Navigating Exponentially Large Spaces in Biology: Methods for Directed Evolution and smFRET Time Series Analysis

Jonathan Eiseman Bronson

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2010

ABSTRACT

Navigating Exponentially Large Spaces in Biology: Methods for Directed Evolution and smFRET Time Series Analysis

Jonathan Eiseman Bronson

The recent explosion of high throughput technologies in many fields of biology has necessitated the use of sophisticated algorithms to guide experimental design and analyze results. This thesis explores two such fields: directed protein evolution and single molecule fluorescence resonance energy transfer analysis. Although the methodologies and applications of the fields differ greatly, they are both limited by a process which scales exponentially with problem size. In the former case, the problem is determining which combination of amino acids should be mutated to enhance or create protein function. In the latter case, the problem is inferring the number of conformations a molecule explores during an experiment and the probability of being in each state at each time point in the experiment. Methods to address both problems will be presented in this thesis.

Contents

List of Tables

List of Figures

Acknowledgments

First and foremost, I would like to thank my thesis advisors: Chris Wiggins, Ruben Gonzalez, Jr. and David Reichman. Everything I know about machine learning and statistics I learned from Chris. His leadership and guidance throughout all the smFRET work has been invaluable. Thank you, Chris, for everything you taught me, and for giving an experimental student the chance to change directions and learn a new and beautiful field of science. Ruben's lab generated all the smFRET data which we analyzed, and Ruben was very influential in shaping our approach to smFRET analysis. In addition, without his helpful input I would have never been able to write data analysis software tailored to the specific needs of the experimental smFRET community. Dave was there to help whenever I had a problem, no matter how large or small. I also was fortunate enough to get a desk in his research group, giving me wonderful exposure to the world of condensed matter physics and, more importantly, his awesome students. I have had so much fun working alongside Richard, Brenda, Tim, Glen, Sy and Carl.

Ruben's student Jingyi Fei carried out the majority of the smFRET experiments I analyzed for this thesis and was heavily involved in planing our approach to smFRET inference and designing the software I wrote. Chris' student Jake Hofman taught me how to use MATLAB and most of what I know about Bayesian inference. A special thanks to Virginia Cornish. I began my thesis work in her research lab and was able to work on and design some very exciting directed evolution projects

there. I would like to thank the National Science Foundation for providing me with a graduate research fellowship. It greatly facilitated the process of being a joint student. I would not have been able to work in this interesting interface of theory and experiment without its support. Finally, I would like to thank my family for all their support over the past five years. You're the best!

Jonathan E. Bronson

New York, June 2010

For my family.

Chapter 1

Introduction

High-throughput experiments have become ubiquitous in biology. Pharmaceutical companies can screen thousands of compounds per day looking for new drugs (Walters and Namchuk 2003), cellular biologists can study millions of nucleotide sequences in an experiment using microarrays (Bernstein et al. 2005) and biophysicists can record the individual dynamics of hundreds of molecules on a single glass slide (Fei et al. 2009). Enzyme and metabolic engineering projects have benefited tremendously from high-throughput screens and selections as well (Aharoni et al. 2005; Kirby and Keasling 2009). While automation has greatly increased the size of experiments and experimental analyses possible, the complexity of most problems in biology scales exponentially, making it impossible to explore all possible outcomes. Intelligent search algorithms must be devised to sift through these exponential spaces. This thesis explores two such fields of biology: directed evolution and single molecule FRET data analysis.

Directed evolution seeks to design novel proteins and metabolic pathways by

generating large libraries of protein or cell variants and screening or selecting for desired activity. A screen requires the experimentalist to look through all possible variants for activity (*i.e.*, by passing them through a fluorescence-activated cell sorter (Cormack et al. 1996)), and a selection removes inactive variants (*i.e.*, by killing all cells which cannot produce the desired target (Park et al. 2006)). Libraries of $10^4 - 10^{15}$ can be searched, depending on the experimental method (Bloom et al. 2006; Wilson et al. 1999). Directed evolution has been substantially more effective than computational optimizations and rational design for designing enzymes and improving metabolic pathways (Bloom et al. 2005). The two most common ways to generate DNA libraries are though error prone PCR and DNA shuffling.

Although 10^{15} is an enormous number, there are 20 possible amino acids at each position in a protein and, therefore, 20^N possible N residue long proteins. With a library of 10^{15}, only $\sim$12 different sites on a protein could be exhaustively searched simultaneously, and most proteins are hundreds of residues long. As a result, much work is spent optimizing libraries, either through judicious choices of amino acid types or positions (Neylon 2004), or by combining the results of hits in a way to maximize potential synergies (Stemmer 1994) (which is the logic behind DNA shuffling).

Nature already has a highly optimized mechanism to evolve novel functionalities: the genetic algorithm. By harnessing the cell to generate targeted DNA libraries and to selectively replicate only cells with functional gene products, it might be possible to drastically improve the results of directed evolutions. Many directed evolution projects utilize the cell for either library generation or selection;

however, "smart cells" capable of generating well designed libraries and systematically performing selections have yet to be realized. This thesis will consider some ways to harness the power of the cell for purposes of directed evolution.

Learning from time series data, such as single molecule FRET data, also presents a challenge which grows exponentially with data length. For a molecule which can adopt K conformations, there are K^T possible trajectories the molecule could explore in a T step time series. Trying to infer the most probable trajectory from a noisy time series by enumeration on a useful time scale would be impossible for all but trivially small systems (*e.g.*, $\mathrm{K} = 2$, $T = 25$). By appealing to graphical modeling, and specifically hidden Markov modeling (Andrec et al. 2003), it is possible to cut that space down to K^2T possible trajectories, which is computationally tractable.

Finding an appropriate model for the data is only the beginning of the inference process. In most experiments, both the trajectory of the molecule and the model parameters describing the molecule are unknown. Usually the number of states in the data is also unknown and must be learned as well. The are many ways of inferring this information from the data. The standard approach is to use the principle of maximum likelihood and the expectation maximization algorithm (MacKay 2003; Bishop 2006). This method has two known pathologies: a strong propensity to overfit (*i.e.* find more states than are supported by the observed data) and convergence to divergent solutions (*i.e.* the algorithm can converge to solutions where a state has zero variance and infinite likelihood, rendering the analysis meaningless). Alternative strategies, which do not suffer from these pathologies, will be

considered in this thesis.

This thesis is organized in two parts. Part one considers ways to harness the cell for improved directed evolution. Chapter 2 describes an attempt to create a cell line which can replicate a plasmid with a high error rate, without affecting the mutation rate of the host chromosome, by using the T7 bacteriophage's DNA replisome to create a DNA replication system in *E. coli* which is orthogonal to the host's DNA replication machinery. Chapter 3 describes an approach to make *in vivo* logic gates, using the yeast three-hybrid assay (Lin et al. 2000). This project is not immediately applicable to directed evolution; however, it addresses the larger issue of cellular logic, which ultimately will need to be considered for sophisticated *in vivo* directed evolution projects.

Part two describes two different methods for smFRET inference. Chapter 4 presents overviews of hidden Markov modeling and single molecule FRET. Chapters 5 and 6 describe how the principle of maximum evidence and the variational Bayesian expectation maximization algorithm can be used to solve the hidden Markov model without the problems associated with maximum likelihood. Chapter 7 proposes a novel method for single molecule FRET inference, which builds on the work of Chapters 5 and 6 but allows for more accurate inference and inference of problems which were previously impossible. Chapter 8 will discuss possible future directions for this project.

Part I

Directed Evolution

Chapter 2

The Use of T7 DNA Polymerase for Error Prone PCR

2.1 Background

2.1.1 epPCR

Error prone PCR (epPCR) is one of the most commonly used methods to create random point mutations in a targeted region of DNA (Romero and Arnold 2009). The method was first developed in 1989 using Taq polymerase (since it lacks a 3' → 5' exonuclease proofreading activity) under conditions that promote poor fidelity lacks (Leung et al. 1989). The first directed evolution experiment using epPCR came three years later (Rice et al. 1992). Numerous directed evolution experiments have been carried out since, including the directed evolution of thermostable enzymes for industrial use, novel binding proteins with potential medicinal appli-

cations, and β-lactamases to understand the evolution of bacterial drug resistance (Giver et al. 1998; Binz et al. 2005; Goldberg et al. 2003).

Typically mutation rates are tuned to create very few (i.e. 1–3) mutations per gene per round of epPCR, although error rates an order of magnitude higher are used in some experiments. Higher error rates can create gene sequences which are enriched for positively coupled mutations, but far fewer gene sequences will result in functional proteins (Drummond et al. 2005). Often epPCR is combined with other methods as well, such as DNA shuffling (Stemmer 1994).

Although epPCR is an effective way to generate mutations in a gene target, implementation of the method is somewhat tedious and suboptimal for evolving new protein function. Each round of epPCR requires the researcher to run the epPCR reaction on the target gene, ligate the gene into an expression vector, transform the library of vectors into a cell line and screen or select for functional proteins. Often the best hits from the screen or selection are collected, purified and used as a template for another round of epPCR. A typical directed evolution experiment might require several dozen rounds of epPCR (Goldberg et al. 2003). In these experiments, transforming the DNA into the cell line is what limits the size of the library that can be screened or selected. Transformation efficiencies of up to 10^{10} transformants per μg DNA can be achieved in *E. coli* using electroporation (Dower et al. 1988). Transformation efficiencies using other methods and/or cell lines are typically lower. Since the transformation process is highly stressful to the cell, random mutation of the cellular genome and genetic recombination between cell and transformed plasmid are common during transformations (Foster 2005). Often

the most time consuming step of a directed evolution experiment is screening out false positives that result from recombination between the genome and the library plasmids.

To circumvent some of these problems, researchers have tried using *mutator strains*: cell lines which lack DNA repair mechanisms and/or have highly error prone DNA polymerases (Nguyen and Daugherty 2003). The problem with these strains is that they mutate a cell's entire genome in addition to the target gene, leading to a high false positive rate. It would be far more desirable to have a cell line which only mutates the gene of interest and leaves the cellular genome untouched. Some progress has been made in this regard using an error prone DNA polymerase I (PolI) in *E. coli* (Fabret et al. 2000; Camps et al. 2003). Only three point mutations, D424A, I709N and A759R, were required to increase the error rate of PolI 80,000-fold. PolI plays a minor role in genome replication, but replicates the first few kilobases of the ColE1 origin of replication (ori) found in most commercial plasimds. Provided the gene is only a few kilobases long and can be expressed on a ColE1 plasmid, this method provides a way to selectively increase the mutation rate of the target gene. While PolI's role in genome replication is minor, it is significant enough that the genome mutation error rate is still elevated in these strains[1].

A more desirable option would be to create a cell in which specific plasmids were replicated with DNA replication machinery completely orthogonal to the rest of the cell. Not only would this allow researchers in directed evolution experiments

[1]Estimates for the increased error rate vary, but in my personal experience with the strain created by Camps *et. al.*, the increased background mutation rate was high enough to preclude its use in directed evolution experiments.

to tune the mutation rate of a gene without affecting the cellular genome mutation rate, but it would also provide a synthetic biology platform to study DNA replication (Baker et al. 2006). A synthetic biology model system for DNA replication would be especially helpful to researchers studying DNA replication and cellular maintenance of plasmid copy numbers, because the modifications to the cellular DNA replication machinery necessary to test many hypothesises would disrupt normal cellular function.

2.1.2 T7 bacteriophage

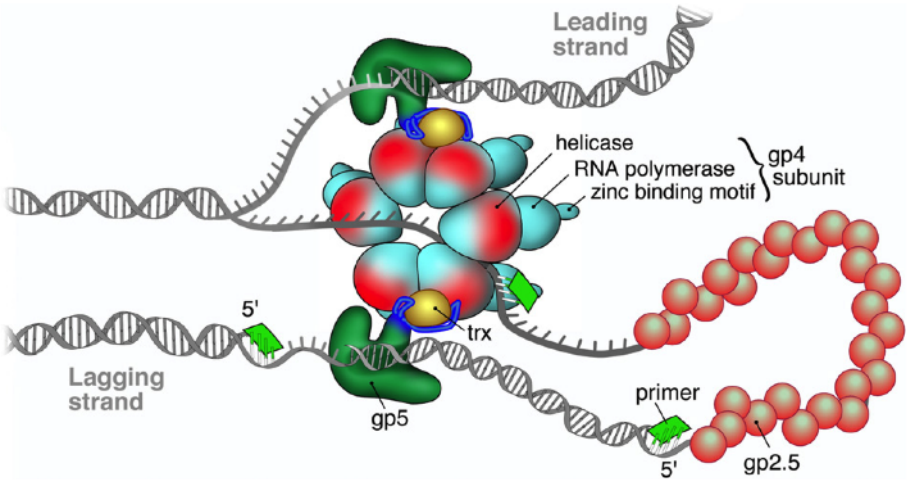

Figure 2.1: Cartoon depiction of T7 DNA replisome during bidirectional DNA replication. The replisome is one of the smallest DNA replisomes known, consisting of only four proteins: T7 DNA polymerase (gp5), T7 helicase/primase (gp4), T7 single stranded binding protein (gp2.5) and *E. coli* thioredoxin increases the processivity of gp5. Figure reproduced from (Perumal et al. 2009)

T7 is an icosahedral bacteriophage, with a capsid diameter of 60 nm and 40 kb double stranded genome (Kruger and Schroeder 1981). It is a lytic virus,

capable of creating ∼200 progeny within 15 minutes of infecting a host *E. coli* cell. More importantly for this work, it has among the simplest DNA replisomes, comprising only four proteins. Three of the proteins, the DNA polymerase (gp5), helicase/primase (gp4) and single stranded binding protein (gp2.5), are encoded by the T7 genome (Perumal et al. 2009). A processivity factor, thioredoxin (trx), is supplied by the host. The T7 RNA polymerase (gp1) is necessary to initiate DNA replication at the T7 ori (oriT7). It has been shown that these genes are sufficient to replicate DNA containing oriT7 both *in vitro* (Fischer and Hinkle 1980) and on a plasmid *in vivo* (Rabkin and Richardson 1988). Replication of DNA via the T7 replisome proceeds bidirectionally.

The crystal structure of gp5 has been solved (Doublie et al. 1998). It has been shown to have high structural homology with the *E. coli* PolI (Ollis et al. 1985). Sequence alignments of PolI and gp5 show the three residues required to make PolI error prone, D424, I709 and A759 correspond to the semi-conserved residues E228, L479 and T523, suggesting that mutating these residues will make gp5 error prone as well. Alternative, it has been shown that simply inactivating the 3' → 5' exonuclease increased error rates in gp5 (Tabor and Richardson 1990).

2.2 Experimental objectives

This project consists of four objectives:

1. Clone gp1, gp2.5, gp4 and gp5 onto a plasmid which expresses the genes at appropriate levels (I will refer to the plasmid containing these genes as pT7).

2. Stably transform pT7 and a plasmid containing oriT7 (pOriT7) into a strain of *E. coli*.

3. Make gp5 error prone.

4. Use the error prone gp5 in an *in vivo* directed evolution experiment, such as the evolution of a β-lactamase to hydrolyze a novel drug target (Camps et al. 2003).

Gp4 has been shown to be toxic to *E. coli* in high concentrations (Rosenberg et al. 1992; Patel et al. 1992). Personal communications with F.W. Studier suggest that a variant of gp4 containing a M64L point mutation (termed gp4A') may be more appropriate for pT7 than gp4. DNA is still efficiently replicated with gp4A', with less toxicity to the host.

The proposed plasmids for this system are diagrammed in Fig. 2.2. Given that the T7 genes are slightly toxic to the host and that they are normally expressed at levels designed for lytic growth of the virus, controlling copy number of the T7 genes is an important consideration. Finding the appropriate expression levels of the replisome genes can be accomplished by making a replisome expression level library. If pOriT7 and pT7 have different antibiotic resistance markers, which is necessary to ensure neither is lost by the host, then the ability of pT7 to replicate pOriT7 can be used as a selection for functional T7 replisomes. Copy number of the pOriT7 can be controlled in a number of ways, perhaps most easily by encoding anti-sense RNA for gp5 on pOriT7 (Dias and Stein 2002).

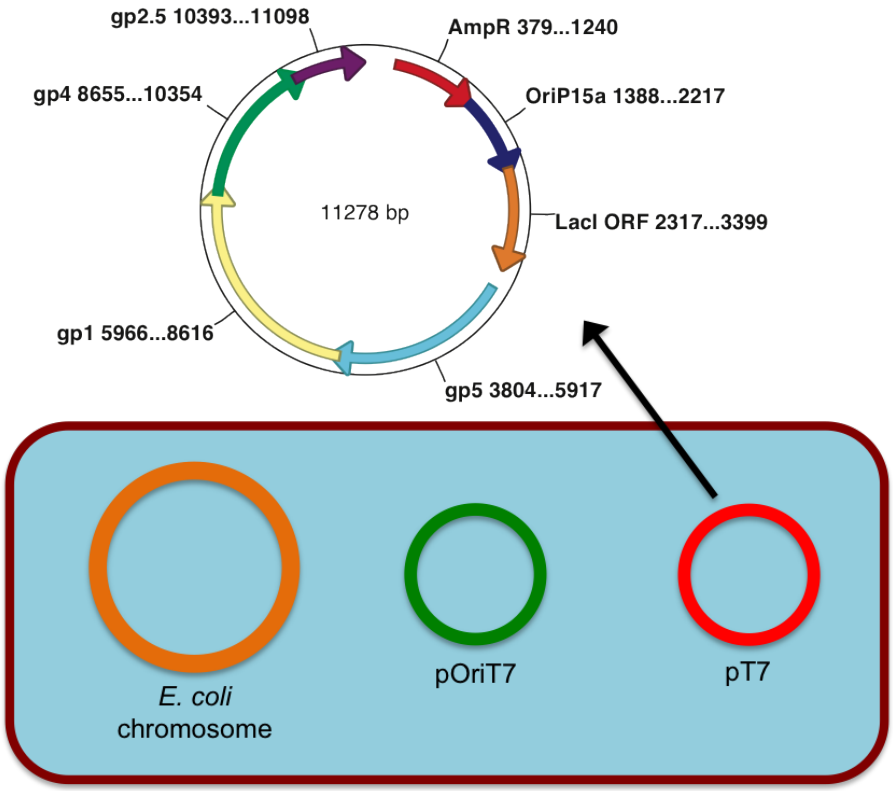

Figure 2.2: (Bottom) Plasmids needed for T7 based orthogonal DNA replication. The *E. coli* host (brown rectangle) contains its own genetic material (orange), pT7 (green) and pOriT7 (red). The pOriT7 contains the genes for error prone replication of pOriT7, which contains the gene(s) of the directed evolution experiment. (Top) A plasmid map of pT7.

2.3 Construction of the T7 expression vector

There are many factors which affect how much of a protein is made in a cell: plasmid copy number (del Solar et al. 1998), promoter strength (Studier and Moffatt 1986), intergenic RNA (Pfleger et al. 2006) and ribosome binding site (RBS) strength (Barrick et al. 1994). I chose the pACYC177 plasmid (NEB) as a starting point to construct the pT7 expression vector (Rose 1988). It has the p15A replicon, which

has a relatively low copy number at 10-12 plasmids per cell. Using standard molecular biology methods I inserted the medium strength *trc* promoter (Amann et al. 1983) and strong rrnB anti-termination region (Li et al. 1984) from the pTrcHis2 vector (Invitrogen). The *trc* promoter is constitutive but regulated by the $LacI^q$ repressor (Calos 1978). To simplify subcloning into this vector, an 800 bp stuffer flanked by SfiI sites was inserted between the promoter and terminator. The resulting plasmid is shown in Fig. 2.3. SfiI is a useful restriction enzyme for making large DNA libraries. Its recognition site (GGCCNNNN'NGGCC) is long, so the enzyme is selective, the three bp sticky-ends it creates can have any sequence and it only cuts two restriction sites at once, so most vectors are either cut completely or left uncut.

To confirm the activity of the plasmid, I subcloned the LacZ gene into the vector and detected its presence using a standard ONPG hydrolysis assay (Strathern 2005). In the presence of the LacZ gene product, colorless ONPG is hydrolyzed into galactose and bright yellow ortho-nitrophenol. The 96-welled plate in Fig. 2.3 shows the results of cells containing the pT7-LacZ in both the absence and presence of IPTG. Each condition was assayed eight times.

Because the expression system is prokaryotic, the entire T7 replisome can be expressed as a single polycistronic mRNA. It has been shown that the insertion of a well chosen library of intergenic RNA can affect protein expression levels by a factor of 100 (Pfleger et al. 2006) and varying the RBS can influence protein expression levels by a factor of 3,000 (Barrick et al. 1994). A RBS library can be encoded into the primers used to PCR the T7 replisome genes, whereas additional

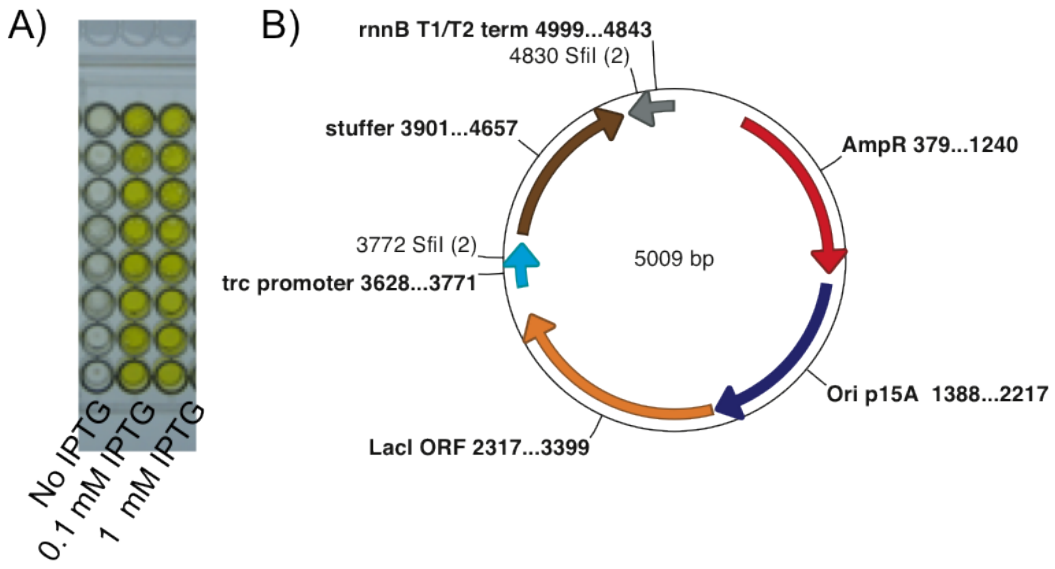

Figure 2.3: The vector built to express the T7 replisome. This expression vector will ultimately become p7, once the replisome is subcloned into the expression site. (A) ONPG hydrolysis assay confirming the inducible promoter was successfully subcloned into the vector. The LacZ, which hydrolyzes colorless ONPG to yellow ONP, was subcloned into the expression site. It is only expressed in the presence of IPTG. Either 0.1 or 1 mM IPTG is sufficient for full induction. Each condition was assayed eight times. (B) Map of the expression vector showing details of the promoter and gene regulatory elements.

intergenic RNA would require additional gene assembly. I opted to attempt to create a library of T7 replisomes with expression levels varied via the RBS.

