Jose, D.; Phelps, C.; Marcus, A. H.; von Hippel, P. H., A Single-Molecule View of the Assembly Pathway, Subunit Stoichiometry and Unwinding Activity of the Bacteriophage T4 Primosome (Helicase-Primase) Complex. *Biochemistry* **2013**, *52*, 3157-3170.
2. Phelps, C.; Lee, W.; Jose, D.; von Hippel, P. H.; Marcus, A. H., Single-Molecule Fret and Linear Dichroism Studies of DNA ‘Breathing’ and Helicase Binding at Replication Fork Junctions. *Proc Natl Acad Sci U S A* **2013**, *110*, 17320-17325.
3. Myong, S., M. M. Bruno, A. M. Pyle, T. Ha, Spring-Loaded Mechanism of DNA Unwinding by Hepatitis C Virus Ns3 Helicase. *Science* **2007**, *317*, 513-16.
4. Murphy, M. C.; Rasnik, I.; Cheng, W.; Lohman, T. M.; Ha, T., Probing Single-Stranded DNA Conformational Flexibility Using Fluorescence Spectroscopy. *Biophys. J.* **2004**, *86*, 2530-2537.

5. Rasnik, I.; Myong, S.; Cheng, W.; Lohman, T. M.; Ha, T., 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*, 395-408.

6. Rasnik, I.; Myong, S.; Ha, T., Unraveling Helicase Mechanisms One Molecule at a Time. *Nucleic Acids Res.* **2006**, *34*, 4225-4231.

7. Morten, M. J.; Peregrina, J. R.; Figueira-Gonzalez, M.; Ackermann, K.; Bode, B. E.; White, M. F.; Penedo, J. C., Binding Dynamics of a Monomeric Ssb Protein to DNA: A Single-Molecule Multi-Process Approach. *Nucleic Acids Res* **2015**, *doi: 10.1093/nar/gkv1225*, 1 - 18.

8. Lee, W.; Gillies, J. P.; Jose, D.; Israels, B. A.; von Hippel, P. H.; Marcus, A. H., Single-Molecule Fret Studies of the Cooperative and Non-Cooperative Binding Kinetics of the Bacteriophage T4 Single-Stranded DNA Binding Protein (Gp32) to Ssdna Lattices at Replication Fork Junctions. *Nucleic Acids Res* **2016**, *doi: 10.1093/nar/gkw863*, 1-20.

9. Santoso, Y.; Joyce, C. M.; Potapova, O.; Le Reste, L.; Hohlbein, J.; Torella, J. P.; Grindley, N. D. F.; Kapanidis, A. N., Conformational Transitions in DNA Polymerase I Revealed by Single-Molecule Fret. *Proc Natl Acad Sci U S A* **2010**, *107*, 715-720.

10. von Hippel, P. H.; Johnson, N. P.; Marcus, A. H., 50 Years of DNA 'Breathing': Reflections on Old and New Approaches. *Biopolymers* **2013**, *99*, 923-954.

11. Chung, H. S.; McHale, K.; Louis, J. M.; Eaton, W. A., Single-Molecule Fluorescence Experiments Determine Protein Folding Transition Path Times. *Science* **2012**, *335*, 981-984.

12. Heuer, A., Information Content of Multitime Correlation Functions for the Interpretation of Structural Relaxation in Glass-Forming Systems. *Phys. Rev. E* **1997**, *56*, 730-740.

13. Yang, S.; Cao, J., Two-Event Echos in Single-Molecule Kinetics: A Signature of Conformational Fluctuations. *J. Phys. Chem. B* **2001**, *105*, 6536-6549.

14. Barsegov, V.; Chernyak, V.; Mukamel, S., Multitime Correlation Functions for Single Molecule Kinetics with Fluctuating Bottlenecks. *J. Chem. Phys.* **2002**, *116*, 4240-4251.

15. Senning, E. N.; Lott, G. A.; Fink, M. C.; Marcus, A. H., Ii. Kinetic Pathways of Switching Optical Conformations in Dsred by 2d Fourier Imaging Correlation Spectroscopy. *J. Phys. Chem. B* **2009**, *113*, 6854-6860.