Each T7 replisome gene was PCRed individually and the RBS library was encoded in the PCR primers. As shown in Fig. A.2, the six bases beginning -7 upstream from a gene's start codon should either be an A or G, and the start codon can be either an A or a G. Primers were constructed to have a 50% probability of being A or G at each of these sites. Primer sequences are listed in Table A.2. With seven binary options for each of four genes, the library contained $(2^7)^4 =$

$\sim 2 \times 10^8$ members. This library was constructed so that each T7 replisome gene could be digested with BglI (NEB). The sticky ends of each gene were designed to be uniquely complementary, so that the genes and the expression vector could be ligated together in a one pot reaction, depicted in Fig. 2.4. This gene assembly strategy has been successfully employed before, but only to create a library with 125 members (Guet et al. 2002).

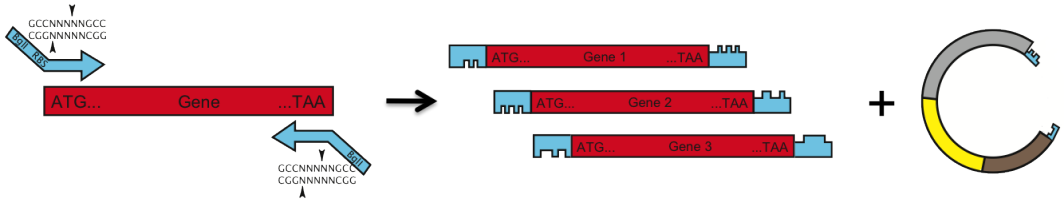

Figure 2.4: Gene assembly strategy. Each primer contains a BglI restriction site. Forward primers contain a RBS library as well. Once the genes are PCRed with these primers, they can be digested with BglI and ligated together in a one pot reaction.

All the genes were PCRed directly from a 500x diluted T7 virus stock, except the gp4A' gene, which was PCRed from a plasmid. Both the virus and plasmid were obtained from F.W. Studier. The genes were PCRed using Vent polymerase (NEB) and standard PCR conditions:

100 μL rxn:
— 85 μL deionized H_2O
— 10 μL 10x ThermolPol buffer (NEB)
— 1 μL Template (1–10 ng of plasmid)
— 1 μL Primer 1 (100 μM)
— 1 μL Primer 2 (100 μM)
— 1 μL dNTPs (should be 200-400 μM for each dNTP)
— 1 μL Vent (NEB)

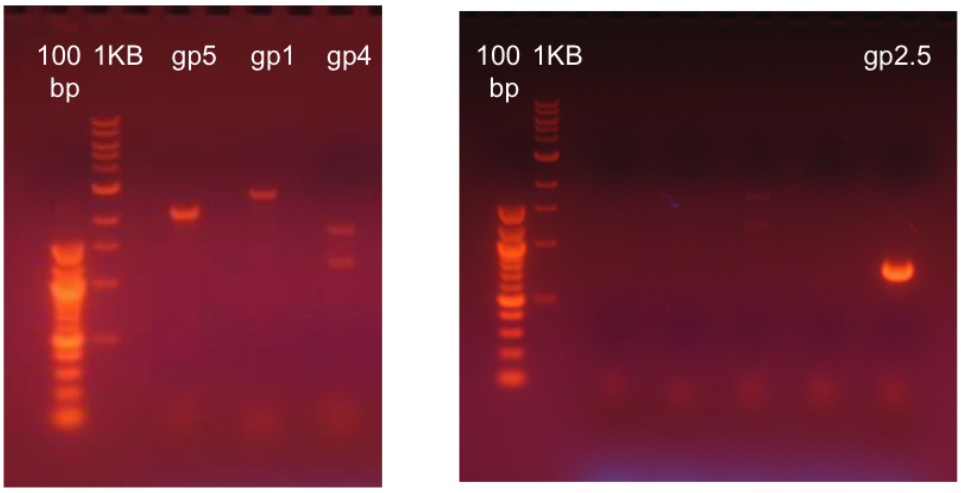

Figure 2.5: DNA gels showing PCR products of the four T7 DNA replisome genes. Gp1, gp2.5, gp4 and gp5 are 2.75, 0.70, 1.70 and 2.12 KB, respectively (see Table A.1). DNA fragment lengths for the 100 bp and 1 KB ladders can be found in Fig. A.3. Empty lanes were from failed PCRs, and should be ignored.

I was able to PCR all four T7 replisome genes using this protocol. The gene products are shown in Fig. 2.5. There was a contaminating band in gp4/4A', which was removed via gel purification. The 4 genes were purified, digested with BglI and ligated together. The ligation reaction was PCR amplified using the outside primers of gp5 and pg2.5 (which should amplify the entire four gene replisome), but no PCR product formed. The ligation followed by PCR amplification was tried numerous times, varying many different parameters: PCR cycling temperatures, PCR cycling times, PCR primers, template concentrations, primer concentrations, dNTP concentrations, DNA polymerases, PCR volume, PCR additives (DMSO, BSA, $MgSO_4$), DNA concentrations during ligation, ligation duration, and ligase concentration. I tried using the FailSafe PCR system (Epicentre Biotechnologies), and I tried constructing the replisome via fusion PCR (Kuwayama et al. 2002), also

varying many of these same parameters, but was unable to construct an operon of the T7 DNA replisome. After several months of unsuccessful plasmid construction, I decided to abandon the project.

2.4 Concluding thoughts

It is unfortunate that I was unable to progress past the gene construction phase of this project. The creation of an orthogonal DNA replication system within a bacteria is an exciting prospect for directed evolution, synthetic biology and the study of DNA replication. Successful completion of this project would have been especially exciting in 2006, when synthetic biology was beginning to take off as a field and I was attempting this experiment. My failure in this project can be attributed to a combination of factors. The most important were likely my inexperience as a molecular biologist, a lack of colleagues with expertise in assembling large DNA fragments (the standard PCR methodologies, which were sufficient for the other projects in the lab, are limited to genes $\leq$ 4 Kb in length (Shevchuk et al. 2004)) and focusing on expression level libraries before I was able to assemble a single T7 replisome.

The technologies for constructing DNA sequences and libraries is rapidly advancing (Baker et al. 2006). Undoubtedly, constructing a gene sequence the size of the T7 replisome will soon be a routine exercise for a molecular biologist. Creating orthogonal DNA replication *in vivo* remains an interesting challenge with many useful scientific applications. I hope the work described here can somehow

be helpful in achieving this goal.

Chapter 3

In vivo Logic

The following chapter is reproduced with minor modifications from: "Transcription factor logic using chemical complementation", by Jonathan E. Bronson, William W. Mazur and Virginia W. Cornish. *Molecular Biosystems* (4):56–58. 2008.

3.1 Abstract

Chemical complementation was used to make a transcription factor circuit capable of performing complex Boolean logic.

3.2 Introduction

Artificial transcription regulation networks are used to quantitatively study biological processes such as quorum sensing, circadian rhythm, cellular memory and biochemical signaling pathways (Chen and Weiss 2005; Elowitz and Leibler 2000;

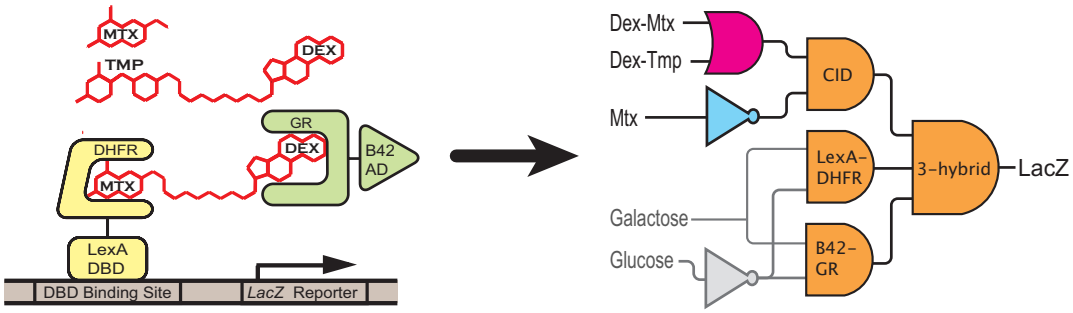

Figure 3.1: (Left) The three-hybrid system. A heterodimeric ligand (Dex-Mtx or Dex-Tmp (red)) bridges a DNA binding protein-receptor protein chimera (LexA-DHFR (yellow)) and a transcriptional activation protein-receptor protein chimera (B42-GR (green)) effectively reconstituting a transcriptional activator and stimulating transcription of a lacZ reporter gene. Transcription can be disrupted by the small molecule Mtx (red). (Right) The three-hybrid system viewed as a three input AND gate. LexA-DHFR and B42-GR are further regulated by the GAL1 promoter, creating a two transcription step circuit. AND, NOT and OR gates directly involved in the three-hybrid logic gate are shown in orange, blue and fuchsia, respectively. Inputs regulating production of three-hybrid components are shown in gray.

Gardner et al. 2000; You et al. 2004). In biotechnology, they are used for the biosynthesis of natural products such as resveratrol and the malaria drug precursor artemisinic acid (Beekwilder et al. 2006; Ro et al. 2006). The creation of "smart cells", which are engineered to perform sophisticated decision making such as the ability to recognize and invade tumor cells (Anderson et al. 2006), is based on artificial transcription networks as well. Often, these designed networks are treated like electrical circuits with transcription factors functioning as Boolean logic gates (Hasty et al. 2002). Multi-input logic functions, such as AND or OR logic, are currently created using combinations of simpler transcription factors, such as LacI, TetR, cI or LuxR (Hasty et al. 2002; Kaern et al. 2003).

This approach to creating logic gates has several drawbacks though. Very

few small molecule inducible transcription factors have been well characterized and shown to be robust and orthogonal enough to the cells genetic machinery to use in artificial networks, so using them in combination quickly limits the size of the networks that can be built. Additionally, regulating promoters with multiple transcription factors can produce unexpected transcription regulation (Setty et al. 2003). These limitations are most pronounced in eukaryotic systems, which are necessary to study many processes pertinent to human development and disease. We offer a solution to these limitations here, using our previously reported dexamethasone methotrexate (Dex-Mtx) yeast three-hybrid system (Lin et al. 2000; Baker et al. 2002), by showing that chemical complementation can be used to create transcription factor logic gates.

In the yeast three-hybrid system, depicted in Fig. Fig. 3.1, a DNA-binding domain (DBD) and an activation domain (AD) of a transcriptional activator are genetically separated and fused to two receptor proteins that bind their respective ligands with high affinity. A heterodimeric small molecule designed to bind the two receptor proteins effectively dimerizes the DBD and AD, reconstituting the transcriptional activator and activating transcription of a downstream reporter gene. This system builds on previous work on n-hybrid systems and chemical dimerizers (Brakmann and Johnsson 2002; Fields and Song 1989; Licitra and Liu 1996; Spencer et al. 1993). For this study, a B42-glucocorticoid receptor chimera (B42-GR) was used as the AD, a LexA-dihydrofolate reductase chimera (LexA-DHFR) as the DBD, Dex-Mtx (Lin et al. 2000) and dexamethasone-trimethoprim (Dex-Tmp) (Gallagher et al. 2007) as the chemical inducers of dimerization (CIDs) and lacZ

as the transcription reporter. The chimeras were made from *E. coli* DHFR and a variant of the hormone-binding domain of the rat GR with two point mutations. Both chimeric proteins were placed under control of the GAL1 promoter. Both small molecules dimerize this three-hybrid system, however, Dex-Tmp has a higher K_D for DHFR than does Dex-Mtx (Benkovic et al. 1988). Although Dex-Mtx and Dex-Tmp both dimerize this three-hybrid system, Mtx and Tmp have significantly different binding affinities for eukaryotic DHFRs and should be functionally distinguishable molecules in other environments (Baccanari et al. 1982). Dimerization of the transcription factor can be disrupted by the presence of 10 mM Mtx without an observable decline in cell viability (Lin et al. 2000).

This three-hybrid system behaves as a three-input Boolean AND gate with LexA-DHFR, B42-GR and Dex-Mtx and/or Dex-Tmp as the inputs. Regulation of the CID is achieved by its presence or absence from the media. To regulate the DBD and AD, we placed both under control of the GAL1 promoter, creating the two transcription step circuit depicted in Fig. 3.1. We evaluated the three-hybrid logic gate in the context of this circuit. Note the GAL1 promoter is only active in the presence of galactose and strongly repressed in the presence of glucose (Strathern 2005). This circuit is capable of processing five bits of information: the presence or absence of glucose, galactose, Dex-Mtx, Dex-Tmp and Mtx in the cellular environment. The 32 entry truth table for this circuit is shown in Fig. 3.2. Only three combinations of inputs, shown in bold in the table, should result in lacZ transcription. The circuit corresponds to the logical expression ((Dex-Mtx OR Dex-Tmp) AND (NOT Mtx)) AND (Gal AND (NOT Glu)).

3.3 Results

The ability of this circuit to perform the expected logical operations was assessed by growing cells containing the circuit in synthetic complete media with 2% raffinose and all 32 combinations of the inputs. Transcription of lacZ was determined using a standard ONPG hydrolysis assay (Strathern 2005). Each condition was tested in quadruplicate. The averaged values and standard deviations are shown in Fig. 3.2. As expected, all combinations of inputs expected to produce a logical 0 showed activity on the order of 10^0 or 10^1 Miller units. All combinations of inputs expected to produce a logical 1 showed activity on the order of 10^3 Miller units. When both Dex-Mtx and Dex-Tmp are present, the circuit shows a slightly weaker output than it does in the presence of just one or the other, however. This is likely due to the inhibitory effect of high concentrations of chemical dimerizers on the three-hybrid system (Lin et al. 2000). These results show the circuit behaves as predicted and the on states and off states are easily distinguishable.

Glu.	Gal.	D-M	D-T	Mtx	Out	Obs.	Stdev
0	0	0	0	0	0	25	39
0	0	0	0	1	0	15	17
0	0	0	1	0	0	15	26
0	0	0	1	1	0	2	66
0	0	1	0	0	0	-10	60
0	0	1	0	1	0	-16	25
0	0	1	1	0	0	7	25
0	0	1	1	1	0	-6	40
0	1	0	0	0	0	39	38
0	1	0	0	1	0	36	10
0	**1**	**0**	**1**	**0**	**1**	**1896**	**218**
0	1	0	1	1	0	8	24
0	**1**	**1**	**0**	**0**	**1**	**1671**	**235**
0	1	1	0	1	0	16	35
0	**1**	**1**	**1**	**0**	**1**	**1315**	**131**
0	1	1	1	1	0	46	85

Glu.	Gal.	D-M	D-T	Mtx	Out	Obs.	Stdev
1	0	0	0	0	0	-3	22
1	0	0	0	1	0	18	23
1	0	0	1	0	0	38	102
1	0	0	1	1	0	-21	30
1	0	1	0	0	0	38	95
1	0	1	0	1	0	34	24
1	0	1	1	0	0	10	30
1	0	1	1	1	0	1	31
1	1	0	0	0	0	23	15
1	1	0	0	1	0	82	35
1	1	0	1	0	0	92	92
1	1	0	1	1	0	33	96
1	1	1	0	0	0	10	10
1	1	1	0	1	0	50	55
1	1	1	1	0	0	27	24
1	1	1	1	1	0	45	68

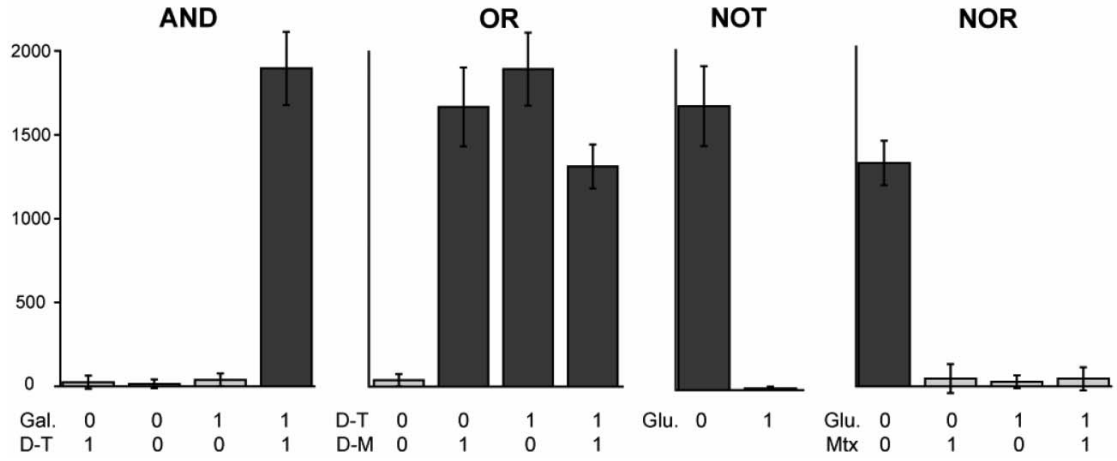

Figure 3.2: (Top) The 32 entry truth table for the three-hybrid genetic circuit. This circuit obeys the logical expression ((Dex-Mtx OR Dex-Tmp) AND (NOT Mtx)) AND (Gal AND (NOT Glu)). The table was split in two for formatting purposes only. A 0 indicates the absence of the input from the media and a 1 indicates the presence of the input in the media. Combinations of inputs that produce a transcription output are shown in bold. The observed outputs (in Miller units), averaged over four trials, and standard deviations of the measurements are shown to the right of the expected outputs. Inputs (left to right) are: 2% glucose, 2% galactose, 1 mM Dex-Mtx, 10 mM Dex-Tmp and 10 mM Mtx. (Bottom) Graphs demonstrating that when several inputs are held constant, the three-hybrid circuit reduces to simpler one and two bit logic functions. On states are shown in dark gray and off state are shown in light gray. Error bars show standard deviation. Glucose, Dex-Mtx and Mtx were all set to 0 for the AND gate. Glucose and Mtx were set to 0 and galactose was set to 1 for the OR gate. Dex-Tmp and Mtx were set to 0 and galactose and Dex-Mtx were set to 1 for the NOT gate. Galactose, Dex-Mtx and Dex-Tmp were set to 1 for the NOR gate.

If simpler logic gates are desired, this three-hybrid circuit can be converted to an AND, OR, NOT or NOR logic gate by holding several of the inputs constant. AND logic is created when glucose and Mtx are off and galactose and either CID are used as inputs. OR logic is created when glucose and Mtx are off, galactose is on and both CIDs are used as inputs. NOR logic is created when galactose and either CID is on and glucose and Mtx are used as inputs. NOT logic is created when galactose is on, Dex-Mtx or Dex-Tmp is on and either glucose or Mtx is used as the input. The outputs of several of these logic gates are shown in Fig. 3.2. YES logic (small molecule or protein inducible transcription) may be produced several ways as well.

3.4 Conclusions

These results demonstrate that chemical complementation can be used to create multiple input transcription factor logic gates. Both complex circuits and simple one or two bit logic gates can be created. The on states and off states of our genetic circuit behaved robustly and with the expected logics. Although not shown here, increasing levels of logical sophistication can be added by having the cell enzymatically modify the chemical inducer of dimerization or by having multiple three-hybrid systems with different DBD-ligand receptor small molecule pairs or different AD-ligand receptor small molecule pairs (Brakmann and Johnsson 2002). Since transcription factors based on chemical complementation are created using known receptor-small molecule pairs and protein chimeras that do not require al-

losteric interactions, it is possible to rapidly generate new, modular transcription factors. The transcription output of one gate can be an input for another, so chemical complementation logic gates are easily connected to each other. All of these features suggest chemical complementation is a useful platform to build artificial transcription factor networks in yeast.

As it becomes possible to create larger transcription factor networks, more complicated cellular decision making will be possible as well. For example, it might be desirable to create a yeast strain that could monitor the conditions in a fermentor, determine whether they were more favorable for producing ethanol or glycerol and turn on/off the appropriate biosynthetic pathways. It would not be possible to make such a strain without creating a sophisticated genetic circuit inside it. The next step in this project will be to to construct three-hybrid NAND gates and characterize our current system in greater depth to further enhance the utility of three-hybrid transcription factors.

3.5 Acknowledgements

We thank Ron Weiss and Milan Stojanovic for their helpful comments while preparing this manuscript. We thank Hening Lin for the synthesis of Dex-Mtx and Sarah Gallagher for the synthesis of Dex-Tmp. This research was supported by the NIH (GM62867-01A1) and NSF (CHE 99-84928).

3.6 Methods

Standard protocols for yeast genetics were used (Strathern 2005). Synthetic defined media were purchased from Qbiogene. ONPG, amino acids, D-raffinose and D-galactose were purchased from Sigma-Aldrich. D-Glucose was purchased from Mallinckrodt Chemicals. Yeast was grown in U-bottomed 96-well plates (VWR) while shaking at 200 rpm in a 30 degree incubator for two days before taking measurements. Spectroscopic measurements were taken with a SpectraMaxPlus 384 spectrophotometer (Molecular Devices). The yeast strain used in this study was the V781Y strain previously described by Baker *et. al.* (Baker et al. 2003). It contains P_{gal1}-LexA-*e*DHFR and 8*lexAop-lacZ* integrated into the chromosome at the *ade*4 and *ura*3 loci, respectively, as well as a 2μ plasmid containing *Pgal*1-*B*42-(*GSG*)2-*rGR*2 and a tryptophan auxotrophy marker. Synthesis of Dex-Mtx is described by Lin *et. al.* (Lin et al. 2000). Synthesis of Dex-Tmp is described in Gallagher *et. al.* (Gallagher et al. 2007).

Part II

smFRET Analysis

Chapter 4

Introduction

4.1 smFRET

When the emission spectrum of a polar chromophore (donor) overlaps with the absorption spectrum of another polar chromophore (acceptor), electromagnetic excitation of the donor can induce a transfer of energy to the acceptor via a non-radiative, dipole-dipole coupling process termed Förster resonance energy transfer (FRET) (Förster 1948). The transfer efficiency between donor and acceptor scales with the distance between molecules (r) as $1/r^6$, with FRET efficiencies most sensitive to r in the range of $1 - 10$nm. Because of this extraordinary sensitivity to distance, FRET efficiency can serve as a molecular ruler, allowing an experimentalist to measure the separation between donor and acceptor by stimulating the donor with light and measuring emission intensities of both the donor (I_D) and acceptor (I_A) (Stryer and Haugland 1967). Usually a summary statistic called the "FRET ratio" (given by FRET $= I_A/(I_D + I_A)$) is used to report on molecular distance

rather than the "raw", 2-channel I_A/I_D data[1]. A summary of the FRET process is shown in Fig. 4.1.

When the donor and acceptor are attached to individual proteins, nucleic acids or other molecular complexes, the FRET signal can be used to report on the dynamics of the molecule to which the donor and acceptor are attached; and when the experiment is crafted to monitor individual molecules rather than ensembles of them, the process is termed single molecule FRET (smFRET). For most biological studies smFRET must be used rather than FRET, since the majority of molecular dynamics cannot be observed from ensemble averages. Often the molecule of interest adopts a series of locally stable conformations during a smFRET time series. From these data, the experimentalist would like to learn (1) the number of locally stable conformations in the data (*i.e.*, states) and (2) the transition rates between states. Although it is theoretically possible use the FRET signal to quantitatively measure the distance between parts of a molecule during a time series, there are usually too many variables affecting FRET efficiency to do this in practice (Schuler et al. 2005). Consequently, smFRET is usually used to extract quantitative information about kinetics (i.e. rate constants) but only qualitative information about distances.

The phenomena of FRET has been studied for over half a century, but the first smFRET experiments were only carried out about fifteen years ago (Ha et al. 1996). The field has been growing exponentially since, and hundreds of smFRET

[1]It is unclear from the literature whether the FRET ratio is used because it a more reliable reporter of donor/acceptor separation than the 2-channel data or if the 2-channel data were merely too hard to analyze when FRET experiments were analyzed by hand and the FRET ratio is a maladaptive statistic which has persisted out of tradition.

papers are published every year now (Joo et al. 2008). Diverse topics such as protein folding (Deniz et al. 2000), RNA structural dynamics (Zhuang et al. 2002) and DNA-protein interactions (Roy et al. 2009) have been investigated via smFRET. The size and complexity of smFRET experiments have grown substantially since the original publication by Ha *et. al.* A modern smFRET experiment can require thousands of time series to be analyzed (Fei et al. 2009). Such large data sets require automated methods to analyze the data and provide a lens of objectivity.

There is a consensus in the smFRET inference field that the data should be modeled with a hidden Markov model (HMM), however, there is debate as to how best to perform inference using the HMM. Chapters 5, 6 and 7 will discuss possible inference methods. The rest of this chapter will provide background information to provide a context for chapters 5, 6 and 7.

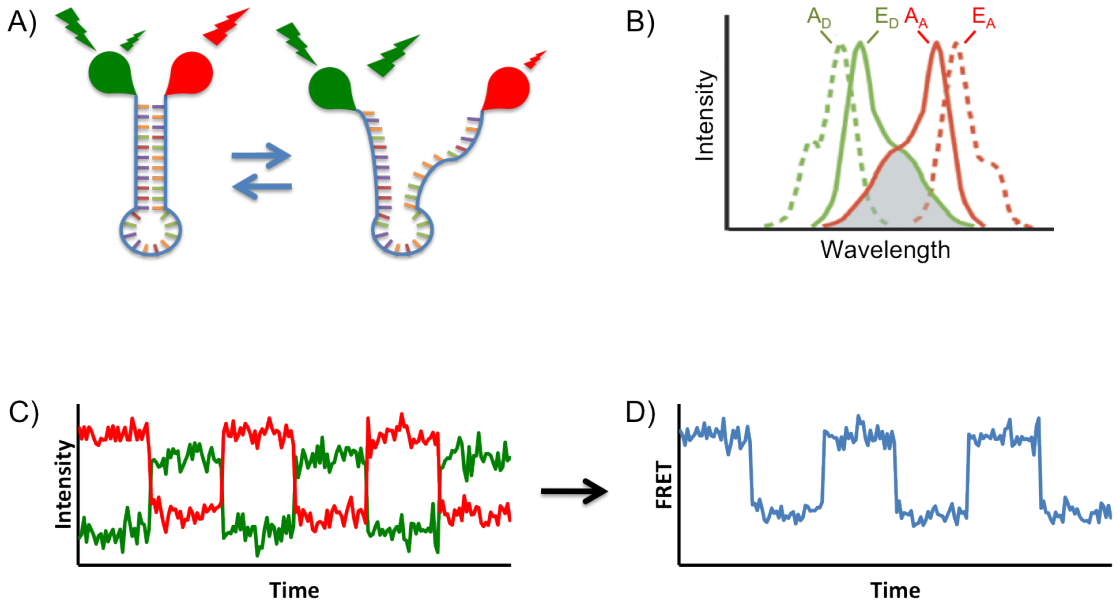

Figure 4.1: **(A)** Cartoon of a smFRET experiment. A DNA hairpin is shown with FRET donor (green balloon) and acceptor (red balloon) choromophores attached. The DNA can adopt two conformations: (1) the zipped hairpin with donor and acceptor near each other (left) and (2) unzipped single stranded DNA with donor and acceptor far apart (right). When the donor is excited with green light in conformation 1, the majority of energy is transferred to the acceptor, causing the donor to fluoresce dimly and the acceptor to fluoresce brightly. In conformation 2, the FRET probes are too far apart for efficient FRET, so when the donor is excited it fluoresces brightly and the acceptor fluoresces dimly. **(B)** The absorption/emission spectrum for a typical donor/acceptor FRET pair (A_D, E_D, A_A and E_A, respectively). Stimulation of the donor with short wavelength light (green for the dye used in this thesis) causes it to fluoresce at a slightly longer wavelength of light. Overlap between E_D and A_A allows efficient FRET. The acceptor fluoresces at an even longer wavelength of light (red for the dye used in this thesis). **(C)** A smFRET time series for the cartoon in A. A CCD camera would be set up to separately record the wavelengths where E_D and E_A are at their maxima. As the DNA transitions between zipped and unzipped, the relative emission intensities of the FRETing dyes switches (more intense red = zipped, more intense green = unzipped), allowing the experimentalist to observe the DNA zipping/unzipping dynamics. **(D)** The 1D FRET transformation of the time series from C. A more intense signal means the FRET pair is closer together. This is the presentation of FRET data which is most commonly analyzed.

4.2 The smFRET time series as a HMM

In early smFRET studies, data were analyzed either "by eye" (Tan et al. 2003), where the experimentalist would assign states to each data point by inspecting the data, or by thresholding, where the experimentalist sets cutoffs between smFRET states (Blanchard et al. 2004a). In 2003, Talaga and and coworkers proposed that a smFRET time series would be well approximated by a hidden Markov model (HMM) (Andrec et al. 2003). The HMM models temporal data using the following assumptions:

1. Time is discrete.

2. At each time step (t) the system is in one of K discrete states.

3. These states cannot be observed (i.e. they are hidden), but at each time step there is a noisy observable which is a function of the current hidden state. The observable can be discrete or continuous.

4. After each time step the system can transition to a new state or remain in its current state. The probability of transitioning is a function of the current state.

5. Both the observed datum at each time step and the probability of transitioning to a new state depend only on the current hidden state of the system. In other words, these probabilities are completely independent of the past, given the current hidden state.

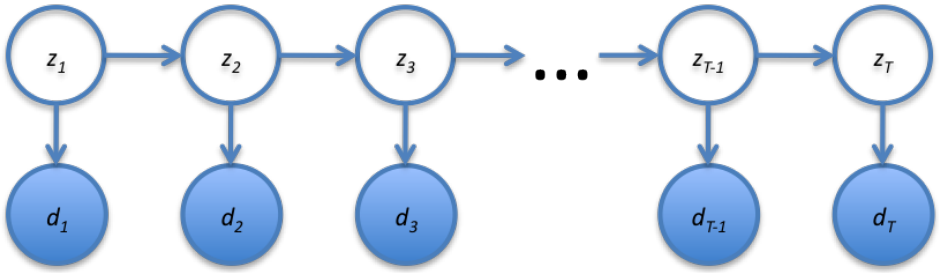

Figure 4.2: Graphical model of a HMM corresponding to Eq. 4.1. Hidden variables (z_t) are shown as empty circles. Observed data (d_t) are shown as filled circles. An arrow from A to B denotes that the probability of B is a function of A. At each time step, the system is in one of K hidden states (i.e. z_t can equal $\{1, 2, \ldots, \mathrm{K}\}$) and produces a noisy observable (d_t), which is a function of the hidden state occupied by z_t. The system can transition after each time step, with transition probabilities that are also functions of the state of z_t.

A graphical model of the HMM is shown in Fig. 4.2. Using the HMM, the modeler wishes to learn the probability of being in each of the K hidden states at each point in time, given the observed data. Often one wishes to know the most probable hidden state trajectory from the data (*i.e.*, the idealized trace). This is known as the Viterbi path of the HMM (Viterbi 1967).

In this thesis, observable data sets will be denoted $\mathbf{D}$, individual data points will be denoted $d_1, d_2, \ldots, d_T$, sets of hidden states will be denoted $\mathbf{Z}$ and individual hidden states will be denoted $z_1,\ z_2, \ldots, z_T$. The parameters of the HMM will be denoted $\vec{\theta}$. For the HMM, the joint probability of $\mathbf{D}$ and $\mathbf{Z}$ is then given by

$$p(\mathbf{D}, \mathbf{Z}|\vec{\theta}, \mathrm{K}) = p(z_1|\vec{\theta}, \mathrm{K}) \left[\prod_{t=2}^{T} p(z_t|z_{t-1}, \vec{\theta}, \mathrm{K})\right] \prod_{t=1}^{T} p(d_t|z_t, \vec{\theta}, \mathrm{K}). \tag{4.1}$$

The probability of $\mathbf{D}$ is found using the sum rule of probability (Eq. B.2):

$$p(\mathbf{D}|\vec{\theta}, \mathrm{K}) = \sum_{\mathbf{Z}} p(\mathbf{D}, \mathbf{Z}|\vec{\theta}, \mathrm{K}). \tag{4.2}$$

Much of what is believed about a smFRET time series for a molecule with a set of locally stable conformations is represented by the HMM. The conformation of the molecule is hidden from the observer. The FRET signal observed is a function of the conformation of the molecule, and one wishes to use the observed data to report on the conformation of the molecule. The probability of transitioning from one molecular conformation to another is a function of the current molecular conformation (i.e. the DNA in Fig. 4.1. is more likely to be zipped at time $t+1$ if it is zipped at time t than if it is unzipped at time t). The CCD camera used to collect smFRET data[2] automatically time bins the data, making a discrete time series.

For smFRET data, each hidden state (molecular conformation) is assumed to give rise to data with Gaussianly distributed noise[3]. The mean and standard deviation of each hidden state (μ_k and σ_k, respectively) are typically different for every state. The values of μ_k and σ_k can be the same for two different states, however, provided that the transition probabilities from those states are different (i.e. if the molecule can somehow switch between fast and slow transitioning conformations). Transition probabilities are modeled as multinomial distributions — at each time step, the molecule throws a weighted die to determine what state it transitions to next. Transition probabilities are stored in a matrix, appropriately called the transition matrix (A). The value of the j^{th} row and k^{th} column of A ($a_{j,k}$) holds the probability of transitioning to the k^{th} hidden state at the next time step given that

[2]FRET can also be observed by other instruments, such as confocal microscopes. However, the majority of smFRET experiments, including all the ones in this thesis, use a CCD camera.

[3]Since FRET data must be between 0 and 1 it cannot actually be Gaussian, but the approximation appears to work well and is generally accepted as a valid approximation.

the system is currently in the j^{th} hidden state (i.e. $a_{j,k} = p(z_{t+1} = k|z_t = j)$). (The probability of not transitioning while in state k is given by $a_{k,k}$.) As an example, in a two state system in which there is a 10% chance of transitioning to state 2 when the system is in state 1 and a 25% chance of transitioning to state 1 when the system is in state 2, A would look like

$$\mathrm{A} = \begin{pmatrix} 0.90 & 0.10 \\ 0.25 & 0.75 \end{pmatrix}.$$

In addition to the transition matrix, there is another $1 \times \mathrm{K}$ vector of parameters, $\vec{\pi}$. Each entry stores the probability that the HMM starts in the k^{th} hidden state. This variable is needed since the first hidden state of the HMM cannot depend on its state at time $t-1$. In practice, $\vec{\pi}$ is unimportant for smFRET analysis (aside from being necessary for the HMM calculations) since it reflects the state the molecule was in the moment the experimentalist started the experiment, rather than any biologically significant quantity.

4.3 A Bayesian primer

4.3.1 Bayes' rule

Bayes' rule (also known as Bayes' theorem and Bayes' law) is an equation which relates a conditional probability, $p(A|B)$, to its inverse, $p(B|A)$. According to Bayes' rule:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}. \tag{4.3}$$

This equation can easily be derived by noting that, according to the product rule of probability (Eq. B.3), $p(A, B)$ is equal to both $p(A|B)p(B)$ and $p(B|A)p(A)$. Equating the two and dividing by p(B) yields Eq. 4.3. Bayes' rule arises in many problems where one wants to know $p(A|B)$, but one only has access to $p(B|A)$. The classic example is the following medical paradox (Ross 2008). Imagine a disease which infects 1% of the population. There is a test for the disease which is 99% accurate (i.e. 99 of 100 people with the disease test positive and 99 of 100 people without the disease test negative). You take the test and the results come back positive, what is the probability you have the disease?

Here we want to know p(disease|test positive) but all we know is p(test positive|disease). Using Bayes' rule, we can find the desired probability.

$$\begin{aligned}
\text{p(disease|test positive)} &= \frac{\text{p(test positive|disease)p(disease)}}{\text{p(test positive)}} \\
&= \frac{\text{p(test positive|disease)p(disease)}}{\text{p(test positive|disease)p(disease)+p(test positive|healthy)p(healthy)}} \\
&= \frac{0.99*0.01}{0.99*0.01+0.01*0.99} \\
&= 0.5
\end{aligned}$$

Despite the test's 99% accuracy rate, you only have a 50% chance of having the disease! This somewhat counterintuitive result can be rationalized when one realizes that the overwhelming majority of people taking the test are healthy and 1 in 100 of these healthy people will produce false positives.

4.3.2 Data modeling with Bayes' rule

Bayes' rule is important for data analysis because often we have some data ($\mathbf{D}$) and a model (m) which we believe describes the data. We want to learn the parameters

($\vec{\theta}$) which describe the model. In the case of smFRET data, $\mathbf{D}$ is the time series, m is a K state HMM with Gaussian observables and $\vec{\theta}$ contains the means and standard deviations of the Gaussians as well as the transition probabilities between states. Since we do not know the values of $\vec{\theta}$, we treat them probabilistically[4] and ask, what is the probability of the parameters, given the data and model, $p(\vec{\theta}|\mathbf{D}, m)$[5]? This is a difficult calculation, since $p(\vec{\theta}|\mathbf{D}, m)$ has no obvious functional form (i.e. there is no obvious way to write a mathematical expression for $p(\vec{\theta}|\mathbf{D}, m)$). The inverse calculation, $p(\mathbf{D}|\vec{\theta}, m)$, is, in general, straightforward though. For example, calculating the probability of flipping {heads, heads, tails} with a coin given that it is weighted to land on heads 70% of the time is much simpler than calculating the probability that the coin is weighted to land on heads 70% of the time given that you observed {heads, heads, tails} in three coin flips[6]. In order to calculate $p(\vec{\theta}|\mathbf{D}, m)$ we turn to Bayes rule,

$$p(\vec{\theta}|\mathbf{D}, m) = \frac{p(\mathbf{D}|\vec{\theta}, m)p(\vec{\theta}|m)}{p(\mathbf{D}|m)}. \tag{4.4}$$

Note that Bayes' rule is still algebraically exact when an extra variable, m, is included as a given for all terms in the equation. Both $p(\mathbf{D}|\vec{\theta}, m)$ and $p(\vec{\theta}|m)$ have functional forms, which are specified by the choice of model. The $p(\mathbf{D}|m)$ can

[4] In the Bayesian approach to statistics, all unknown parameters and variables are assigned probability distributions. A frequentist would take issue with this approach, and argue instead that only a best estimate of parameters should be inferred from the data. This is a large debate in the field of statistics, but is well outside the scope of this work.

[5] This probability is often written as $p(\vec{\theta}|\mathbf{D})$ to avoid clutter, leaving the parameters' dependence on the model implicit. The model dependency is written explicitly here, since much of this work will be about model selection.

[6] The former probability is given by $0.7 * 0.7 * 0.3 = 0.147$. The latter requires Bayes' rule.

be calculated using the sum rule of probability:

$$p(\mathbf{D}|m) = \int d\vec{\theta} p(\mathbf{D}|\vec{\theta}, m) p(\vec{\theta}|m) \tag{4.5}$$

The integral should be replaced with a summation in Eq. 4.5 if $\vec{\theta}$ has discrete variables instead of continuous ones. In theory $p(\mathbf{D}|m)$ can always be calculated using Eq. 4.5. In practice this calculation is often intractable and approximations must be used. The approximation used in this thesis is described in Sec. 4.4.2.

4.3.3 Bayesian terminology

When Bayes' rule is written as Eq. 4.4, the four probabilities are given special names.

$p(\vec{\theta}|m)$ — this term is called the *prior*. It is the only term in Eq. 4.4 that is independent of $\mathbf{D}$, and can be thought of as one's belief about $\vec{\theta}$ prior to seeing data[7]. There are many ways one can set a prior (Van Dongen 2006). They generally fall in to two large categories: either $p(\vec{\theta}|m)$ is set to incorporate one's beliefs about what $\vec{\theta}$ should be, or $p(\vec{\theta}|m)$ is set to be as non-specific as possible. The former approach makes sense when one has seen many similar data sets, and the latter makes sense when one has little belief about $\vec{\theta}$ and does not want the prior to bias the $p(\vec{\theta}|\mathbf{D}, m)$ learned from the data.

[7]The use of priors is the largest point of contention between Bayesians and Frequentists but, again, this debate is outside the scope of this thesis.

$p(\mathbf{D}|\vec{\theta}, m)$ — this term is called the *likelihood*. It is the probability (or likelihood) of the observed data, given the model and a specific choice of model parameters.

$p(\mathbf{D}|m)$ — this term is called the *evidence*. It can be thought of as a normalization constant for the right hand side of Eq. 4.4. When more than one model might be appropriate for a data set, a comparison of the models' evidence can be used to choose which model to use. For reasons described in Sec. 4.3.4 the model with the largest evidence is most likely the best model for the data.

$p(\vec{\theta}|\mathbf{D}, m)$ — this term is called the *posterior*. It is the probability of the model's parameters after (or posterior to) seeing the data.

Bayes' rule can be thought of as a way to incorporate data to update a world view. The modeler starts with a belief about a model's parameters prior to seeing data. The modeler then views the data, through the likelihood and evidence, and develops a belief of the parameters posterior to seeing the data. This posterior belief takes both the prior and the data into account. The relative weighting of the prior and data on the posterior depends on the strength of the prior chosen and the amount of data observed (for more details, see (Bishop 2006)). In the limit of an infinite amount of data, the prior has no effect on the posterior, since an infinite amount of data should outweigh any preconceived notions about model parameters.

4.3.4 Evidence based model selection

When one is presented with competing models to describe a data set and needs to choose between or among them, the model with the highest evidence is most probably correct. There are two arguments for why higher evidence implies a more probable model. The first is Bayesian. The quantity needed for model selection is the probability of the model given the data, $p(m|\mathbf{D})$. For two competing models m_1 and m_2, if we look at the ratio of $p(m_1|\mathbf{D})$ and $p(m_2|\mathbf{D})$ and apply Bayes' rule we get:

$$\frac{p(m_1|\mathbf{D})}{p(m_2|\mathbf{D})} = \frac{\frac{p(\mathbf{D}|m_1)p(m_1)}{p(\mathbf{D})}}{\frac{p(\mathbf{D}|m_2)p(m_2)}{p(\mathbf{D})}} = \frac{p(\mathbf{D}|m_1)p(m_1)}{p(\mathbf{D}|m_2)p(m_2)} \approx \frac{p(\mathbf{D}|m_1)}{p(\mathbf{D}|m_2)} \tag{4.6}$$

It is okay to make the approximation at the end of Eq. 4.6 provided we do not have a strong prior belief about p(m_1) or p(m_2) — in other words, we assume both models are equally likely prior to seeing data. As long as this assumption is valid, the evidence of m_1 can be compared to the evidence of m_2 to determine which model is more probable.