16. Senning, E. N.; Marcus, A. H., Subcellular Dynamics and Protein Conformation Fluctuations Measured by Fourier Imaging Correlation Spectroscopy. *Annu. Rev. Phys. Chem.* **2010**, *61*, 111-128.

17. Verma, S. D.; Vanden Bout, D. A.; Berg, M. A., When Is a Single Molecule Heterogeneous? A Multidimensional Answer and Its Application to Dynamics near the Glass Transition. *J. Chem. Phys.* **2015**, *143*, 024110.

18. Ono, J.; Takada, S.; Saito, S., Couplings between Hierarchical Conformational Dynamics from Multi-Time Correlation Functions and Two-Dimensional Lifetime Spectra: Application to Adenylate Kinase. *J. Chem. Phys.* **2015**, *142*, 212404.

19. Qian, H.; Elson, E. L., Fluorescence Correlation Spectroscopy with High-Order and Dual-Color Correlation to Probe Nonequilibrium Steady States. *Proc Natl Acad Sci U S A* **2004**, *101*, 2828-2833.

20. Kryvohuz, M.; Mukamel, S., Nonlinear Response Theory in Chemical Kinetics. *J. Chem. Phys.* **2014**, *140*, 034111.

21. Kryvohuz, M.; Mukamel, S., Multidimensional Measures of Response and Fluctuations in Stochastic Dynamical Systems. *Phys. Rev. A* **2012**, *86*, 043818.

22. Pelton, M.; Smith, G.; Scherer, N. F.; Marcus, R. A., Evidence for a Diffusion-Controlled Mechanism for Fluorescence Blinking of Colloidal Quantum Dots. *Proc Natl Acad Sci U S A* **2007**, *104*, 14249-14254.

23. McKinney, S. A.; Joo, C.; Ha, T., Analysis of Single-Molecule Fret Trajectories Using Hidden Markov Modeling. *Biophys. J.* **2006**, *91*, 1941-1951.

24. Reichl, L. E., *A Modern Course in Statistical Physics*, 2nd ed.; John Wiley & Sons, Inc.: New York, 1998.

25. Mukamel, S., *Principles of Nonlinear Optical Spectroscopy*; Oxford University Press: Oxford, 1995.

26. Mukamel, S., D. Abramavicius, L. Yang, W. Zhuang, I. V. Schweigert, D. V. Voronine, Coherent Multidimensional Optical Probes for Electron Correlations and Exciton Dynamics: From Nmr to X-Rays. *Acc. Chem. Res.* **2009**, *42*, 553-562.

27. Khurmi, C.; Berg, M. A., Parallels between Multiple Population-Period Trasient Spectroscopy and Multidimensional Coherence Spectroscopies. *J. Chem. Phys.* **2008**, *129*, 064504-1-17.

28. Kowalczykowski, S. C.; Lonberg, N.; Newport, J. W.; von Hippel, P. H., Interactions of Bacteriophage T4-Coded Gene 32 Protein with Nucleic Acids. *J. Mol. Biol.* **1981**, *145*, 75-104.

29. Jose, D.; Weitzel, S. E.; Baase, W. A.; von Hippel, P. H., Mapping the Interactions of the Single-Stranded DNA Binding Protein of Bacteriophage T4 (Gp32) with DNA Lattices at Single Nucleotide Resolution: Gp32 Monomer Binding. *Nucleic Acids Res* **2015**, *43*, 9276-9290.

30. McGhee, J. D.; von Hippel, P. H., Theoretical Aspects of DNA-Protein Interactions: Cooperative and Non-Cooperative Binding of Large Ligands to a One-Dimensional Homogeneous Lattice. *J. Mol. Biol.* **1974**, *86*, 469-489.

31. Epstein, I. R., Cooperative and Non-Cooperative Binding of Large Ligands to a Finite One-Dimensional Lattice. A Model for Ligand-Oligonucleotide Interactions. *Biophys. Chem.* **1978**, *8*, 327-339.

32. Noé, F.; Fischer, S., Transition Networks for Modeling the Kinetics of Conformational Change in Macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154-162.