The second argument is a bit more philosophical, and is sometimes referred to as the Occam's razor argument (MacKay 2003). According to this argument, if the model is too simple it will not be able to explain the data, so the evidence will be small. If the model is too complex, then the probability of observing that specific data given the model is also small, because a complex model can account for many data sets making the probability of observing any one of them less likely. There is some evidence that humans naturally evaluate competing models this way (Kemp and Tenenbaum 2008).

To make this argument more concrete, consider the following problem. *There is a random number generator which can generate random numbers between 1 and N. You observe the following sequence of numbers:*

{*1,2,1,1,3,2,3,1,3,3,1,1,1,2,1,3,1,2,3,2,2,3,1,3,2*}

What is the value of N?

Most people will intuitively say N=3. The reasoning is that if N is less than three, the model is too simple to explain the data and, therefore, $p(\mathbf{D}|m)$ is small[8]. Random number generators where N is greater than three are possible, but the probability of observing a data set with no numbers greater than three is very unlikely. For instance, N could be 4, but the probability of observing zero 4s in the 25 number string above would be $0.75^{25} = 0.00075254$, making it unlikely to observe that data given a model with N=4.

4.4 Solving the HMM

We now return to the problem of solving the HMM described in Sec. 4.2. From now on, the only type of model we will be concerned with is a HMM with Gaussian observables, multinomial transition probabilities and K states. Since these models only differ in the number of states, K will be used as shorthand for a model with K states. We want to learn the hidden states occupied by the system during the time series ($\mathbf{Z}$) and the means, standard deviations and transition rates between states ($\vec{\theta}$) from the observed time series. Once $\vec{\theta}$ is known, the most probable $\mathbf{Z}$ is

[8]In this case $p(\mathbf{D}|m)$ is actually 0 because the data is impossible, not just improbable given this model for N< 3. In most real examples a simple model could in theory explain the data, but it would explain it poorly, making $p(\mathbf{D}|m)$ small.

easily calculated using Viterbi's method (Viterbi 1967). Consequently, once we find $\vec{\theta}$ we will known $\mathbf{Z}$ as well. The Bayesian approach to solving this problem would be to assign probability distributions to all the parameters in $\vec{\theta}$ and calculate the posterior, $p(\vec{\theta}|\mathbf{D}, \mathrm{K})$, using Eq. 4.4. Once the posterior is known, a best estimate of the parameters ($\vec{\theta}_*$) can be taken from the posterior's mode:

$$\vec{\theta}_* = \max_{\vec{\theta}} p(\vec{\theta}|\mathbf{D}, \mathrm{K}). \tag{4.7}$$

Moreover, by knowing the full $p(\vec{\theta}|\mathbf{D}, \mathrm{K})$, one also knows the margin of error for this estimate. The more sharply $p(\vec{\theta}|\mathbf{D}, \mathrm{K})$ is peaked around $\vec{\theta}_*$ the better the estimate for $\vec{\theta}_*$. Unfortunately, the expression for the evidence for the HMM,

$$p(\mathbf{D}|\mathrm{K}) = \sum_{\mathbf{Z}} \int d\vec{\theta} p(\vec{\theta}|\mathrm{K}) p(z_1|\vec{\theta}, \mathrm{K}) \left[\prod_{t=2}^{T} p(z_t|z_{t-1}, \vec{\theta}, \mathrm{K})\right] \prod_{t=1}^{T} p(d_t|z_t, \vec{\theta}, \mathrm{K}) \tag{4.8}$$

is intractable for all known choices of $p(\vec{\theta}|\mathrm{K})$, so some form of approximation must be used to estimate $p(\vec{\theta}|\mathbf{D}, \mathrm{K})$.

4.4.1 Maximum likelihood

The maximum likelihood (ML) estimate for $\vec{\theta}_*$ is given by:

$$\vec{\theta}_* = \operatorname*{argmax}_{\vec{\theta}} p(\mathbf{D}|\vec{\theta}, \mathrm{K}). \tag{4.9}$$

There are two arguments for why one would want to take the ML estimate of $\vec{\theta}_*$. The frequentist argument asserts that finding the parameters which make the likelihood of the data greatest should be done as a first principle of statistics. The second argument relies on Bayes' rule. Since

$$p(\vec{\theta}|\mathbf{D}, \mathrm{K}) = \frac{p(\mathbf{D}|\vec{\theta}, \mathrm{K}) p(\vec{\theta}|\mathrm{K})}{p(\mathbf{D}|\mathrm{K})},$$

there is a good chance that the values of $\vec{\theta}$ which maximize $p(\mathbf{D}|\vec{\theta}, \mathrm{K})$ will maximize $p(\vec{\theta}|\mathbf{D}, \mathrm{K})$ as well.

The main advantage of ML is that it is easy to implement. For the HMM, the Expectation Maximization (EM) algorithm (Dempster et al. 1977) can be used to efficiently find the ML estimate of $\vec{\theta}_*$. The EM algorithm starts with a guess for $\vec{\theta}_*$. It uses the guessed $\vec{\theta}_*$ to assign the values of the hidden states, $\mathbf{Z}$. These new values of $\mathbf{Z}$ are then used to calculate new values of $\vec{\theta}_*$. The process iterates until the likelihood converges. Convergence to a local optimum is guaranteed. A local optimum is a set of $\vec{\theta}$ which are better than any other set of $\vec{\theta}$ nearby in parameter space, but are still worse than $\vec{\theta}_*$, which is located in a different region of parameter space. Consequently, the EM algorithm must be run many times with many different initial guesses for $\vec{\theta}_*$ (sometimes called "random restarts"). One is never guaranteed to find the true $\vec{\theta}_*$, but the chances of finding it are much higher if many random restarts are used.

There are two well known problems with ML, both of which are illustrated in Fig. 4.3. The first problem with ML is there is no form of model selection. As K is increased, the likelihood of the fit of data will always monotonically increase. The reason for this is that extra states can be used to fit the noise of the observed data. Since the likelihood is the probability of the observed data given $\vec{\theta}$ and K, adding extra states to fit the noise will always increase the likelihood. This problem is known as overfitting, and is highly undesirable for two reasons. First, because the model is tuning parameters to the observed noise in the data, the $\vec{\theta}_*$ learned from a model which overfits will not accurately reflect the true parameters of the

system being studied. Second, a model which overfits will be a poor predictor of future observed data, since it is not correctly modeling the true dynamics of the system.

To further illustrate the problem of overfitting, consider its most extreme possible case: every data point is assigned to its own state. Such a fit of the data would have the highest likelihood possible, since the observed data is completely accounted for by the model. Such a model has absolutely no explanatory or predictive power however; it is merely a restatement of the data.

In order to properly use ML, it must be combined with some form of model selection, most commonly cross validation or the addition of a penalty term. In cross validation, the observed data is split into two groups: training and testing data. The training data is used to learn $\vec{\theta}_*$. The testing data is then fit with the $\vec{\theta}_*$ learned. The likelihood of the training data will monotonically increase as K is increased, but the likelihood of the testing data will peak for the model of correct complexity (see Fig. 4.3, left panel). Although generally effective, cross validation can be computationally intensive, and prevents the model from learning from all of the data that is collected.

Adding a penalty term to the likelihood, which grows with model complexity, is computationally quicker than cross validation and allows the model to learn from the full data set. The two most common penalty terms are the AIC and BIC (Bishop 2006). The AIC says choose the model for which $\log(p(\mathbf{D}|\vec{\theta}, \mathrm{K})) - \mathcal{M}$ is largest, where $\mathcal{M}$ is the number of free parameters in the model. The BIC says choose the model for which $\log(p(\mathbf{D}|\vec{\theta}, \mathrm{K})) - \frac{1}{2}\mathcal{M}\log(n)$ is the largest, where n is the number

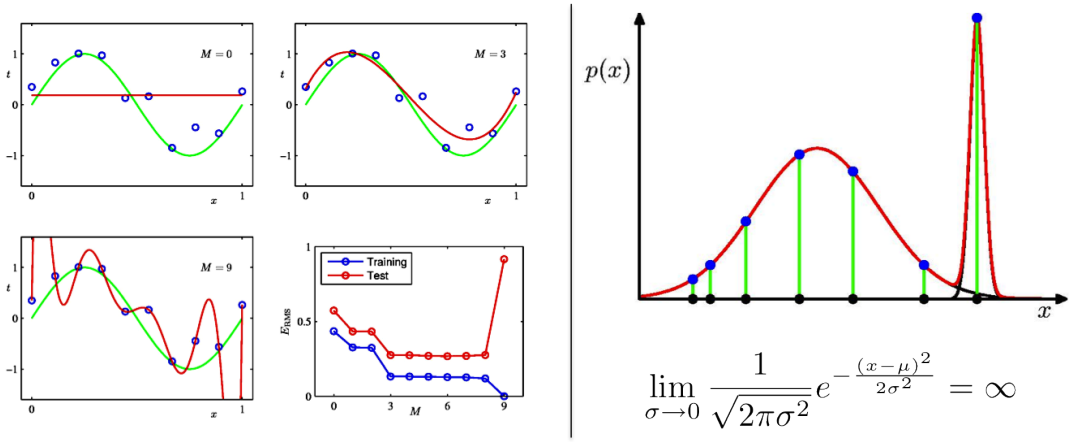

Figure 4.3: Illustration of the two known pathologies ML. **(Left)** the problem of overfitting. The same data set is shown three times. Starting with the lower left subplot and going clockwise: a 0^{th}, 3^{rd} and 9^{th} order polynomial fit of the same data set. The true curve (green) gives rise to noisy observations (blue), which are then fit by the m^{th} degree polynomials (red). Lower right subplot: (blue) the squared error between the model and observed data (training error) for $m = 0 - 9$, and (red) the squared error on new noisy data drawn from the green curve (test data, not shown on subplots $1 - 3$). The testing error is used to determine the correct model in cross validation. As the degree of the polynomial increases, the squared error decreases because the model is better able to account for the noise. (The likelihood increases with m as well, since the model is accounting for more of the observed data). Although it has the lowest training error, the $m = 9$ model is substantially overfit, and the model learned does not resemble the true curve at all. As expected, the testing error is very high for the overfit model. **(Right)** the problem of divergent solutions. A pathological fit of a Gaussian mixture model is shown. One hidden state is assigned to a single data point, giving the associated Gaussian 0 variance and infinite likelihood at the point of that datum. The rest of the fit of the data has no meaning, since the likelihood is already infinite. This is a problem whenever ML is used on a hidden mixture model with continuous observables. Figures reproduced from (Bishop 2006) Figs. 1.4 & 9.7.

of data points and $\mathcal{M}$ is still the number of free model parameters. Penalty terms such as these tend to be coarse, general fixes for the problem of overfitting and often favor a model of the wrong complexity. The derivation of the BIC and conditions under which its use is appropriate are contained in Sec. B.3.

The second problem with ML occurs only in the case of a model with multiple hidden states and a continuous observable (such as the HMM). If one hidden state is assigned to exactly one data point, then the variance of that state will be zero, and the likelihood of that datum will be infinite. This pathology is known as a "divergent solution" (Fig. 4.3, right panel). Since the likelihood of the data will be infinite regardless of how poorly the rest of the data is fit, divergent solutions make meaningful inference impossible. The EM algorithm can be modified to detect divergent solutions and fix them by setting the variance of diverging states to be very large. This correction requires the modeler to impose a subjective criteria to detect divergence though. For some data sets, divergent solutions can be a substantial problem when fit by ML.

4.4.2 Maximum evidence

The principle of maximum evidence (ME), can be thought of as an extension of ML for model selection. Where ML asks which parameters make a given model most probable, ME asks which model makes a given data set most probable (see Sec. 4.3.4 for an explanation of why ME selects the most probable model). Often the evidence is an intractable quantity to calculate (*e.g.* it requires a summation which grows exponentially with the size of the data set) and approximations to the evidence must be used. One such approximation, the variational Bayes expectation maximization algorithm (VBEM), provides an estimate of both the model's evidence and the posterior parameters of the model, allowing the modeler to simultaneously select the model of the correct complexity and fit the data using the model's posterior.

The basis of the VBEM algorithm is explained with the following simple algebraic identity (Bishop 2006). Since Bayesian analysis treats unknown states ($\mathbf{Z}$) and unknown parameters ($\vec{\theta}$) the same way this section will lump them both into $\mathbf{X}$ for notationally simplicity. Let $q(\mathbf{X})$ be any probability distribution which only depends on $\mathbf{X}$.

$$\begin{aligned}
\log p(\mathbf{D}|\mathrm{K}) &= \int q(\mathbf{X}) \log\left(p(\mathbf{D}|\mathrm{K})\right) d\mathbf{X} && (4.10)\\
&= \int q(\mathbf{X}) \log\left(\frac{p(\mathbf{D},\mathbf{X}|\mathrm{K})}{p(\mathbf{X}|\mathbf{D},\mathrm{K})}\right) d\mathbf{X} && (4.11)\\
&= \int q(\mathbf{X}) \log\left(\frac{p(\mathbf{D},\mathbf{X}|\mathrm{K})q(\mathbf{X})}{p(\mathbf{Z}|\mathbf{D},\mathrm{K})q(\mathbf{X})}\right) d\mathbf{X} && (4.12)\\
&= \int q(\mathbf{X}) \log\left(\frac{p(\mathbf{D},\mathbf{X}|\mathrm{K})}{q(\mathbf{X})}\right) d\mathbf{X} \\
&\quad - \int q(\mathbf{X}) \log\left(\frac{p(\mathbf{X}|\mathbf{D},\mathrm{K})}{q(\mathbf{X})}\right) d\mathbf{X} && (4.13)\\
&= \mathcal{L}(q) + D_{KL}(q(\mathbf{Z},\vec{\theta})||p(\mathbf{Z},\vec{\theta}|\mathbf{D},\mathrm{K})) && (4.14)
\end{aligned}$$

Summations over the discrete components of $\mathbf{X}$ should be included in these equations, but were omitted for notational simplicity. The equality in Eq. 4.10 results from the requirement that $q(\mathbf{X})$ be a normalized probability, Eq. 4.11 rewrites $p(\mathbf{D}|\mathrm{K})$ in terms of a conditional probability and Eq. 4.14 reinserts $\{\mathbf{Z},\vec{\theta}\}$ for $\mathbf{X}$ and renames the two terms in Eq. 4.13 as the lower bound of the log(evidence) and Kullback-Leibler divergence, respectively.

Using Jensen's inequality, it can be shown that

$$D_{KL}(q||p) \geq 0, \tag{4.15}$$

with equality when $q = p$ (Bishop 2006). Since D_{KL} is always non-negative,

$$\log\left(p(\mathbf{D}|\mathrm{K})\right) \geq \mathcal{L}(q), \tag{4.16}$$

which proves that $\mathcal{L}(q)$ is a lower bound on the model's log(evidence). Moreover, because of the equality condition for Eq. 4.15, $\mathcal{L}(q)$ is maximized when $q(\mathbf{Z}, \vec{\theta})$ is equal to the posterior distribution for the model's parameters and hidden states (i.e $q(\mathbf{Z}, \vec{\theta})$ is an approximating function for the posterior). Therefore, the optimization simultaneously performs model selection (calculation of $p(\mathbf{D}|\mathrm{K})$) and inference (calculation of $p(\mathbf{Z}, \vec{\theta}|\mathbf{D}, \mathrm{K})$).

The net effect of Eqs. 4.10 – 4.15 is to replace an intractable calculation with a tractable bound optimization problem. The only assumption about $q(\mathbf{X})$ needed to make the optimization tractable is that it factorizes into a function of $\mathbf{Z}$ and a function of $\vec{\theta}$ ($q(\mathbf{X}) = q(\mathbf{Z})q(\vec{\theta})$) (Ji et al. 2006; Bishop 2006). The VBEM optimization is similar to the EM optimization. Instead of iteratively using guesses for $\vec{\theta}_*$ to set $\mathbf{Z}$ and guesses for $\mathbf{Z}$ to set $\vec{\theta}_*$ VBEM iterates between the following update equations:

$$\mathrm{VBE} : q(\mathbf{Z}) = \frac{1}{\mathcal{Z}_{\mathbf{Z}}} \exp\left(\mathbb{E}_{q(\vec{\theta})}\left[\log\left(p(\mathbf{D}, \mathbf{Z}|\vec{\theta}, \mathrm{K})p(\vec{\theta}|\mathrm{K})\right)\right]\right) \tag{4.17}$$

$$\mathrm{VBM} : q(\vec{\theta}) = \frac{1}{\mathcal{Z}_{\vec{\theta}}} \exp\left(\mathbb{E}_{q(\mathbf{Z})}\left[\log\left(p(\mathbf{D}, \mathbf{Z}|\vec{\theta}, \mathrm{K})p(\vec{\theta}|\mathrm{K})\right)\right]\right). \tag{4.18}$$

Here $\mathbb{E}$ denotes the expected value with respect to the subscripted quantity and $\mathcal{Z}$ is a normalization constant. Whereas the $\log(p(\mathbf{D}|\mathrm{K}))$ is a log of a sum/integral, Eqs. 4.17 & 4.18 are both the sum/integral of a log. This difference is what renders $\log(p(\mathbf{D}|\mathrm{K}))$ intractable, but Eqs. 4.17 & 4.18 tractable.

For the HMM used in smFRET analysis, $\vec{\theta}$ comprises four types of variables. The noise of each states is Gaussianly distributed, with mean μ_k and precision

Λ_k[9]. The probability that the time series started in the k^{th} state ($\vec{\pi}$) is modeled as a multinomial distribution. The rows of the transition matrix ($\{a_{i,1}\ldots,a_{i,K}\}$) are modeled as multinomial distributions as well. The prior used for calculations in this thesis, and in other work (Ji et al. 2006), models $p(\mu_k,\Lambda_k)$ as a Gaussian-Gamma distribution,

$$p(\mu_k,\lambda_k)=\sqrt{\frac{u_\beta^k\lambda_k}{2\pi}}e^{-\frac{1}{2}u_\beta^k\lambda_k(\mu_k-u_\mu^k)^2}\frac{1}{\Gamma(u_v^k/2)}(2u_W^k)^{-u_v^k/2}\lambda_k^{(u_v^k/2)-1}e^{-\frac{\lambda_k}{2u_W^k}} \tag{4.19}$$

$p(\vec{\pi})$ as a Dirichlet distribution,

$$p(\vec{\pi})=\frac{\Gamma(\sum_{k=1}^K u_\pi^k)}{\prod_{k=1}^K\Gamma(u_\pi^k)}\prod_{k=1}^K\pi_k^{u_\pi^k-1}, \tag{4.20}$$

and the rows of A as Dirichlet distributions

$$p(a_{j1},...,a_{jK})=\frac{\Gamma(\sum_{k=1}^K u_a^{jk})}{\prod_{k=1}^K\Gamma(u_a^{jk})}\prod_{k=1}^K a_{jk}^{u_a^{jk}-1}. \tag{4.21}$$

Note that setting probability distributions over $\vec{\theta}$ moves the inference problem from finding the parameters of the model to finding the parameters of the probability distributions over the parameters of the model. These parameters are known as "hyperparameters". The terms $\vec{u}_\pi$, $\vec{u}_a$, $\vec{u}_\beta$, $\vec{u}_\mu$, $\vec{u}_v$, and $\vec{u}_W$ (collectively termed $\vec{u}$) in Eqs. 4.19−4.21 are the hyperparameters for the model of the HMM used in this thesis. For this model

$$\begin{aligned}p(\mathbf{D},\mathbf{Z}|\vec{\theta})p(\vec{\theta}) &= p(z_1|\vec{\pi})\left[\prod_{t=2}^T p(z_t|z_{t-1},A)\right]\prod_{t=1}^T p(d_t|z_t,\vec{\mu},\vec{\lambda})\times \\ &\quad p(\vec{\pi}|\vec{u}_\pi)p(A|\vec{u}_a)p(\vec{\mu}|\vec{u}_\mu,\vec{u}_\beta,\vec{\lambda})p(\vec{\lambda}|\vec{u}_v,\vec{u}_W)\end{aligned} \tag{4.22}$$

where the dependence on K has been omitted for clarity. The full graphical model for this HMM is shown in Fig. 4.4.

[9] Λ_k is the inverse of covariance ($\Lambda=1/\sigma^2$). Using Λ instead of σ simplifies some of the algebra of the VBEM equation.

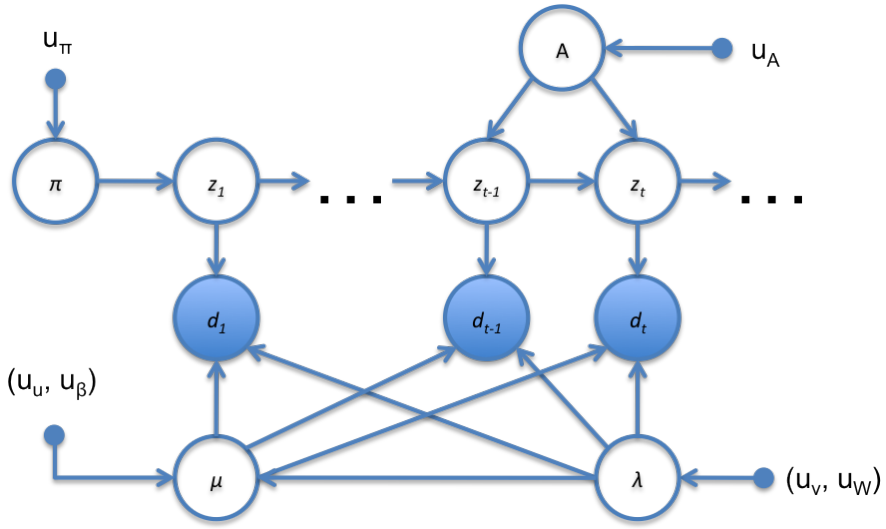

Figure 4.4: Graphical model representation of the HMM, corresponding to the factorization of the probability distribution in Eq. 4.22. Each vertical slice represents a time slice $t = 1, \ldots, T$, for which there is an observed FRET ratio d_t, given a hidden conformational state $z_t \in 1, \ldots, K$. Transitions between conformational states are represented by the dependencies between z_t and z_{t-1}. Parameters are also modeled as random variables, with arrows indicating the dependence of the observed (shaded) and hidden (unshaded) variables. Hyperparameters are shown as small solid circles.

In summary, the goal of variational Bayes is to simultaneously calculate a model's evidence, $p(\mathbf{D}|\mathrm{K})$, and the posterior parameter distribution for the model, $p(\vec{\theta}|\mathbf{D}, \mathrm{K})$. The former allows one to perform model selection and the latter allows one to model the observed data. The VBEM algorithm, implemented via in Eqs. 4.17 & 4.18, is designed to find the set of hyperparameters, $\vec{u}$, which parameterize the posterior, $p(\vec{\theta}|\mathbf{D}, \mathrm{K})$, given a choice of $\vec{u}$ for the prior and given the observed data.

4.4.3 Advantages of ME over ML

There are several reason which the ME method described in Sec. 4.4.2 is better than the ML method described in Sec. 4.4.1.

- ME naturally provides model selection, ML does not. Model selection with ML requires cross validation (time & data intensive) or the use of penalty terms (inaccurate).

- ME does not suffer from the problem of divergent solutions. ML is a point estimate of $\vec{\theta}$, which allows the EM algorithm to converge to solutions with zero variance and infinite likelihood. By looking at both $p(\vec{\theta}|\mathrm{K})$ and $p(\mathbf{D}|\vec{\theta},\mathrm{K})$, ME prevents divergent solutions. The reason for this is that the evidence calculation is only interested in distributions over $\vec{\theta}$. Only for point estimates of $\vec{\theta}$ can $p(\mathbf{D}|\vec{\theta},\mathrm{K}) = \infty$ and, like any other continuous distribution, $p(\vec{\theta}|\mathrm{K}) = 0$ for all point estimates of $\vec{\theta}$ so $p(\mathbf{D}|\vec{\theta},\mathrm{K})p(\vec{\theta}|\mathrm{K})$ does not contribute to the evidence at these divergent points.

- By returning the full $p(\vec{\theta}|\mathbf{D},\mathrm{K})$, ME provides a simple mechanism both to extract an estimate of $\vec{\theta}_*$ and an estimate for the error on $\vec{\theta}_*$. To estimate the error on $\vec{\theta}_*$ using ML requires splitting the data set up into many independent segments, calculating $\vec{\theta}_*$ for each one and using the variance on the $\vec{\theta}_*$ calculated to estimate error bars.

- When more states are fit to the data than are supported by the data, ME will leave some states unpopulated, but ML will usually populate all available

states. The reasons ME does not populate these states are somewhat abstract (Bishop 2006, Ch. 3). Essentially, the posterior calculated in ME can be thought of as the prior plus modifications to the prior made as a result of seeing the data. If there is no data to support populating a state, then the posterior of that state will be identical to the prior (which is the same thing as saying the state was unpopulated in the posterior).

4.4.4 Other estimation methods

It should be noted that there are other inference methods aside from ML and the version of ME discussed here. The most notable are *maximum a posteriori* (MAP) estimates (Gauvain and Lee 1994) and Monte Carlo (MC) techniques (Neal 1993). MAP is similar to ML, but seeks to find the argmax$_{\vec{\theta}}$ of $p(\mathbf{D}|\vec{\theta},\mathrm{K})p(\vec{\theta}|\mathrm{K})$, rather than the argmax$_{\vec{\theta}}$ of $p(\mathbf{D}|\vec{\theta},\mathrm{K})$. MAP avoids the divergent solutions problem of ML. It suffers from a lack of intrinsic model selection and is still a point estimate of the posterior, making it a less desirable inference method than ME. MC uses computer simulations to calculate a model's evidence and/or posterior. It can be very accurate given an infinite amount of computing time, but is too computationally intensive to be of much use for smFRET inference in practice.

4.5 Current methods

There are currently two software packages commonly used for smFRET data analysis: QUB (Qin et al. 1997; 2000)and HaMMy (McKinney et al. 2006). Both

programs model the smFRET time series using a HMM and both programs solve the HMM via ML. Consequently, both programs suffer from the known problems of ML, discussed above. QUB was originally created to analyze ion channel data and HaMMy was created specifically for smFRET analysis.

The next two chapters of this thesis present an alternative approach to smFRET analysis, based on ME. I developed an open source MATLAB software package, termed vbFRET, as well. A screenshot of the graphical user interface (GUI) is shown in Fig. 4.5. The code is available for download at http://vbfret.sourceforge.net/.

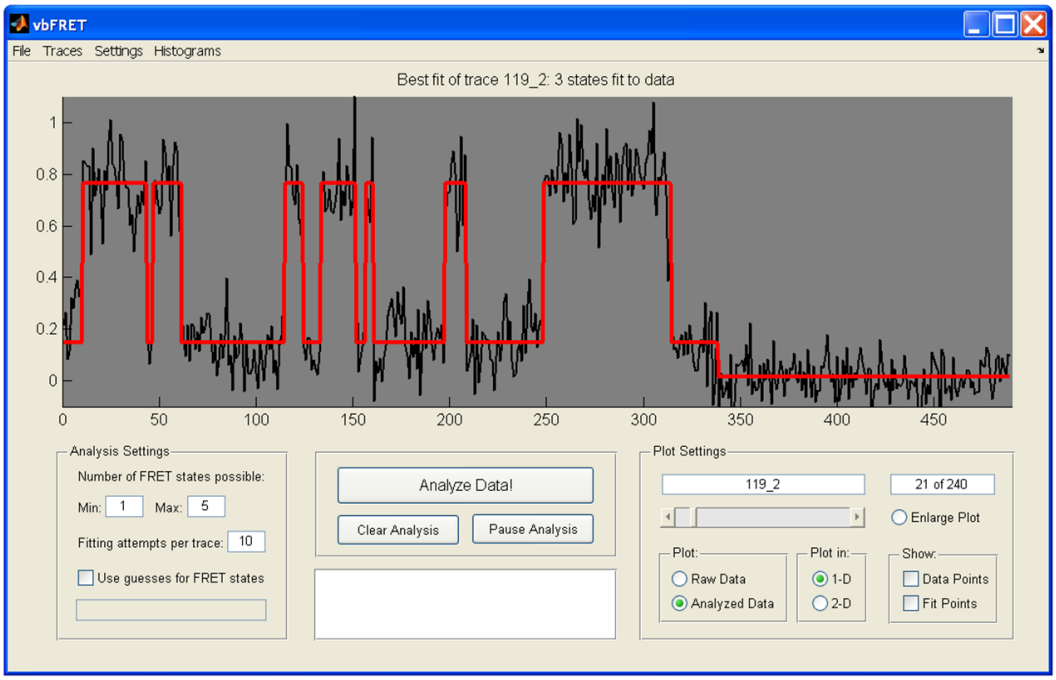

Figure 4.5: The vbFRET GUI

Chapter 5

vbFRET

The following chapter is reproduced with minor modifications from: "Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data", by Jonathan E. Bronson, Jingyi Fei, Jake M. Hofman, Ruben L. Gonzalez Jr., and Chris H. Wiggins. *Biophysical Journal* (97):3196–3205. 2009.

In addition, the inference algorithm described here was used to analyze the data in: "Allosteric collaboration between elongation factor G and the ribosomal L1 stalk direct tRNA movements during translation", by Jingyi Fei, Jonathan E. Bronson, Jake M. Hofman, Rathi L. Srinivas, Chris H. Wiggins and Ruben L. Gonzalez, Jr. *Proceedings of the National Academy of Sciences* (106):15702-15707. 2009. The process of analyzing this data was extremely helpful in refining the inference algorithm.

5.1 Abstract

Time series data provided by single-molecule Förster resonance energy transfer (smFRET) experiments offer the opportunity to infer not only model parameters describing molecular complexes, *e.g.* rate constants, but also information about the model itself, *e.g.* the number of conformational states. Resolving whether or how many of such states exist requires a careful approach to the problem of *model selection*, here meaning discriminating among models with differing numbers of states. The most straightforward approach to model selection generalizes the common idea of maximum likelihood — selecting the *most likely parameter values* — to maximum evidence: selecting the *most likely model*. In either case, such inference presents a tremendous computational challenge, which we here address by exploiting an approximation technique termed *variational Bayesian expectation maximization*. We demonstrate how this technique can be applied to temporal data such as smFRET time series; show superior statistical consistency relative to the maximum likelihood approach; compare its performance on smFRET data generated from experiments on the ribosome; and illustrate how model selection in such probabilistic or generative modeling can facilitate analysis of closely related temporal data currently prevalent in biophysics.

5.2 Introduction

Single-molecule biology has triumphed at creating well-defined experiments to analyze the workings of biological materials, molecules, and enzymatic complexes. As

the molecular machinery studied become more complex, so too do the biological questions asked and, necessarily, the statistical tools needed to answer these questions from the resulting experimental data. In a number of recent experiments, researchers have attempted to infer mechanical parameters (*e.g.*, the typical step size of a motor protein), probabilistic parameters (*e.g.*, the probability per turn that a topoisomerase releases from its DNA substrate), or kinetic parameters (*e.g.*, the folding/unfolding rates of a ribozyme) via statistical inference (Koster et al. 2006; Moffitt et al. 2009; Munro et al. 2007; Zhuang et al. 2000; 2002; Fei et al. 2008; Yildiz et al. 2004; Wiita et al. 2007; Myong et al. 2005). Often the question of interest is not only one of selecting model parameters but also selecting the model, including from among models which differ in the number of parameters to be inferred from experimental data. The most straightforward approach to model selection generalizes the common idea of maximum likelihood (ML) — selecting the *most likely parameter values* — to maximum evidence (ME): selecting the *most likely model*.

Here we focus on model selection in a specific example of such a biological challenge: revealing the number of enzymatic conformational states in single molecule FRET (smFRET) data. FRET (Jares-Erijman and Jovin 2003; Joo et al. 2008; Roy et al. 2008; Schuler and Eaton 2008) refers to the transfer of energy from a donor fluorophore (which has been excited by short-wavelength light) to an acceptor fluorophore (which then emits light of a longer wavelength) with an efficiency which decreases as the distance between the fluorophores increases. The distance-dependence of the energy transfer efficiency implies that the quantification of the

light emitted at both wavelengths from a fluorophore pair may be used as a proxy for the actual distance (typically ∼1–10 nm) between these fluorophores. Often a scalar summary statistic (*e.g.* the "FRET ratio" $I_A/(I_A+I_D)$ of the acceptor intensity to the sum of the acceptor and donor intensities) is analyzed as a function of time, yielding time series data which are determined by the geometric relationship between the two fluorophores in a non-trivial way. When the donor and acceptor are biochemically attached to a single molecular complex, one may reasonably interpret such a time series as deriving from the underlying conformational dynamics of the complex.

If the complex of interest transitions from one locally stable conformation to another, the experiment is well modeled by a hidden Markov model (HMM) (Rabiner 1989), a probabilistic model in which an observed time series (here, the FRET ratio) is conditionally dependent on a hidden, or unobserved, discrete state variable (here, the molecular conformation). HMMs have long been used in ion channel experiments in which the observed dynamic variable is voltage, and the hidden variable represents whether the channel is open or closed (Qin et al. 1997; 2000). More recently, Talaga proposed adapting such modeling for FRET data (Andrec et al. 2003), and Ha and coworkers developed HMM software designed for FRET analysis (McKinney et al. 2006). Such existing software for biophysical time series analysis implement ML on individual traces and require users either to guess the number of states present in the data, or to overfit the data intentionally by asserting an excess number of states. Resulting errors commonly are then corrected via heuristics particular to each software package. It would be advantageous to

avoid subjectivity (as well as extra effort) on the part of the experimentalist necessary in introducing thresholds or other parameterized penalties for complex models, as well as to derive principled approaches likely to generalize to new experimental contexts and data types. To that end, our aim here is to implement ME directly, avoiding overfitting even within the analysis of each individual trace rather than as a post-processing correction.

We begin by describing the general problem of using probabilistic or *generative* models for experimental data (generically denoted $\mathbf{D}$) in which one specifies the probability of the data given a set of parameters of biophysical interest (denoted $\vec{\theta}$) and possibly some hidden value of the state variable of interest (denoted $\mathbf{Z}$). We then present one particular framework, variational Bayesian expectation maximization (VBEM), for estimating these parameters while at the same time finding the optimal number of values for the hidden state variable $\mathbf{Z}$. (Bold variables are reserved for those extensive in the number of observations.) We next validate the approach on synthetic data generated by an HMM, with parameters chosen to simulate data comparable to experimental smFRET data of interest. Having validated the technique, we apply it to experimental smFRET data and interpret our results. We close by highlighting advantages of the approach; suggesting related biophysical time series data which might be amenable to such analysis; and outlining promising avenues for future extension and developments of our analysis.

5.3 Parameter and model selection

Since the techniques we present here are natural generalizations of those which form the common introduction to statistical techniques in a broad variety of natural sciences, we first remind the reader of a few key ideas in inference necessary before narrowing to the description of smFRET data, briefly discussing ML methods for parameter inference and ME methods for model selection. Note that, since the ML-ME discussion does not rely on whether or not the model features hidden variables, for the sake of simplicity we first describe in the context of models without hidden variables.

5.3.1 Maximum likelihood inference

The context in which most natural scientists encounter statistical inference is that of ML; in this problem setting, the model is specified by an expression for the *likelihood* $p(\mathbf{D}|\vec{\theta}, \mathrm{K})$ — *i.e.*, the probability of the vector of data $\mathbf{D}$ given some unknown vector of parameters of interest $\vec{\theta}$. (While this is not often stated explicitly, this is the framework underlying minimization of χ^2 or sums-of-squared errors; *cf.* Sec. B.2 for a less cursory discussion.) In this context the ML estimate of the parameter $\vec{\theta}$ is

$$\vec{\theta}_* = \underset{\vec{\theta}}{\operatorname{argmax}}\, p(\mathbf{D}|\vec{\theta}, \mathrm{K}). \tag{5.1}$$

ML methods are useful for inference of parameter settings under a fixed model (or model complexity), *e.g.* a particular parameterized form with a fixed number of parameters. However, when one would like to compare competing models (in ad-

dition to estimating parameter settings), ML methods are generally inappropriate, as they tend to "overfit", because likelihood always increases with greater model complexity.

This problem is conceptually illustrated in the case of inference from FRET data as follows: if a particular system has a *known* number of conformational states, say $K = 2$, one can estimate the parameters (the transition rates between states and relative occupation of states per unit time) by maximizing the likelihood, which gives a formal measure of the "goodness of fit" of the model to the data. Consider, however, an overly complex model for the same observed data with $K = 3$ conformational states, which one might do if the number of states is itself unknown. The resulting parameter estimates will have a higher likelihood or "better" fit to the data under the maximum likelihood criterion, as the additional parameters have provided more degrees of freedom with which to fit the data. The difficulty here is that maximizing the likelihood fails to accurately quantify the desired notion of a "good fit" which should agree with past observations, generalize to future ones and model the underlying dynamics of the system. Indeed, consider the pathological limit in which the number of states K is set equal to the number of FRET time points observed. The model will exactly match the observed FRET trace, but will generalize poorly to future observations. It will have failed to model the data at all and nothing will have been learned about the true nature of the system; the parameter settings will simply be a restatement of observations.

The difficulty in the above example is that one is permitted both to select the model complexity (the number of parameters in the above example) and to estimate

single "best" parameter settings, which results in overfitting. While there are several suggested solutions to this problem (reviewed in (Bishop 2006; MacKay 2003)), we present here a Bayesian solution for modeling FRET data which is both theoretically principled and practically effective (Sec. 5.3.2). In this approach, one extends the concepts behind maximum likelihood to that of maximum *marginal* likelihood, or *evidence*, which results in an alternative quantitative measure of "goodness of fit" that explicitly penalizes overfitting and enables one to perform model selection. The key conceptual insight behind this approach is that one is prohibited from selecting single "best" parameter settings for models considered, and rather maintains probability distributions over *all* parameter settings.

5.3.2 Maximum evidence inference

The ML framework generalizes readily to the problem of choosing among different models. This includes not only models of different algebraic forms, but also among *nested* models in which one model is a parametric limit of another, *e.g.* models with hidden variables or in polynomial regression. (A two state model is a special case of a three state model with an empty state; a second order polynomial is a special case of a third order polynomial with one coefficient set to 0.) In this case we introduce an index K over possible models, *e.g.*, the order of the polynomial to be fit or, here, the number of conformational states, and hope to find the value of K_* which maximizes the probability of the data given the model, $p(\mathbf{D}|\mathrm{K})$:

$$\mathrm{K}_* = \underset{K}{\operatorname{argmax}}\, p(\mathbf{D}|\mathrm{K}) = \underset{K}{\operatorname{argmax}} \int d\vec{\theta} p(\mathbf{D}|\vec{\theta}, \mathrm{K}) p(\vec{\theta}|\mathrm{K}). \tag{5.2}$$

The quantity $p(\mathbf{D}|\mathrm{K})$ is referred to as the *marginal likelihood*, or *evidence*, as unknown parameters are marginalized (or summed out) over all possible settings. The second expression in Eq. 5.2 follows readily from the rules of probability provided we are willing to model the parameters themselves (in addition to the data) as random variables. That is, we must be willing to prescribe a distribution $p(\vec{\theta}|\mathrm{K})$ from which the parameters are drawn given one choice of the model. Since this term is independent of the data $\mathbf{D}$, it is sometimes referred to as the *prior*; the treatment of parameters as random variables is one of the distinguishing features of Bayesian statistics. (In fact, maximizing the evidence is the principle behind the oft-used Bayesian information criterion (BIC), an asymptotic approximation valid under a restricted set of circumstances. The BIC is explored more thoroughly in Sec. B.3.)

In this form we may interpret the marginal likelihood $p(\mathbf{D}|\mathrm{K})$ as an averaged version of the likelihood $p(\mathbf{D}|\vec{\theta},\mathrm{K})$ over all possible parameter values, where the prior $p(\vec{\theta}|\mathrm{K})$ weights each such value. Unlike the likelihood, the evidence is largest for the model of correct complexity and decreases for models that are either too simple or too complex without the need for any additional penalty terms. There are several explanations for why evidence can be used for model selection (Bishop 2006). Perhaps the most intuitive is to think of the evidence as the probability that the observed data was generated using the given model (which we are allowed to do, since ME is a form of generative modeling). Overly simplistic models cannot generate the observed data and, therefore, have low evidence scores (*e.g.* it is improbable that a two FRET state model would generate data with three distinct

FRET states). Overly complex models can describe the observed data, however, they can generate so many different data sets that the specific observed data set becomes improbable (*e.g.* it is improbable that a 100 FRET state model would generate data that only has 3 distinct FRET states (especially when one considers that the evidence is an average taken over all possible parameter values)).

In addition to performing model selection, we would like to make inferences about model parameters, described by the probability distribution over parameter settings given the observed data, $p(\vec{\theta}|\mathbf{D}, \mathrm{K})$, termed the *posterior* distribution. Bayes' rule equates the posterior with the product of the likelihood and the prior, normalized by the evidence:

$$p(\vec{\theta}|\mathbf{D}, \mathrm{K}) = \frac{p(\mathbf{D}|\vec{\theta}, \mathrm{K})p(\vec{\theta}|\mathrm{K})}{p(\mathbf{D}|\mathrm{K})}. \tag{5.3}$$

While ME above does not give us access to the posterior directly, as we show below, VBEM gives not only an approximation to the evidence but also an approximation to the posterior.

5.3.3 Variational approximate inference

While in principle calculation of the evidence and posterior completely specifies the ME approach to model selection, in practice exact computation of the evidence is often both analytically and numerically intractable. One broad and intractable class is that arising from models in which observed data are modeled as conditionally dependent on an unknown or hidden state to be inferred; these *hidden variables* must be marginalized over (summed over) in calculating the evidence in Eq. 5.2.

(For the smFRET data considered here, these hidden variables represent the unobservable conformational states.) As a result, calculation of the evidence now involves a discrete sum over all states $\mathbf{Z}$ in addition to the integrals over parameter values $\vec{\theta}$:

$$p(\mathbf{D}|\mathrm{K}) = \sum_{\mathbf{Z}} \int d\vec{\theta} p(\mathbf{D}, \mathbf{Z}|\vec{\theta}, \mathrm{K}) p(\vec{\theta}|\mathrm{K}) \tag{5.4}$$

This significantly complicates the tasks of model selection and posterior inference. Computing the terms in Eq. 5.2 and Eq. 5.3 requires calculation of the evidence, direct evaluation of which requires a sum over all K settings for each of T extensive variables $\mathbf{Z}$ (where T is the length of the time series). Such a sum is intractable for even $K = 2$ and modest values of T, *e.g.* on the order of 25. While there exist various methods for numerically approximating such sums, such as Monte Carlo techniques, we appeal here to variational methods for a scalable, robust, and empirically accurate method for approximate Bayesian inference. (For a discussion regarding practical aspects of implementing Monte Carlo techniques, including burn-in, convergence rates, and scaling, *cf.* (Neal 1993).)