33. Colquhoun, D.; Dowsland, K. A.; Beato, M.; Plested, A. J. R., How to Impose Microscopic Reversibility in Complex Reaction Mechanisms. *Biophys. J.* **2004**, *86*, 3510-3518.

34. Perdomo-Ortiz, A.; Widom, J. R.; Lott, G. A.; Aspuru-Guzik, A.; Marcus, A. H., Conformation and Electronic Population Transfer in Membrane Supported Self-Assembled Porphyrin Dimers by Two-Dimensional Fluorescence Spectroscopy. *J. Phys. Chem. B* **2012**, *116*, 10757-10770.

35. Steinbach, P. J.; Ionescu, R.; Matthews, C. R., Anaysis of Kinetics Using a Hybrid Maximum-Entropy / Nonlinear-Least-Squares Method: Application to Protein Folding. *Biophys. J.* **2002**, *82*, 2244-2255.

36. Byrd, R. H.; Nocedal, J.; Waltz, R. A., Knitro: An Integrated Package for Nonlinear Optimization. In *Large-Scale Nonlinear Optimization*, Pillo, G.; Roma, M., Eds. Springer-Verlag: Berlin, Germany, 2006; pp 35-59.

References from Chapter III

1. Johnson A & O'Donnell M (2005) Cellular DNA Replicases: Components and Dynamics at the Replication Fork. *Annu. Rev. Biochem.* 74:283-315.

2. Mueser TC, Hinerman JM, Devos JM, Boyer RA, & Williams KJ (2010) Structural Analysis of Bacteriophage T4 DNA Replication: A review in the Virology Journal Series on Bacteriophage T4 and its Relatives. *Virol. J.* 7:359.

3. Nossal NG (1994) The Bacteriphage T4 DNA Replication Fork. *Molecular Biology of Bacteriphage T4*, eds Karam JD, Kreuzer KN, & Hall DH (American Society for Microbiology, Washington, D.C.), pp 43-54.

4. Nossal NG, Makhov AM, Chastain PD, 2nd, Jones CE, & Griffith JD (2007) Architecture of the Bacteriophage T4 Replication Complex Revealed with Nanoscale Biopointers. *J. Biol. Chem.* 282(2):1098-1108.

5. Alberts BM (1987) Prokaryotic DNA Replication Mechanisms. *Philos. Trans. Royal Soc. London B Biol. Sci.* 317(1187):395-420.

6. Alberts BM & Frey L (1970) T4 Bacteriophage Gene 32: A Structural Protein in the Replication and Recombination of DNA. *Nature* 227:1313-1318.

7. Marceau AH (2012) Functions of Single-Strand DNA-Binding Proteins in DNA Replication, Recombination and Repair. *Meth. Mol. Biol.* 922:1-20.

8. Kowalczykowski SC, Lonberg N, Newport JW, & von Hippel PH (1981) Interactions of Bacteriophage T4-Coded Gene 32 Protein with Nucleic Acids. *J. Mol. Biol.* 145:75-104.

9. Jose D, Weitzel SE, Baase WA, Michael MM, & von Hippel PH (2015) Mapping the Interactions of the Single-Stranded DNA Binding Protein of Bacteriophage T4 (gp32) with DNA Lattices at Single Nucleotide Resolution: Polynucleotide Binding and Cooperativity. *Nucleic Acids Res* 43:9291-9305.

10. Jensen DE, Kelly RC, & von Hippel PH (1976) DNA "Melting" Proteins. II. Effects of Bacteriophage T4 Gene 32-Protein Binding on the Conformation and Stability of Nucleic Acid Structures. *J. Biol. Chem.* 251(22):7215-7228.

11. Kozlov AG*, et al.* (2015) Intrinsically Disordered C-Terminal Tails of *E. coli* Single Stranded DNA Binding Protein Regulate Cooperative Binding to Single Stranded DNA *J. Mol. Biol.* 427(4):763-774.

12. Chen R & Wold MS (2014) Replication Protein A: Single-Stranded DNA's First Responder: Dynamic DNA-Interactions Allow Replication Protein A to Direct Single-Strand DNA Intermediates into Different Pathways for Synthesis or Repair. *Bioessays* 36(12):1156-1161.