To motivate the variational method, we note that we wish not only to select the model by determining K_* but also to find the posterior probability distribution for the parameters given the data, *i.e.*, $p(\vec{\theta}|\mathbf{D}, \mathrm{K})$. This is done by finding the distribution $q(\mathbf{Z}, \vec{\theta})$ which best approximates $p(\mathbf{Z}, \vec{\theta}|\mathbf{D}, \mathrm{K})$, *i.e.*,

$$q_*(\mathbf{Z}, \vec{\theta}) = \underset{q(\mathbf{Z}, \vec{\theta})}{\operatorname{argmin}} D_{KL}\left(q(\mathbf{Z}, \vec{\theta}) || p(\mathbf{Z}, \vec{\theta}|\mathbf{D}, \mathrm{K})\right), \tag{5.5}$$

where D_{KL} is the usual Kullback-Leibler divergence, which quantifies the dissimilarity between two probability distributions. A simple identity (derived in Sec. 6.6)

relates this quantity to the evidence $p(\mathbf{D}|\mathrm{K})$:

$$\log p(\mathbf{D}|\mathrm{K}) = -F[q(\mathbf{Z},\vec{\theta})] + D_{KL}\left(q(\mathbf{Z},\vec{\theta})||p(\mathbf{Z},\vec{\theta}|\mathbf{D},\mathrm{K})\right) \geq -F[q(\mathbf{Z},\vec{\theta})] \quad (5.6)$$

where $F[q(\mathbf{Z},\vec{\theta})]$ is an analytically tractable functional (owing to a simple choice of the approximating distribution $q(\mathbf{Z},\vec{\theta})$). The inequality in Eq. 5.6 results from the property $D_{KL} \geq 0$, with equality if and only if $q(\mathbf{Z},\vec{\theta}) = p(\mathbf{Z},\vec{\theta}|\mathbf{D},\mathrm{K})$. Mathematically, Eq. 5.6 illustrates that minimizing the functional $F[q(\mathbf{Z},\vec{\theta})]$ simultaneously maximizes a lower bound on the evidence and minimizes the dissimilarity between the test distribution q and the parameter posterior distribution. Qualitatively, the best test distribution not only gives the best estimate of the evidence but also the best estimate of the posterior distribution of the parameters themselves. In going from Eq. 5.4 to Eq. 5.6, we have replaced the problem of an intractable summation with that of bound optimization. As is commonly the case in bound optimizations, the closeness of this bound to the true evidence cannot be calculated. The validity of the approximation must be tested on synthetic data (tests we perform in Sec. 5.5).

Calculation of F is made tractable by choosing an approximating distribution q with conditional independence among variables which are coupled in the model given by p; for this reason the resulting technique generalizes mean field theory of statistical mechanics (MacKay 2003). Just as in mean field theory, the variational method is defined by iterative update equations; here the update equations result from setting the derivative of F with respect to each of the factors in the approximating distribution q to 0. This procedure for calculating evidence is known as VBEM, and can be thought of as a special case of the more general expectation

maximization algorithm (EM). (We encourage the reader to enjoy the text (Bishop 2006) for a more pedagogical discussion of EM and VBEM.) Since F is convex in each of these factors, the algorithm provably converges to a local (though not necessarily global) optimum, and multiple restarts are typically employed. Note that this is true for EM procedures more generally, including as employed to maximize likelihood in models with hidden variables (*e.g.*, HMMs). In ML inference, practitioners on occasion use the converged result based on one judiciously chosen initial condition rather than choosing the optimum over restarts; this heuristic often prevents pathological solutions (*cf.* (Bishop 2006, Ch. 9)).

5.4 Statistical inference and FRET

5.4.1 Hidden Markov modeling

The HMM (Rabiner 1989), illustrated in Fig. 4.2, models the dynamics of an observed time series $\mathbf{D}$ (here, the observed FRET ratio) as conditionally dependent on a hidden process $\mathbf{Z}$ (here, the unknown conformational state of the molecular complex). At each time t, the conformational state z_t can take on any one of K possible values, conditionally dependent only on its value at the previous time via the transition probability matrix $p(z_t|z_{t-1})$ (*i.e.*, $\mathbf{Z}$ is a Markov process); the observed data depend only on the current-time hidden state via the emission probability $p(d_t|z_t)$. Following the convention to the field, we model all transition probabilities as multinomial distributions and all emission probabilities as Gaussian distributions (McKinney et al. 2006; Dahan et al. 1999), ignoring for the moment the complica-

tion of modeling a variable distributed on the interval $[0, 1]$ with a distribution of support $(-\infty, \infty)$.

For a smFRET time series with observed data $(d_1, \ldots, d_T) = \mathbf{D}$ and corresponding hidden state conformations $(z_1, \ldots, z_T) = \mathbf{Z}$, the joint probability of the observed and hidden data is

$$p(\mathbf{D}, \mathbf{Z}|\vec{\theta}, \mathrm{K}) = p(z_1|\vec{\theta}, \mathrm{K}) \left[\prod_{t=2}^{T} p(z_t|z_{t-1}, \vec{\theta}, \mathrm{K})\right] \prod_{t=1}^{T} p(d_t|z_t, \vec{\theta}, K) \tag{5.7}$$

where $\mathbf{Z}$ comprises four types of parameters: a K-element vector, $\vec{\pi}$ where the k^{th} component, π_k, holds the probability of starting in the k^{th} state; a $K \times K$ transition matrix, A, where $a_{j,k}$ is the probability of transitioning from the j^{th} hidden state to the k^{th} hidden state (*i.e.* $a_{j,k} = p(z_t = k|z_{t-1} = j)$); and two K-element vectors, $\vec{\mu}$ and $\vec{\lambda}$, where μ_k and λ_k are the mean and precision of the Gaussian distribution of the k^{th} state.

As in Eq. 5.4, the evidence follows directly from multiplying the likelihood by priors and marginalizing:

$$\begin{aligned} p(\mathbf{D}|\mathrm{K}) \;=\; & \sum_{z} \int d\vec{\theta} p(\vec{\pi}|K) p(A|K) p(\vec{\mu}, \vec{\lambda}|K) p(z_1|\vec{\pi}, K) \\ & \times \left[\prod_{t=2}^{T} p(z_t|z_{t-1}, A, K)\right] \prod_{t=1}^{T} p(d_t|z_t, \vec{\mu}, \vec{\lambda}, K). \end{aligned} \tag{5.8}$$

The $p(\vec{\pi}|\mathrm{K})$ and each row of $p(A|K)$ are modeled as Dirichlet distributions; each pair of μ_k and λ_k are modeled jointly as a Gaussian-Gamma distribution. These distributions are the standard choice of priors for multinomial and Gaussian distributions (Bishop 2006). If we also assume that $q(\mathbf{Z}, \vec{\theta})$ factorizes into $q(\mathbf{Z})q(\vec{\theta})$, this HMM can be solved via VBEM (*cf.* (Ji et al. 2006)). Algebraic expressions

for these distributions can be found in Sec. 6.4.1. Their parameter settings and the effect of their parameter settings on data inference can be found in Sec. 6.4.2 and Sec. 6.4.3, respectively. We found that for the experiments considered here, and the range of prior parameters tested, there is little discernible effect of the prior parameter settings on the data inference.

The variational approximation to the above evidence utilizes the dynamic program termed the forward-backward algorithm (Rabiner 1989), which requires $\mathrm{O}(K^2T)$ computations, rendering the computation feasible. (In comparison, direct summation over all terms requires $O(K^T)$ operations.) We emphasize that, while individual steps in the ME calculation are slightly more expensive than their ML counterparts, the scaling with the number of states and observations is identical. As discussed in section 5.3.3, in addition to calculating the evidence the variational solution yields a distribution approximating the probability of the parameters given the data. Idealized traces can be calculated by taking the most probable parameters from these distributions and calculating the most probable hidden state trajectory using the Viterbi algorithm (Viterbi 1967).

5.4.2 Rates from states

HMMs are used to infer the number of conformational states present in the molecular complex as well as the transition rates between states. Here, we follow the convention of the field by fitting every trace individually (since the number and mean values of smFRET states often vary from traces to trace). Unavoidably then, an ambiguity is introduced comparing FRET state labels across multiple traces,

since "state 2" may refer to the high variant of a low state in one trace and to the low variant of a high state in a separate trace. To overcome this ambiguity, rates are not inferred directly from $q(\vec{\theta})$, but rather from the idealized traces $\hat{\mathbf{Z}}$ where

$$\hat{\mathbf{Z}} = \underset{\mathbf{Z}}{\operatorname{argmax}}\, q(\mathbf{Z}|\mathbf{D}, \vec{\theta}_\dagger, K) \tag{5.9}$$

and $\vec{\theta}_\dagger$ are, for ME, the parameters specifying the optimal parameter distribution $q_*(\mathbf{Z}, \vec{\theta})$ or, for ML, the most likely parameters, $\vec{\theta}_*$. The number of states in the data set can then be determined by combining the idealized traces and plotting a 1D FRET histogram or transition density plot (TDP). Inference facilitates the calculation of transition rates by, for example, dwell-time analysis, TDP analysis, or by dividing the sum of the dwell times by the total number of transitions (Cornish et al. 2008; McKinney et al. 2006). In this work, we determine the number of states in an individual trace using ME. To overcome the ambiguity of labels when combining traces, we follow the convention of the field and use 1D FRET histograms and/or TDPs to infer the number of states in experimental data sets and calculate rates using dwell time analysis (Sec. 6.3.3).

5.5 Numerical experiments

We created a software package to implement VBEM for FRET data called vbFRET. Software was written in MATLAB and is available open source, including a point and click GUI. All ME data inference was performed using vbFRET. All ML data inference was performed using HaMMy (McKinney et al. 2006), although we note that any analysis based on ML should perform similarly (see Sec. 6.3.1 for practicalities

regarding implementing ML). Parameter settings used for both programs, methods for creating computer generated synthetic data, and methods for calculating rate constants for experimental data can be found in Sec. 6.3. Following the convention of the field, in subsequent sections the dimensionless FRET ratio is quoted in dimensionless "units" of FRET.

5.5.1 Example: maximum likelihood vs maximum evidence

To illustrate the differences between ML and ME, consider the synthetic trace shown in Fig. 5.1, generated with three noisy states ($K_0 = 3$) centered at $\mu_z = (0.41, 0.61, 0.81)$ FRET. This trace was analyzed by both ME and ML with $K = 1$ (underfit), $K = 3$ (correctly fit), and $K = 5$ (overfit) (Fig. 5.1*A*). In the cases when only one or three states are allowed, ME and ML perform similarly. However, when five states are allowed, ML overfits the data, whereas ME leaves two states unpopulated and correctly infers three states, illustrated clearly via the idealized trace.

Moreover, whereas the likelihood of the overfitting model is larger than that of the correct model, the evidence is largest when only three states are allowed ($p(\mathbf{D}|\vec{\theta}_*, K > K_0) > p(\mathbf{D}|\vec{\theta}_*, K_0)$; however, $p(\mathbf{D}|\mathrm{K})$ peaks at $K = K_0 = 3$). The ability to use the evidence for model selection is further illustrated in Fig. 5.1*B*, in which the same data as in Fig. 5.1*A* are analyzed using both ME and ML with $1 \leq K \leq 10$. The evidence is greatest when $K = 3$; however, the likelihood increases monotonically as more states are allowed, ultimately leveling off after five or six states are allowed.

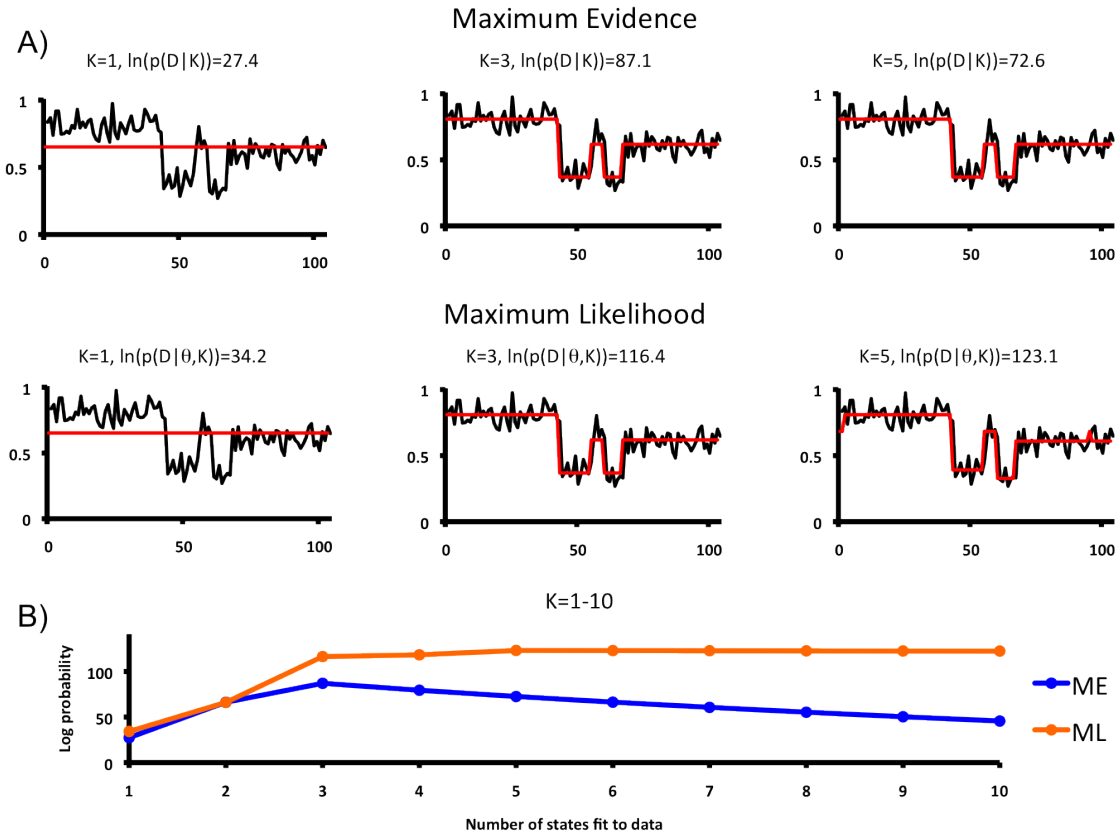

Figure 5.1: A single (synthetic) FRET trace analyzed by ME and ML. The trace contains 3 hidden states. A) (Top) Idealize traces inferred by ME when $K = 1, K = 3$, and $K = 5$, as well as the corresponding log(evidence) for the inference. The data are under resolved when $K = 1$, but for both $K = 3$ and $K = 5$ the correct number of states are populated. (Bottom) Idealized traces inferred by ML when $K = 1, K = 3$, and $K = 5$, as well as the corresponding log(likelihood). Inference when $K = 1$ and $K = 3$ are the same as for ME but the data are overfit when $K = 5$. B) The log evidence from ME (black) and log likelihood from ML (gray) for $1 \leq K \leq 10$. The evidence is correctly maximized for $K = 3$, but the likelihood increases monotonically.

5.5.2 Statistical validation

ME can be statistically validated by generating synthetic data, for which the true trajectory of the hidden state $\mathbf{Z}_0$ is known, and quantifying performance relative to ML. We performed such numerical experiments, generating several thousand synthetic traces, and quantified accuracy as a function of signal-to-noise via four probabilities: (1) accuracy in number of states $p(|\hat{\mathbf{Z}}| = |\mathbf{Z}_0|)$: the probability in any trace of inferring the correct number of states (where $|\mathbf{Z}_0|$ is the number of states in the model generating the data and $|\hat{\mathbf{Z}}|$ is the number of populated states in the idealized trace); (2) accuracy in states $p(\hat{\mathbf{Z}} = \mathbf{Z}_0)$: the probability in any trace at any time of inferring the correct state; (3) sensitivity to true transitions: the probability in any trace at any time that the inferred trace $\hat{\mathbf{Z}}$ exhibits a transition, given that $\mathbf{Z}_0$ does; and (4) specificity of inferred transitions: the probability in any trace at any time that the true trace $\mathbf{Z}_0$ does not exhibit a transition, given that $\hat{\mathbf{Z}}$ does not. We note that, encouragingly, for the ME inference, $|\hat{\mathbf{Z}}|$ always equaled K_* as defined in Eq. 5.2.

We identify each inferred state with the true state which is closest in terms of their means provided the difference in means is less than 0.1 FRET. Inferred states for which no true state is within 0.1 FRET are considered inaccurate. Note that we do not demand that one and only one inferred state be identified with the true state. This effective smoothing corrects overfitting errors in which one true state has been inaccurately described by two nearby states (consistent with the convention of the field for analyzing experimental data).

For all synthetic traces, $K_0 = 3$ with means centered at $\mu_z = (0.25, 0.5, 0.75)$

FRET. Traces were made increasingly noisy by increasing the standard deviation, σ, of each state. Ten different noise levels, ranging from $\sigma \approx 0.02$ to $\sigma \approx 0.15$ were used. Over this noise range, traces vary from trivially resolvable by eye to unresolvable by either inference program — both methods correctly infer $< 45\%$ of transitions by the final value of σ. Trace length, T, varied from $50 \leq T \leq 500$ time steps, drawn randomly from a uniform distribution. One time step corresponds to one time-binned unit of an experimental trace, which is typically 25–100 msec for most CCD camera based experiments. Fast-transitioning (mean lifetime of 4 time steps between transitions) and slow-transitioning (mean lifetime of 15 time steps between transitions) traces were created and analyzed separately. Transitions were equally likely from all hidden states to all hidden states. For each of the 10 noise levels and 2 transition speeds, 100 traces were generated (2,000 traces in total). Traces for which $K_0 = 2$ (Fig. 6.7) and $K_0 = 4$ (Fig. 6.8) were created and analyzed as well. The results were qualitatively similar and can be found in Sec. 6.5.

As expected, both programs performed better on low noise traces than on high noise traces. ME correctly determined the number of FRET states more often than ML in all cases except for the noisiest fast-transitioning trace set (Fig. 5.2, top left). Of the 2,000 traces analyzed here using ME and ML, ME overfit 1 and underfit 232. ML overfit 767 and underfit 391. In short, ME essentially eliminated overfitting of the individual traces, whereas ML overfit 38% of individual traces. Over 95% (all but 9) of ME underfitting errors occurred on traces with FRET state noise > 0.09, whereas ML underfitting was much more evenly distributed (at least 30 traces at

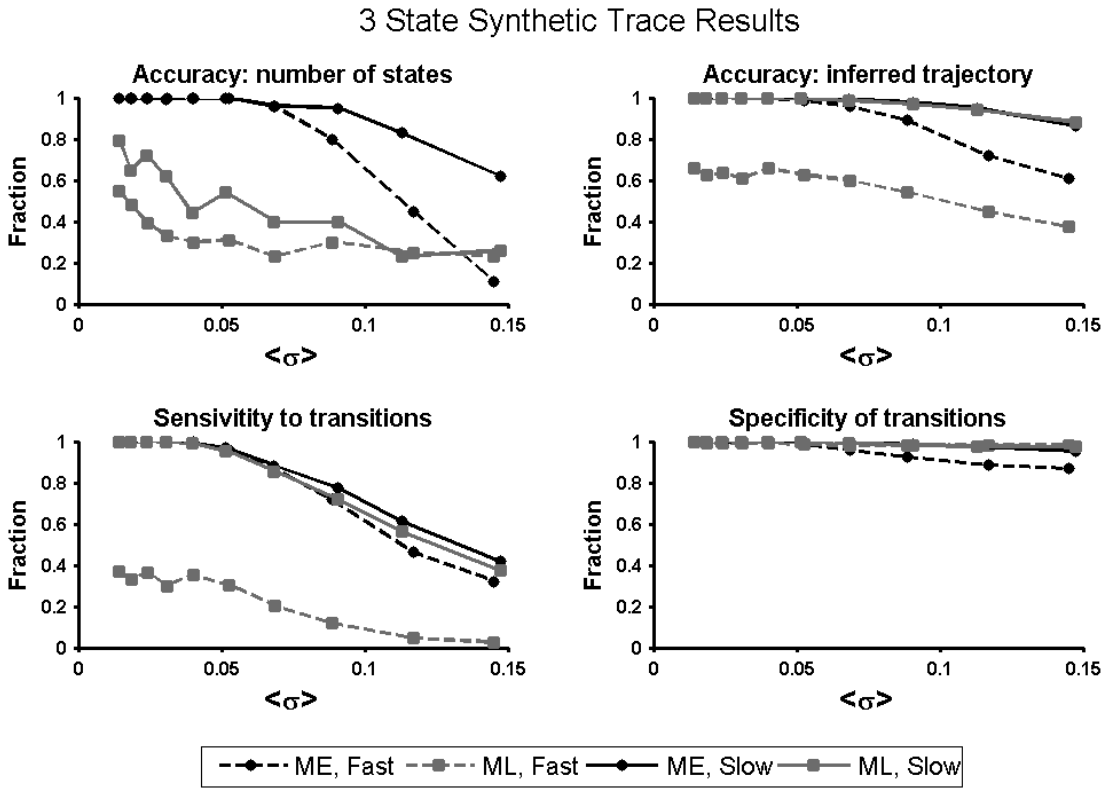

Figure 5.2: Comparison of ME and ML as a function of increasing hidden state noise. Fast transitioning (hidden state mean lifetime of 4 time steps) and slow transitioning (hidden state mean lifetime of 15 time steps) traces were created and analyzed separately. Each data point represents the average value taken over 100 traces. (Top left) $p(|\hat{\mathbf{Z}}| = |\mathbf{Z}_0|)$: the probability in any trace of inferring the correct number of states. (Top right) $p(\hat{\mathbf{Z}} = \mathbf{Z}_0)$: the probability in any trace at any time of inferring the correct state. (Bottom left) Sensitivity to true transitions: the fraction of time the correct FRET state was inferred during FRET trajectories. (Bottom right) Specificity of inferred transitions: the probability in any trace at any time that the true trace $\mathbf{Z}_0$ does not exhibit a transition, given that $\hat{\mathbf{Z}}$ does not. Error bars on all plots were omitted for clarity and because the data plotted represent mean success rates for Bernoulli processes (and, therefore, determine the variances of the data as well).

every noise level were underfit by ML). The underfitting of noisy traces by ME may be a result of the intrinsic resolvability of the data, rather than a shortcoming of the inference algorithm; as the noise of two adjacent states becomes much larger than the spacing between them, the two states become indistinguishable from a single noisy state (in the limit, there is no difference between a one state and two state system if the states are infinitely noisy). The causes of the underfitting errors by ML are less easily explained, but suggest that the ML algorithm has not converged to a global optimum in likelihood (for reasons explained in Sec. 6.3.2).