13. Nguyen B*, et al.* (2014) Diffusion of Human Replication Protein A Along Single-Stranded DNA. *J. Mol. Biol.* 426:3246-3261.

14. Roy R, Kozlov AG, Lohman TM, & Ha T (2009) SSB Protein Diffusion on Single-Stranded DNA Stimulates RecA Filament Formation. *Nature* 461:1092-1097.

15. Lee W*, et al.* (2016) Single-Molecule FRET Studies of the Cooperative and Non-Cooperative Binding Kinetics of the Bacteriophage T4 Single-Stranded DNA Binding Protein (gp32) to ssDNA Lattices at Replication Fork Junctions. *Nucleic Acids Res* 44:10691-10710.

16. Phelps C, Lee W, Jose D, von Hippel PH, & Marcus AH (2013) Single-Molecule FRET and Linear Dichroism Studies of DNA 'Breathing' and Helicase Binding at Replication Fork Junctions. *Proc Natl Acad Sci U S A* 110:17320-17325.

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

18. Nair G, Zhao J, & Bawendi MG (2011) Biexciton Quantum Yield of Single Semiconductor Nanocrystals from Photon Statistics *Nano Letters* 11:1136-1140.

19. McKinney SA, Joo C, & Ha T (2006) Analysis of Single-Molecule FRET Trajectories using Hidden Markov Modeling. *Biophys. J.* 91:1941-1951.

20. Phelps C, Israels B, Marsh MC, von Hippel PH, & Marcus AH (2016) Using Multi-Order Time Correlation Functions (TCFs) to Elucidate Biomolecular Reaction Pathways from Microsecond Single-Molecule Fluorescence Experiments *J. Phys. Chem. B* 120:13003-13016.

21. Heuer A (1997) Information Content of Multitime Correlation Functions for the Interpretation of Structural Relaxation in Glass-Forming Systems. *Phys. Rev. E* 56:730-740.

22. Yang S & Cao J (2001) Two-Event Echos in Single-Molecule Kinetics: A Signature of Conformational Fluctuations. *J. Phys. Chem. B* 105:6536-6549.

23. Barsegov V, Chernyak V, & Mukamel S (2002) Multitime Correlation Functions for Single Molecule Kinetics with Fluctuating Bottlenecks. *J. Chem. Phys.* 116:4240-4251.

24. Kryvohuz M & Mukamel S (2012) Multidimensional Measures of Response and Fluctuations in Stochastic Dynamical Systems. *Phys. Rev. A* 86:043818.

25. Kryvohuz M & Mukamel S (2014) Nonlinear Response Theory in Chemical Kinetics. *J. Chem. Phys.* 140:034111.

26. Senning EN, Lott GA, Fink MC, & Marcus AH (2009) II. Kinetic Pathways of Switching Optical Conformations in DsRed by 2D Fourier Imaging Correlation Spectroscopy. *J. Phys. Chem. B* 113:6854-6860.

27. Senning EN & Marcus AH (2010) Subcellular Dynamics and Protein Conformation Fluctuations Measured by Fourier Imaging Correlation Spectroscopy. *Annu. Rev. Phys. Chem.* 61:111-128.

28. Verma SD, Vanden Bout DA, & Berg MA (2015) When is a Single Molecule Heterogeneous? A Multidimensional Answer and its Application to Dynamics Near the Glass Transition. *J. Chem. Phys.* 143:024110.

29. Ono J, Takada S, & Saito S (2015) Couplings Between Hierarchical Conformational Dynamics from Multi-Time Correlation Functions and Two-Dimensional Lifetime Spectra: Application to Adenylate Kinase. *J. Chem. Phys.* 142:212404.

30. Khurmi C & Berg MA (2008) Parallels Between Multiple Population-Period Trasient Spectroscopy and Multidimensional Coherence Spectroscopies. *J. Chem. Phys.* 129:064504-064501-064517.

31. Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy* (Oxford University Press, Oxford).

32. Palmer AG & Thompson NL (1987) Molecular Aggregation Characterized by High Order Autocorrelation in Fluorescence Correlation Spectroscopy. *Biophys. J.* 52:257-270.

33. Delius H, Mantell NJ, & Alberts B (1972) Characterization by Electron Microscopy of the Complex Formed Between T4 Bacteriophage Gene 32-Protein and DNA. *J. Mol. Biol.* 67(3):341-350.

34. McGhee JD & von Hippel PH (1974) Theoretical Aspects of DNA-Protein Interactions: Cooperative and Non-Cooperative Binding of Large Ligands to a One-Dimensional Homogeneous Lattice. *J. Mol. Biol.* 86:469-489.

35. Jose D, Weitzel SE, Baase WA, & von Hippel PH (2015) Mapping the Interactions of the Single-Stranded DNA Binding Protein of Bacteriophage T4 (gp32) with DNA Lattices at Single Nucleotide Resolution: gp32 Monomer Binding. *Nucleic Acids Res* 43:9276-9290.

36. Reichl LE (1998) *A Modern Course in Statistical Physics* (John Wiley & Sons, Inc., New York) 2nd Ed.

37. Casas-Finet JR, Fischer KR, & Karpel RL (1992) Structural Basis for the Nucleic Acid Binding Cooperativity of Bacteriophage T4 Gene 32 Protein: The (Lys/Arg)3(Ser/Thr)2 (LAST) Motif. *Proc Natl Acad Sci U S A* 89(3):1050-1054.

38. Brianzi P & Frontini M (1991) On the Regularized Inversion of the Laplace Transform. *Inverse Problems* 7:355-368.

39. Song Y-Q*, et al.* (2002) T1-T2 Correlation Spectra Obtained Using a Fast Two-Dimensional Laplace Transform. *J. Mag. Res.* 154:261-268.

40. Fordham EJ, Sezginer A, & Hall LD (1995) Imaging Multiexponential Relaxation in the (y, loge T1) Plane, with Application to Clay Filtration in Rock Cores. *J. Mag. Res.* 113:139-150.

41. Kowalczykowski SC, Lonberg N, Newport JW, Paul LS, & von Hippel PH (1980) On the Thermodynamics and Kinetics of the Cooperative Binding of Bacteriophage T4-Coded Gene 32 (Helix Destabilizing) Protein to Nucleic Acid Lattices. *Biophys. J.* 32:403-418.

42. Lohman TM & Kowalczykowski SC (1981) Kinetics and Mechanism of the Association of the Bacteriophage T4 Gene 32 (Helix Destabilizing) Protein with Single-Stranded Nucleic Acids. *J. Mol. Biol.* 152:67-109.

References from Chapter IV

1. Crick F, Watson J. Molecular Structure of Nucleic Acids. *Nature*. 1953:1-2. http://www.nature.com/nature/journal/v171/n4356/pdf/171737a0.pdf%5Cnpapers2://publication/uuid/F22F0EF7-45F2-416F-9FBB-4EA542F395C9.

2. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Chromosomal DNA and Its Packaging in the Chromatin Fiber. *Mol Biol cell, 4th Ed*. 2002.

3. Altan-Bonnet G, Libchaber A, Krichevsky O. Bubble Dynamics in Double-Stranded DNA. *Phys Rev Lett*. 2003;90(13):138101. doi:10.1103/PhysRevLett.90.138101

4. von Hippel PH, Marcus AH. 50 years of DNA ‘Breathing’: Reflections on Old and New Approaches. *October*. 2008;141(4):520-529. doi:10.1016/j.surg.2006.10.010.Use

5. Vaisman, Woodgate. Ribonucleotide Discrimination by Translesion Synthesis DNA Polymerases. *Physiol Behav*. 2018;176(3):139-148. doi:10.1016/j.physbeh.2017.03.040

6. Ansari A, Kuznetsov S V., Shen Y. Configurational diffusion down a folding funnel describes the dynamics of DNA hairpins. *Proc Natl Acad Sci U S A*. 2001;98(14):7771-7776. doi:10.1073/pnas.131477798

7. Changeux JP, Edelstein S. Conformational selection or induced fit? 50 Years of debate resolved. *F1000 Biol Rep*. 2011;3(1):1-15. doi:10.3410/B3-19