In analyzing the slow-transitioning traces, the methods performed roughly the same on probabilities (2–4) (always within $\sim$5% of each other). For the fast-transitioning traces, however, ME was much better at inferring the true trajectory of traces (by a factor of 1.5–1.6 for all noise levels) and showed superior sensitivity (factor of 2.7–12.5) to transitions at all noise levels. The two methods showed the same specificity to transitions until a noise level of $\sigma > 0.8$, beyond which ML showed better specificity (factor of 1.06–1.13). Inspection of the individual traces showed that all three of these results were due to ML missing many of the transitions in the data.

These results on synthetic data suggest that when the number of states in the system is unknown, ME clearly performs better at identifying FRET states. For inference of idealized trajectories, ME is at least as accurate as ML for slow-transitioning traces and more accurate for fast-transitioning traces. The performance of ME on fast-transitioning traces is particularly encouraging since detection of a transient biophysical state is often an important objective of smFRET

experiments, as discussed below.

5.6 Results

Having validated inference with vbFRET, we compared ME and ML inference on experimental smFRET data, focusing our attention on the number of states and the transition rates. The data we used for this analysis report on the conformational dynamics of the ribosome, the universally-conserved ribonucleoprotein enzyme responsible for protein synthesis, or translation, in all organisms. One of the most dynamic features of translation is the precisely directed mRNA and tRNA movements that occur during the translocation step of translation elongation. Structural, biochemical, and smFRET data overwhelmingly support the view that, during this process, ribosomal domain rearrangements are involved in directing tRNA movements (Moazed and Noller 1989; Blanchard et al. 2004b; Munro et al. 2007; Kim et al. 2007; Fei et al. 2008; Cornish et al. 2008; Agirrezabala et al. 2008; Julian et al. 2008; Cornish et al. 2009). One such ribosomal domain is the L1 stalk, which undergoes conformational changes between open and closed conformations that correlate with tRNA movements between so-called classical and hybrid ribosome-bound configurations (Fig. 5.3*A*) (Fei et al. 2008; Cornish et al. 2009; Fei et al. 2009; Sternberg et al. 2009).

Using fluorescently-labeled tRNAs and ribosomes, we have recently developed smFRET probes between tRNAs ($\text{smFRET}_{\text{tRNA}-\text{tRNA}}$) (Blanchard et al. 2004b), ribosomal proteins L1 and L9 ($\text{smFRET}_{\text{L1}-\text{L9}}$) (Fei et al. 2008), and ribosomal pro-

tein L1 and tRNA ($\text{smFRET}_{\text{L1-tRNA}}$) (Fei et al. 2009). Collectively, these data demonstrate that, upon peptide bond formation, tRNAs within pre-translocation (PRE) ribosomal complexes undergo thermally-driven fluctuations between classical and hybrid configurations ($\text{smFRET}_{\text{tRNA-tRNA}}$) that are coupled to transitions of the L1 stalk between open and closed conformations ($\text{smFRET}_{\text{L1-L9}}$). The net result of these dynamics is the transient formation of a direct L1 stalk-tRNA contact that persists until the tRNA and the L1 stalk stochastically fluctuate back to their classical and open conformations, respectively ($\text{smFRET}_{\text{L1-tRNA}}$). This intermolecular L1 stalk-tRNA contact is stabilized by binding of elongation factor G (EF-G) to PRE and maintained during EF-G catalyzed translocation (Fei et al. 2008; 2009).

Here we compare the rates of L1 stalk closing (k_{close}) and opening (k_{open}) obtained from ME and ML analysis of $\text{smFRET}_{\text{L1-L9}}$ PRE complex analogs (PMN) under various conditions (which have the same number of FRET states by both inference methods) and the number of states inferred for $\text{smFRET}_{\text{L1-tRNA}}$ PMN complexes by ME and ML. (FRET complexes shown in Fig. 5.3*B*.) These data were chosen for their diversity of smFRET ratios. The $\text{smFRET}_{\text{L1-L9}}$ ratio fluctuates between FRET states centered at 0.34 and 0.56 (*i.e.* a separation of 0.22 FRET), whereas the $\text{smFRET}_{\text{L1-tRNA}}$ ratio fluctuates between FRET states centered at 0.09 and 0.59 FRET(*i.e.* a separation of 0.50 FRET). In addition, $\text{smFRET}_{\text{L1-L9}}$ data were recorded under conditions that favor either fast-transitioning ($\text{PMN}_{\text{fMet+EFG}}$) or slow-transitioning (PMN_{fMet} and PMN_{Phe}) complexes (complex compositions listed in Table 5.6).

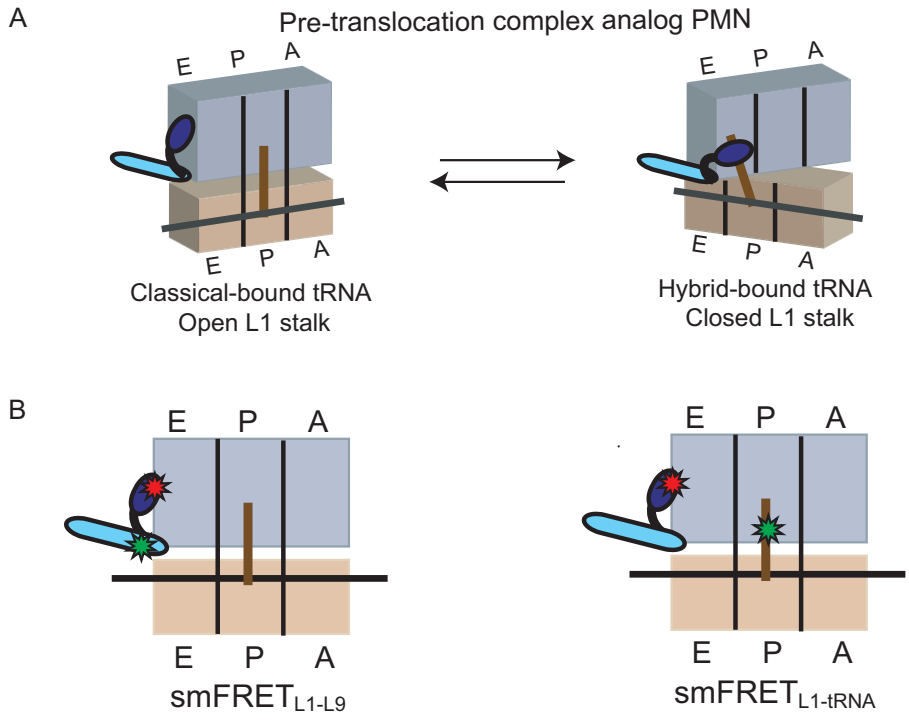

Figure 5.3: Conformational rearrangements within a pre-translocation (PMN) complex and smFRET labeling schemes. (A) Cartoon representation of a PMN complex analog. The small and large ribosomal subunits are shown in tan and lavender, respectively, with the L1 stalk depicted in dark blue, and ribosomal protein L9 in cyan. The aminoacyl-, peptidyl- and deacylated-tRNA binding sites are labeled as A, P and E, respectively, and the P-site tRNA is depicted as a brown line. PMN complex analogs are generated by adding the antibiotic puromycin to a post-translocation complex carrying a deacylated-tRNA at the E site and a peptidyl-tRNA at the P site. The resulting PMN complex analog exists in a thermally-driven dynamic equilibrium between two major conformational states in which the P-site tRNA fluctuations between classical and hybrid configurations correlate with the L1 stalk fluctuations between open and closed conformations. (B) Two labeling schemes have been developed in order to investigate PMN complex dynamics using smFRET. PMN complexes are cartooned as in (A) with Cy3 and Cy5 depicted as green and red stars, respectively. $\text{smFRET}_{\text{L1}-\text{L9}}$ (left), which involves a Cy5 label on ribosomal protein L1 within the L1 stalk and a Cy3 label on ribosomal protein L9 at the base of the L1 stalk, reports on the intrinsic conformational dynamics of the L1 stalk. $\text{smFRET}_{\text{L1}-\text{tRNA}}$ (right), which involves a Cy5 label on ribosomal protein within the L1 stalk as in $\text{smFRET}_{\text{L1}-\text{L9}}$ and a Cy3 label on the P-site tRNA, reports on the formation and disruption of a direct interaction between the closed L1 stalk and the hybrid bound P-site tRNA.

First, we compared the smFRET$_{\text{L1}-\text{L9}}$ data obtained from PMN$_{\text{fMet}}$, PMN$_{\text{Phe}}$, and PMN$_{\text{fMet+EFG}}$. As expected from previous studies (Fei et al. 2009), 1D histograms of idealized FRET values from both inference methods showed two FRET states centered at 0.34 and 0.56 FRET (and one additional state due to photobleaching, for a total of three states). When individual traces were examined for overfitting, however, ML inferred four or five states in $20.1\% \pm 3.7\%$ of traces in each data set whereas ME inferred four or five states in only $0.9\% \pm 0.5$ of traces. Consequently, more post-processing was necessary to extract transition rates from idealized traces inferred by ML.

Data set*	Method	k_{close}	k_{open}
PMN$_{\text{Phe}}$†	ME	0.66 ± 0.05	1.0 ± 0.2
	ML	0.65 ± 0.06	1.0 ± 0.3
PMN$_{\text{fMet}}$‡	ME	0.53 ± 0.08	1.7 ± 0.3
	ML	0.52 ± 0.06	1.8 ± 0.3
PMN$_{\text{fMet+EFG}}$	ME	3.1 ± 0.6	1.3 ± 0.2
$(1\mu M)$§	ML	2.1 ± 0.4	1.0 ± 0.2
PMN$_{\text{fMet+EFG}}$	ME	2.6 ± 0.6	1.5 ± 0.1
$(0.5\mu M)$§	ML	2.0 ± 0.3	1.0 ± 0.1

* Rates reported here are the average and standard deviation from three or four independent data sets. Rates were not corrected for photobleaching of the fluorophores.

† PMN$_{\text{Phe}}$ was prepared by adding the antibiotic puromycin to a post-translocation complex carrying deacylated-tRNA$^{\text{fMet}}$ at the E site and fMet-Phe-tRNA$^{\text{Phe}}$ at the P site, and thus contains a deacylated-tRNA$^{\text{Phe}}$ at the P site.

‡ PMN$_{\text{fMet}}$ was prepared by adding the antibiotic puromycin to an initiation complex carrying fMet-tRNA$^{\text{fMet}}$ a the P site, and thus contains a deacylated-tRNA$^{\text{fMet}}$ at the P site.

§ 1.0 μM and 0.5 μM EF-G in the presence of 1 mM GDPNP (a non-hydrolyzable GTP analog) were added to PMN$_{\text{fMet}}$, respectively.

Table 5.1: Comparison of smFRET$_{\text{L1}-\text{L9}}$ transition rates inferred by ME and ML.

Our results (Table 5.6) demonstrate that there is very good overall agreement between the values of k_{close} and k_{open} calculated by ME and ML. For the relatively slow-transitioning PMN_{fMet} and PMN_{Phe} data, the values of k_{close} and k_{open} obtained from ME and ML are indistinguishable. For the relatively fast-transitioning $\text{PMN}_{\text{fMet+EFG}}$ data, however, the values of k_{close} and k_{open} obtained differ slightly between ME and ML. Since the true transition rates of the experimental $\text{smFRET}_{\text{L1-L9}}$ data can never be known, it is impossible to assess the accuracy of the rate constants obtained from ME or ML in the same way we could with the analysis of synthetic data. While we cannot say which set of k_{close} and k_{open} values are most accurate for this fast-transitioning data set, our synthetic results would predict a larger difference between rate constants calculated by ME and ML for faster-transitioning data and suggest that the values of k_{close} and k_{open} calculated with ME have higher accuracy (Fig. 5.2).

Consistent with previous reports (Fei et al. 2008), ML infers two FRET states centered at $f_{\text{low}} \equiv 0.09$ and $f_{\text{high}} \equiv 0.59$ FRET (plus one photobleached state) for all $\text{smFRET}_{\text{L1-tRNA}}$ data sets. Conflicting with these results, however, ME infers three FRET states (plus a photobleached state) for these data sets. Two of these FRET states are centered at f_{low} and f_{high}, as in the ML case, while the third "putative" state is centered at $f_{\text{mid}} \equiv 0.35$ FRET, coincidentally at the mean between f_{low} and f_{high}. Indeed, TDPs constructed from the idealized trajectories generated by ME or ML analysis of the $\text{PMN}_{\text{fMet+EFG}}$ $\text{smFRET}_{\text{L1-tRNA}}$ data set evidence the appearance of a new, highly populated state at f_{mid} in the ME-derived TDP that is virtually absent in the ML-derived TDP (Fig. 5.5). Consistent with the

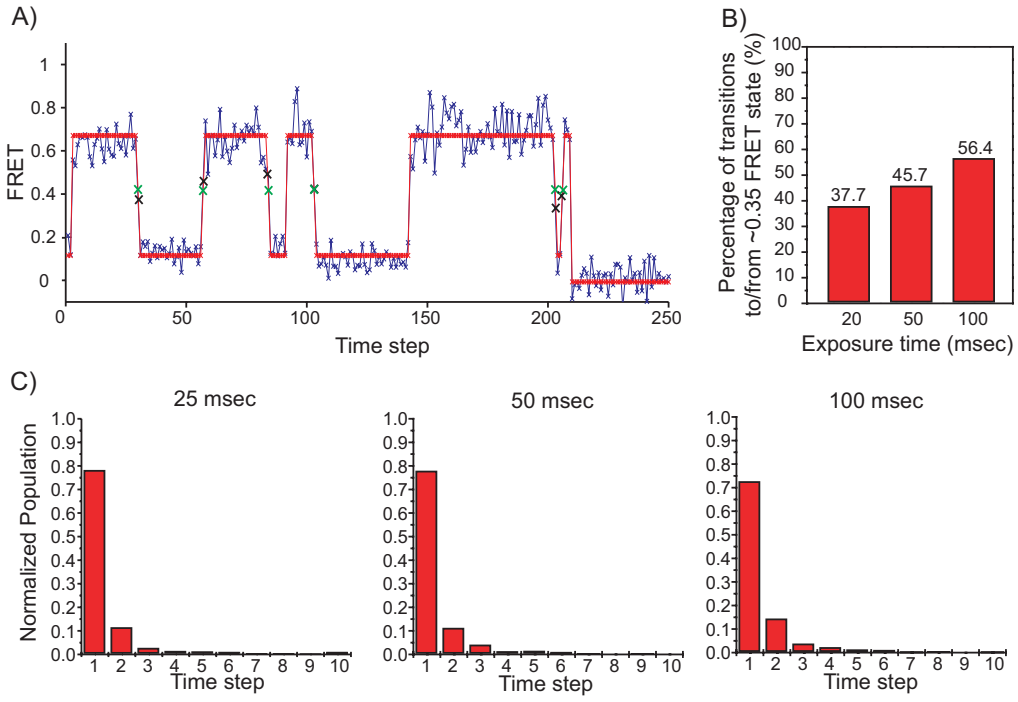

Figure 5.4: Analysis of the smFRET$_{\mathrm{L1-tRNA}}$ f_{mid} state. A) A representative smFRET$_{\mathrm{L1-tRNA}}$ trace idealized by ME, taken from the 50 msec exposure time data set. Both the observed data and idealized path are shown. Individual data points, real and idealized, are shown as Xs. To emphasize the data at or near f_{mid}, the Xs are enlarged and the observed data are shown in black. (B) Bar graph of the percentages of transitions to or from the f_{mid} state under 25 msec, 50 msec and 100 msec CCD integration time. (C) Normalized population histograms of dwell time spent at the f_{mid} state under 25 msec, 50 msec, and 100 msec CCD integration time.

TDPs, ∼46% of transitions in the ME-analyzed smFRET$_{\mathrm{L1-tRNA}}$ trajectories are either to or from the new f_{mid} state (Fig. 5.4*B*). This f_{mid} state is extremely short lived; ∼75% of the data assigned to f_{mid} consist of a single observation, *i.e.* with a duration at or below the CCD integration time CCD blurring artifact (here, 50 msec) (Fig. 5.4*C*). A representative ME-analyzed smFRET$_{\mathrm{L1-tRNA}}$ trace is shown in Fig. 5.4*A*.

There are at least two possible origins for this putative new state. The first is a very short-lived (*i.e.* lifetime $\leq$ 50 msec), *bona fide*, previously unidentified intermediate conformation of the PMN complex. The second is that f_{mid} data are artifactual, resulting from the binning of the continuous-time FRET signal during CCD collection. Each time binned data point represents the average intensity of thousands or more photons. If a transition occurs 25 msec into a 50 msec time step, half the photons will come from the f_{low} state and half from the f_{high} state, resulting in a datum at approximately their mean. This type of CCD blurring artifact would be lost in the noise of closely spaced FRET states, but would become more noticeable as the FRET separation between states increases.

To distinguish between these two possibilities, we recorded $\mathrm{PMN_{fMet+EFG}}$ $\mathrm{smFRET_{L1-tRNA}}$ data at half and double the integration times (*i.e.* 25 msec and 100 msec). If the f_{mid} state is a true conformational intermediate then: (1) the percentage of transitions exhibiting at least one data point at or near f_{mid} should increase as the integration time decreases, and (2) the number of consecutive data points defining the dwell time spent at or near f_{mid} should increase as the integration time decreases. Conversely, if the f_{mid} state arises from a time averaging artifact, then: (1) the percentage of transitions containing at least one data point at or near f_{mid} should increase as the integration time increases, as longer integration times increase the probability that a transition will occur during the integration time, and (2) the number of consecutive data points defining the dwell time spent at or near the f_{mid} state should be independent of the integration time, as transitions occurring within the integration time will always be averaged to generate a single

data point.

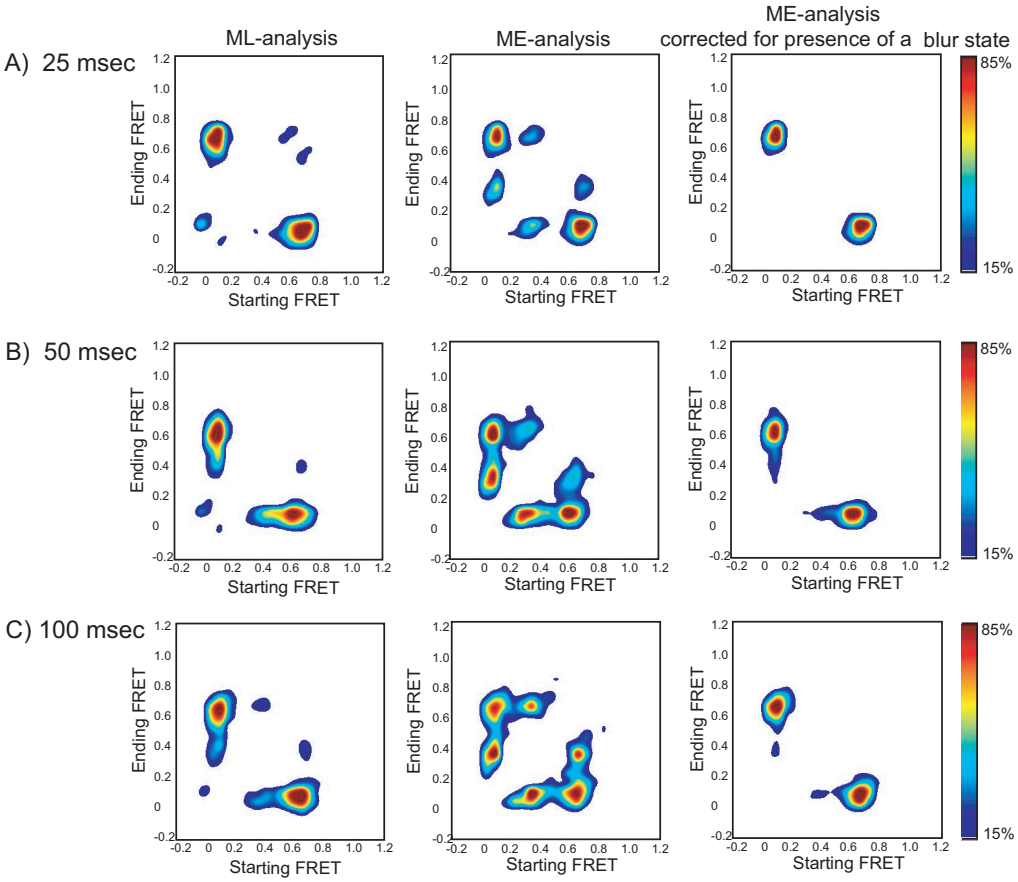

Figure 5.5: Transition density plots (TDP) of smFRET$_{\mathrm{L1-tRNA}}$ PMN$_{\mathrm{fMet+EF-G}}$ derived from ME and ML analysis with different CCD integration times. TDPs are contour plots showing the kernel density estimation of the transitions in idealized traces (with starting and ending FRET values of the transitions as the X and Y axes, respectively). Note that transitions to short-lived or nearby states count with equal weight as those to long-lived states in a TDP. This should not be confused with a time-density plot, which illustrates the probability of observing a pair of experimental values at two different times $p(y(t), y(t + \delta t))$, which can be made from the FRET data themselves without appealing to statistical inference. The plots show ML (left), ME (middle) and ME analysis corrected for the presence of a blur state (right). Contours are plotted from tan (lowest population) to red (highest population). Different CCD integration times were used for recoding these data sets: (A) 25 msec, (B) 50 msec, and (C) 100 msec. For interpretation of the significance of these TDPs, *cf.* Sec. 5.6.