8. Alberts BM, Frey L. T4 bacteriophage gene 32: A structural protein in the replication and recombination of DNA. *Nature*. 1970;227(5265):1313-1318. doi:10.1038/2271313a0

9. Jose D, Weitzel SE, Baase WA, Von Hippel PH. Mapping the interactions of the single-stranded DNA binding protein of bacteriophage T4 (gp32) with DNA lattices at single nucleotide resolution: gp32 monomer binding. *Nucleic Acids Res*. 2015;43(19):9276-9290. doi:10.1093/nar/gkv817

10. Tarantola A. *Inverse Problem Theory*.; 2004. http://www.ipgp.fr/~tarantola/Files/Professional/SIAM/InverseProblemTheory.pdf%0Apapers2://publication/uuid/62A7A164-F0E3-4298-AA2B-222FF673BF4E.

11. Phelps C, Israels B, Marsh MC, Von Hippel PH, Marcus AH. Using Multiorder Time-Correlation Functions (TCFs) to Elucidate Biomolecular Reaction Pathways from Microsecond Single-Molecule Fluorescence Experiments. *J Phys Chem B*. 2016;120(51). doi:10.1021/acs.jpcb.6b08449

12. Murphy MC, Rasnik I, Cheng W, Lohman TM, Ha T. Probing Single-Stranded DNA Conformational Flexibility Using Fluorescence Spectroscopy. *Biophys J*. 2004;86(4):2530-2537. doi:10.1016/S0006-3495(04)74308-8

13. Ambia-Garrido. A model for Structure and Thermodynamics of ssDNA and dsDNA Near a Surface: a Coarse Grained Approach. *Bone*. 2011;23(1):1-7. doi:10.1161/CIRCULATIONAHA.110.956839

14. Chandradoss SD, Haagsma AC, Lee YK, Hwang J-H, Nam J-M, Joo C. Surface Passivation for Single-molecule Protein Studies. *J Vis Exp*. 2014;(86):4-11. doi:10.3791/50549

15. Marttila AT, Laitinen OH, Airenne KJ, et al. Recombinant NeutraLite Avidin: A non-glycosylated, acidic mutant of chicken avidin that exhibits high affinity for biotin and low non-specific binding properties. *FEBS Lett*. 2000. doi:10.1016/S0014-5793(00)01119-4

16. Joo C, McKinney SA, Nakamura M, Rasnik I, Myong S, Ha T. Real-Time Observation of RecA Filament Dynamics with Single Monomer Resolution. *Cell*. 2006. doi:10.1016/j.cell.2006.06.042

17. Rasnik I, McKinney SA, Ha T. Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat Methods*. 2006;3(11):891-893. doi:10.1038/nmeth934

18. D. V. O'Connor WRW. Deconvolution of Fluorescence Decay Curves. A Critical Comparison of Techniques. 1979;83(10). doi:10.1021/j100473a019

19. O'Connor D V., Ware WR, Andre JC. Deconvolution of fluorescence decay curves. A critical comparison of techniques. *J Phys Chem*. 2005;83(10):1333-1343. doi:10.1021/j100473a019

20. Chandler D, Percus JK. Introduction to Modern Statistical Mechanics . *Phys Today*. 1988. doi:10.1063/1.2811680

21. Glasser D, Horn FJM, Meidan R. Properties of certain zero column-sum matrices with applications to the optimization of chemical reactors. *J Math Anal Appl*. 1980;73(2):315-337. doi:10.1016/0022-247X(80)90281-4

22. Reichl LE, Luscombe JH. A Modern Course in Statistical Physics, 2nd Edition *Am J Phys*. 1999. doi:10.1119/1.19118

23. Sanborn ME, Connolly BK, Gurunathan K, Levitus M. Fluorescence Properties and Photophysics of the Sulfoindocyanine Cy3 Linked Covalently to DNA. 2007:11064-11074. doi:10.1021/jp072912u

24. Jia K, Wan Y, Xia A, Li S, Gong F, Yang G. Characterization of photoinduced isomerization and intersystem crossing of the cyanine dye Cy3. *J Phys Chem A*. 2007;111(9):1593-1597. doi:10.1021/jp067843i

25. Ciuba MA, Levitus M. Manganese-Induced Triplet Blinking and Photobleaching of Single Molecule Cyanine Dyes. 2013:3495-3502. doi:10.1002/cphc.201300634

26. Colquhoun D, Dowsland KA, Beato M, Plested AJR. How to impose microscopic reversibility in complex reaction mechanisms. *Biophys J*. 2004;86(6):3510-3518. doi:10.1529/biophysj.103.038679

27. Phelps C, Israels B, Jose D, Marsh MC, Von Hippel PH, Marcus AH. Using microsecond single-molecule FRET to determine the assembly pathways of T4 ssDNA binding protein onto model DNA replication forks. *Proc Natl Acad Sci U S A*. 2017;114(18). doi:10.1073/pnas.1619819114

28. McIntosh DB, Duggan G, Gouil Q, Saleh OA. Sequence-dependent elasticity and electrostatics of single-stranded DNA: Signatures of base-stacking. *Biophys J*. 2014. doi:10.1016/j.bpj.2013.12.018

29. Goddard NL, Bonnet G, Krichevsky O, Libchaber A. Sequence dependent rigidity of single stranded DNA. *Phys Rev Lett*. 2000;85(11):2400-2403. doi:10.1103/PhysRevLett.85.2400

30. Bonnet. Kinetics of conformational fluctuations in DNA hairpin-loops. *Proc Natl Acad Sci*. 1998. doi:10.1088/1367-2630/15/11/113010

31. Gene T, Kelly RC, Jensen DE, Hippel PHVON. “Melting” Proteins. 1976.

32. Jose D, Weitzel SE, Baase WA, Von Hippel PH. Mapping the interactions of the single-stranded DNA binding protein of bacteriophage T4 (gp32) with DNA lattices at single nucleotide resolution: gp32 monomer binding. *Nucleic Acids Res*. 2015;43(19):9276-9290. doi:10.1093/nar/gkv817

References from Chapter V

1. Jones CE, Mueser TC, Nossal NG. Bacteriophage T4 32 Protein Is Required for Helicase-dependent Leading Strand Synthesis When the Helicase Is Loaded by the T4 59 Helicase-loading Protein. *J Biol Chem*. 2004;279(13):12067-12075. doi:10.1074/jbc.M313840200

2. Kowalczykowski SC, Paul LS, Lonberg N, Newport JW, McSwiggen JA, von HP. Cooperative and noncooperative binding of protein ligands to nucleic acid lattices: experimental approaches to the determination of thermodynamic parameters [published erratum appears in Biochemistry 1986 Dec 30;25(26):8473]. *Biochemistry*. 1986.

3. Shamoo Y, Friedman AM, Parsons MR, Konigsberg WH, Steltz TA. Crystal structure of a replication fork single-stranded DNA binding protein (T4 gp32} complexed to DNA. 1995;3(July):362-366.

4. Newport JW, Lonberg N, Kowalczykowski SC, von Hippel PH. Interactions of bacteriophage T4-coded gene 32 protein with nucleic acids. II. Specificity of binding to DNA and RNA. *J Mol Biol*. 1981;145(1):105-121. doi:10.1016/0022-2836(81)90336-3

5. Marintcheva B, Marintchev A, Wagner G, Richardson CC. Acidic C-terminal tail of the ssDNA-binding protein of bacteriophage T7 and ssDNA compete for the same binding surface. *Proc Natl Acad Sci U S A*. 2008;105(6):1855-1860. doi:10.1073/pnas.0711919105

6. Rouzina I, Pant K, Karpel RL, Williams MC. Theory of electrostatically regulated binding of T4 gene 32 protein to single- and double-stranded DNA. *Biophys J*. 2005;89(3):1941-1956. doi:10.1529/biophysj.105.063776

7. Kowalczykowski SC, Lonberg N, Newport JW, von Hippel PH. Interactions of bacteriophage T4-coded gene 32 protein with nucleic acids. *J Mol Biol*. 1981;145(1):75-104. doi:10.1016/0022-2836(81)90335-1

8. Jose D, Weitzel SE, Baase WA, Von Hippel PH. Mapping the interactions of the single-stranded DNA binding protein of bacteriophage T4 (gp32) with DNA lattices at single nucleotide resolution: gp32 monomer binding. *Nucleic Acids Res*. 2015;43(19):9276-9290. doi:10.1093/nar/gkv817

9. Moerner WE, Fromm DP. Methods of single-molecule fluorescence spectroscopy and microscopy. *Rev Sci Instrum*. 2003;74(8):3597-3619. doi:10.1063/1.1589587

10. Nguyen B, Ciuba MA, Kozlov AG, Levitus M, Lohman TM. Protein Environment and DNA Orientation Affect Protein induced Cy3 Fluorescence Enhancement (PIFE). *Biophys J*. 2019;117(1):66-73. doi:10.1016/j.bpj.2019.05.026

11. Lerner E, Ploetz E, Hohlbein J, Cordes T, Weiss S. A Quantitative Theoretical Framework For Protein-Induced Fluorescence Enhancement − Fo¨rster-Type Resonance Energy Transfer (PIFE-FRET). 2016. doi:10.1021/acs.jpcb.6b03692

12. Jensen DE. DNA "melting" proteins. II. Effects of bacteriophage T4 gene 32 protein binding on the conformation and stability of nucleic acid structures. *Proc Natl Acad Sci U S A*. 1976;114(31):8253. doi:10.1073/pnas.1706196114

13. Pant K, Karpel RL, Williams MC. Kinetic regulation of single DNA molecule denaturation by T4 gene 32 protein structural domains. *J Mol Biol*. 2003;327(3):571-578. doi:10.1016/S0022-2836(03)00153-0

14. Casas-Finet JR, Karpel RL. Bacteriophage T4 Gene 32 Protein: Modulation of Protein-Nucleic Acid and Protein-Protein Association by Structural Domains. *Biochemistry*. 1993;32(37):9735-9744. doi:10.1021/bi00088a028

15. Mcghee JD, Hippel PHVON. Theoretical Aspects of DNA-Protein Interactions : Co-operative and Non-co-operative Binding of Large Ligands to a One-dimensional Homogeneous Lattice. 1974:469-489.

References from Appendix

Chapter III SI Section

1. Reichl LE (1998) *A Modern Course in Statistical Physics* (John Wiley & Sons, Inc., New York) 2nd Ed.

2. Phelps C, Israels B, Marsh MC, von Hippel PH, & Marcus AH (2016) Using Multi-Order Time Correlation Functions (TCFs) to Elucidate Biomolecular Reaction Pathways from Microsecond Single-Molecule Fluorescence Experiments *J. Phys. Chem. B* 120:13003-13016.

3. Brianzi P & Frontini M (1991) On the Regularized Inversion of the Laplace Transform. *Inverse Problems* 7:355-368.

4. Song Y-Q*, et al.* (2002) T1-T2 Correlation Spectra Obtained Using a Fast Two-Dimensional Laplace Transform. *J. Mag. Res.* 154:261-268.

5. Fordham EJ, Sezginer A, & Hall LD (1995) Imaging Multiexponential Relaxation in the (y, loge T1) Plane, with Application to Clay Filtration in Rock Cores. *J. Mag. Res.* 113:139-150.

6. Lee W*, et al.* (2016) Single-Molecule FRET Studies of the Cooperative and Non-Cooperative Binding Kinetics of the Bacteriophage T4 Single-Stranded DNA Binding Protein (gp32) to ssDNA Lattices at Replication Fork Junctions. *Nucleic Acids Res* 44:10691-10710.

7. Mitchell M (1996) *An Introduction to Genetic Algorithms* (The MIT Press, Cambridge, Massachusetts).

8. Byrd RH, Nocedal J, & Waltz RA (2006) KNITRO: An Integrated Package for Nonlinear Optimization. *Large-Scale Nonlinear Optimization*, eds Pillo G & Roma M (Springer-Verlag, Berlin, Germany), pp 35-59.

9. Lott GA*, et al.* (2011) Conformation of Self-Assembled Porphyrin Dimers in Liposome Vesicles by Phase-Modulation 2D Fluorescence Spectroscopy. *Proc Natl Acad Sci U S A* 108(40):16521-16526.