Consistent with the view that the f_{mid} state arises from time-averaging over the integration time, Fig. 5.4*B* demonstrates that the percentage of transitions containing at least one data point at or near f_{mid} increases as the integration time increases. This manifests as the increase in the density of transitions starting or ending at f_{mid} as the integration time decreases for the ME-derived TDPs in Fig. 5.5. These data are further supported by the results presented in Fig. 5.4*C*, demonstrating that the number of consecutive data points defining the dwell time of the f_{mid} state is remarkably insensitive to the integration time. We conclude that the f_{mid} state identified by ME is composed primarily of a time-averaging artifact which we refer to as "camera blurring" and the ME-inferred f_{mid} state as the "blur state". Although ML infers four or five states in 35% of the traces (compared to only 25% for ME), for some reason ML significantly suppresses, but does not completely eliminate, detection of this blur state in the individual smFRET trajectories. At present, we cannot determine whether this is a result of the ML method itself (*i.e.* overfitting noise in one part of the trace may cause it to miss a state in another) or due to the specific implementation of ML in the software we used (Sec. 6.3.1). In retrospect, the presence of blur states should not be surprising, since they follow trivially from the time-averaging that results from averaging over the CCD integration time. In Sec. 6.2, we propose a method for correcting these blur artifacts.

The observation that ML analysis does not detect a blur state that is readily identified by ME analysis is in line with our results on synthetic data in which ME consistently outperforms ML in regards to detecting the true number of states

in the data, particularly in fast-transitioning data, and strongly suggests that ME will generally capture short-lived intermediate FRET states that ML will tend to overlook. While this feature of ML might be desirable in terms of suppressing blur states such as the one we have identified in the $\text{smFRET}_{\text{L1}-\text{tRNA}}$ data set, it is undesirable in terms of detecting *bona fide* intermediate FRET states that may exist in a particular data set.

5.7 Conclusions

These synthetic and experimental analyses confirm that ME can be used for model selection (identification of the number of smFRET states) at the level of individual traces, improving accuracy and avoiding overfitting. Additionally, ME inference solved by VBEM provides q_*, an estimate of the true parameter and idealized trace posterior, making possible the analysis of kinetic parameters, again at the level of individual traces. As a tool for inferring idealized traces, ME produces traces which are visually similar to those of ML; in the case of synthetic data generated to emulate experimental data, ME performs with comparable or superior accuracy. The idealized trajectories inferred by ME required substantially less post-processing, however, since ME usually inferred the correct number of states to the data and, consequently, did not require states with similar idealized values within the same trace to be combined in a post-processing step. The superior trajectory inference, accuracy, and sensitivity to transitions of ME on fast transitioning synthetic traces suggests that the differences in transition rates calculated for fast transitioning

experimental data is a result of superior fitting by ME as well.

In some experimental data, ME detected a very short lived blur state, which comparison of experiments at different sampling rates suggests results from a camera time averaging artifact. Once detected by ME, the presence of this intermediate state is easily confirmed by visual inspection, but yet was not identified by ML inference. Although not biologically relevant in this instance, this result suggests that ME inference is able to uncover real biological intermediates in smFRET data that would be missed by ML.

We conclude by emphasizing that this method of data inference is in no way specific to smFRET. The use of ME and VBEM could improve inference for other forms of biological time series where the number of molecular conformations is unknown. Some examples include motor protein trajectories with an unknown number of chemomechanical cycles (*i.e.* steps), DNA/enzyme binding studies with an unknown number of binding sites and molecular dynamics simulations where important residues exhibit an unknown number of rotamers.

All code used in this analysis, as well as a point and click GUI interface, is available open source via http://vbFRET.sourceforge.net.

5.8 Acknowledgments

It is a pleasure to acknowledge helpful conversations with Taekjip Ha and Vijay Pande, Harold Kim and Eric Greene for comments on the manuscript, mathematical collaboration with Alexandro D. Ramirez, and Subhasree Das for man-

aging the Gonzalez laboratory. This work was supported by a grant to CHW from the NIH (5PN2EY016586-03) and grants to RLG from the Burroughs Wellcome Fund (CABS 1004856), the NSF (MCB 0644262), and the NIH-NIGMS (1RO1GM084288-01). CHW also acknowledges the generous support of Mr. Ennio Ranaboldo.

Chapter 6

vbFRET II

The work described in Ch. 5 was too large to fit in one publication. The remainder of the work, originally published as the supporting material from Ch. 5, is reproduced here with minor modifications.

6.1 2D inference

Instead of analyzing the 1D FRET ratio, it is also possible to model the donor and acceptor molecule intensities directly. Such analysis can be accomplished by treating the donor and acceptor signals as a 2D vector, which is then fit by a 2D Gaussian. (The VBEM solution to the HMM with multidimensional Gaussian observables is solved in (Ji et al. 2006), and requires only a minor change to the code used in the rest of this work.) Because information is necessarily lost by transforming the 2D donor / acceptor signal into a 1D ratio, the 2D data may yield more information about the FRETing complex and, therefore, better inference.

However, it is also possible that the 2D data provide information only about the photophysics rather than the biophysics, i.e., that the only biophysically meaningful quantity is the donor / acceptor transfer efficiency reflected in the FRET ratio. While it is outside the scope of this paper to assess the relative merits of 1D and 2D FRET analysis, it is worth considering whether evidence could be used to evaluate the relative accuracies of inferences performed in 1 or 2 dimensions.

Intuitively, one should not expect evidence to be an appropriate evaluative quantity in this situation: evidence allows one to select between competing models for a fixed data set, i.e. a selection between $p(\mathbf{D}|m_1)$ and $p(\mathbf{D}|m_2)$ (which is how evidence is used in the rest of this work). Here, we are asking the for $p(\mathbf{D}_1|m_1)$ versus $p(\mathbf{D}_2|m_2)$, where $\mathbf{D}_1$ and $\mathbf{D}_2$ are the 1D FRET ratio and donor / acceptor data. Because the space of 2D data sets is so much larger than the space of 1D data sets, the evidence for the 2D inference may be lower, regardless of the quality of the inference.

To test this hypothesis, the following data set was devised. Ten $K = 3$ traces were generated as described in Sec. 6.3.4, each of length $T = 200$, with 1D FRET state means of $\sim 0.11 \pm 0.014$ FRET. The traces were then replicated 5 times, and both the donor and acceptor signals were modified by multiplication with the a linearly decreasing envelope function $A(t) = 1 - (t/T)S$, where $S = \{0, 0.15, 0.30, 0.45, 0.60\}$. When $S = 0$, the 2D traces should have more information than do the 1D FRET transformations. By the time $S = 0.60$, the 2D traces should be poorly described by 2D Gaussian HMMs (since the means of the donor / acceptor signals at the end of the traces are 40% of their original values), but the 1D traces

will still look the same as when $S = 0$ (since the multiplying factor cancels out of the FRET ratio). A sample trace is shown in Fig. 6.1.

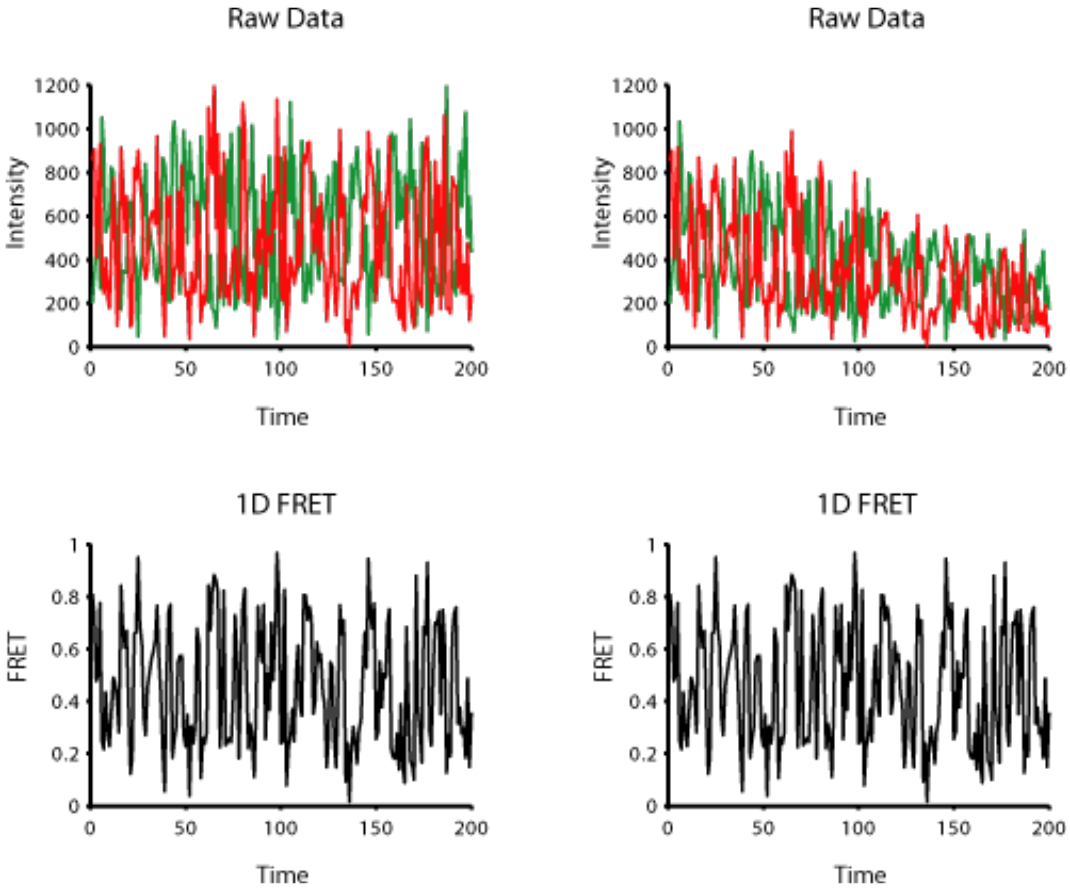

Figure 6.1: (Top left) one of the $S = 0$ traces fit using a 2D Gaussian HMM and (bottom left) the 1D FRET transform. (Right) the same trace, but multiplied by the $A(t) = 1 - 0.60(t/T)$ vector. By the end of the 2D trace the means of the donor and acceptor signal intensities are 40% of their original values while the 1D transformation (bottom right) is unchanged.

The results are shown in Fig. 6.2. The mean and standard deviation of $\ln(p(\mathbf{D}|m))$ for each set of traces are shown in Table 6.1. Consistent with expectations, as S increases from 0 to 0.60, the 1D inference is unchanged, but the accuracy of the inferred trajectory and the sensitivity to transitions decrease for the 2D in-

Table 6.1: $\ln(p(\mathbf{D}|m))$ for 1D and 2D inference.

S:	0	0.15	0.30	0.45	0.60
1D	45.1 ± 18.3	45.1 ± 18.3	45.1 ± 18.3	45.1 ± 18.3	45.1 ± 18.3
2D	-501.2 ± 19.8	-502.5 ± 18.7	-509.1 ± 16.5	-515.5 ± 16.2	-504.6 ± 18.3

ference. Inspection of individual traces shows that when $S = 0$, 2D inference is better or equal to 1D inference, by all four accuracy metrics in Fig. 6.2, for 9 out of 10 traces (data not shown). When $S = 0.60$, accuracy of the inferred trajectory and the sensitivity to transitions is worse for 2D inference than for 1D inference for all traces (specificity of transitions is slightly better for 2D inference, but that is a result of missing transitions in the data).

Regardless of the quality of inference, $\ln(p(\mathbf{D}|m))$ for the 2D inference is lower than for the 1D inference, consistent with our intuition about the far greater number of possible 2D data sets, with $\log p(\mathbf{D}_1|m_1) \approx 40$ and $\log p(\mathbf{D}_2|m_2) \approx -500$. Similar results were observed for all synthetic data sets we have tested (other data sets not shown) and reflect that evidence *cannot* be used to assess the quality of 1D versus 2D data inference. It should be noted, however, that evidence based model selection does work for choosing among different 2D models of varying complexity, as evidenced in Fig. 6.2, since the comparison is again between $p(D|M_1)$ and $p(D|M_2)$.

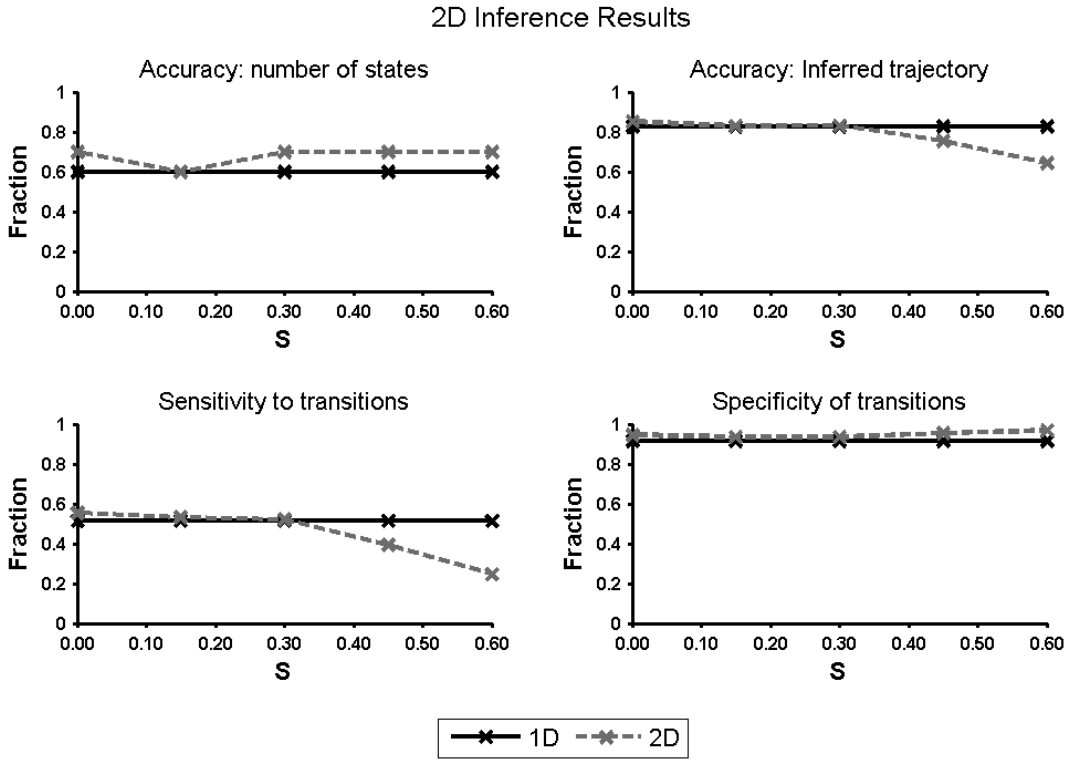

Figure 6.2: Accuracy of 1D inference versus 2D inference. Accuracy metrics are the same as those used in Fig. 5.2.

6.2 Proposed method to correct camera blurring

Single data point artifacts caused by stochastic photophysical fluctuations of fluorophore intensity are a well known and common problem in smFRET data (Roy et al. 2008). These artifacts can be corrected for by applying smoothing algorithms or rolling averages over the data (Blanchard et al. 2004b; Cornish et al. 2009) or ignoring FRET states with a dwell time of one time point (Fei et al. 2008). The artifacts we encounter here are different in nature, since they result from time binning the data rather than a photophysical fluctuation in donor/acceptor signal intensity and, therefore, should be corrected for using a different approach. The algorithm

we propose performs a second round of ME inference on the data, using the idealized traces from the first round of ME inference to make the following modification to the raw data: data which could have resulted from time-averaging artifacts (i.e. events lasting exactly one data point and occurring between two distinct idealized values) were moved to the idealized value closest to the value of the suspected time-averaging artifact (the assumption here is that that a single f_{mid} data point should be considered part of the "real" FRET state that the molecular complex spent the most time in during that transitioning time point). We performed this algorithm on the $\mathrm{smFRET}_{\mathrm{L1-tRNA}}$. The TDP for this "cleaned" data shows the blur state at f_{mid} is virtually eliminated, yielding a result that is wholly consistent with that generated by ML (Fig. 5.5). In general, however, it should be cautioned that a *bona fide* intermediate FRET state may well exist and be buried under a strongly-populated blur state. Unless this intermediate FRET state is positively identified and somehow separated from the blur state (i.e. by obtaining data at an increased integration time), eliminating or ignoring FRET states with dwell times exactly equal to one time point may risk overlooking a *bona fide* intermediate FRET state. We note that the vbFRET software package which we have made available allows the user the opportunity to run this second round of ME analysis with possible blur states detected and cleaned as described above.

6.3 Methods

6.3.1 ML inference settings

Following the HaMMy user manual, ML analyses use $K_{max}+2$ states, where K_{max} is either K_0 (the true number of states) in the case of synthetic data or simply 3 in the case of experimental data, as 1D FRET histograms suggest two biophysical states and one photophysical state: the photobleached state. No additional complexity control was applied to the resulting parameters inferred from individual traces. The default guess for the initial distribution of the means μ_k was used, i.e., uniform spacing between 0 and 1 FRET.

Also consistent with default settings, we use the parameters inferred using only one set of initial parameter-guesses. Note that this differs from the usual implementation of expectation-maximization as a technique for performing ML (*cf.*(Bishop 2006)). Expectation-maximization (the maximization technique used in both HaMMy and vbFRET) provably converges to a local optimum, and therefore the maximization typically is performed using many random restarts for parameter values. One possible reason to avoid this procedure is the inescapable pathology of ML for real-valued emissions (*e.g.* in FRET data) and for which the width of each state is an inferred parameter: the optimization is ill-posed since the case in which one observation is assigned to a state of 0 uncertainty is infinitely likely (*cf.*(Bishop 2006) Ch. 9: "These singularities provide another example of the severe overfitting that can occur in a maximum likelihood approach. We shall see that this difficulty does not occur if we adopt a Bayesian approach.").

6.3.2 ME inference settings

In analyzing synthetic and experimental data with ME, we attempt each choice of $K = 1, 2, \ldots, K_{\max} + 2$ with $K_{\max}$ as above. For synthetic data, 25 random initial guesses were used for each of the traces; for experimental data, 100 initializations were used (though, in our experience, little or no change in the optimization was found after 25 initializations). As with all local optimization techniques, including expectation maximization in ML or in ME, we use the parameters which give the optimum over all restarts (here, the set of parameters specifying the approximating distribution q which gives the maximum evidence $p(\mathbf{D}|\mathrm{K})$).

6.3.3 Rate constant calculations

Rates for the $\mathrm{smFRET}_{\mathrm{L1-L9}}$ experimental data, both for ME and ML analyses, were extracted as previously described (Fei et al. 2008; Sternberg et al. 2009). First, the set of all idealized traces over all times is histogrammed into 50 bins, evenly spaced between -0.2 and 1.2 FRET. The counts in the resulting histogram are given to Origin 7.0, which learns a Gaussian mixture model via expectation-maximization, using user-supplied initial guesses for the three means (we used $\mu = (0, 0.35, 0.55)$ FRET). Origin returns true means and variances for each of the 3 states. From these variances the width at half-max for each mixture is determined, defining three acceptable ranges of fret values. (For this experiment, these ranges had widths of approximately 0.05 FRET. We next re-scan the idealized traces and, for each transition from one acceptable range to another, record the dwell time (the total

time spent within the range; any number of inferred transition within one accepted range are ignored, effectively smoothing overfit idealized traces). The cumulative distribution of dwell times from a given state is now given to Origin 7.0 to infer the most likely parameters, asserting exponential decay. The inverse of the inferred time constant is the rate constant reported for that state.

6.3.4 Generating synthetic data

Synthetic data were generated in MATLAB. Rather than testing the inference on data generated precisely by the emissions model (one in which the scalar FRET signal is taken to be normally-distributed in each state), we challenge the inference by using a slightly more realistic distribution: one that is normally-distributed in each of the two fluorophore colors. That is, each synthetic trace was created from a hidden Markov model with 2D Gaussian output (representing the two fluorophore colors). The 2D data $\mathbf{x}_1, \mathbf{x}_2$ were then FRET transformed using $\mathbf{f} = \mathbf{x}_2/(\mathbf{x}_1 + \mathbf{x}_2)$; points such that $f \notin (0, 1)$ were discarded.

The 2D Gaussians are chosen so that, in any state z, the sum of the means is 1000 ($\mu_z^1 + \mu_z^2 = 1000\ \forall z$), roughly corresponding to our experimental data. Variances were drawn from a uniform distribution centered at each dimension's mean over a range given by 10% of the mean. The two components were allowed a nonzero covariance, also drawn from a uniform distribution centered at 0, with a range given by 1/2 the smaller of the two means. We emphasize that these choices are intended both to be consistent with the $\text{smFRET}_{\text{L1-L9}}$ and $\text{smFRET}_{\text{L1-tRNA}}$ data and *not* to match the algebraic expressions in the priors used below, which

would be a less challenging inference task (model specification identically matching the generative process).

Increasingly noisy traces were generated by multiplying the covariance matrix of each hidden state by a constant. Ten constants, chosen log-linearly between 1 and 100, were used. The mean standard deviation of the FRET state noise in the resulting 1D traces varied from, approximately, $0.02 < \sigma < 1.4$.

6.4 Priors

6.4.1 Mathematical expressions for priors

To calculate the model evidence, we treat the components of $\vec{\theta}$ as random variables. The vector $\vec{\pi}$ and each row of A are modeled as Dirichlet distributions:

$$p(\vec{\pi}) = \frac{\Gamma(\sum_{k=1}^{K} u_\pi^k)}{\prod_{k=1}^{K} \Gamma(u_\pi^k)} \prod_{k=1}^{K} \pi_k^{u_\pi^k - 1} \tag{6.1}$$

$$p(a_{j1}, ..., a_{jK}) = \frac{\Gamma(\sum_{k=1}^{K} u_a^{jk})}{\prod_{k=1}^{K} \Gamma(u_a^{jk})} \prod_{k=1}^{K} a_{jk}^{u_a^{jk} - 1} \tag{6.2}$$

The probabilities for each pair of μ_k and λ_k are modeled jointly as a Gaussian-Gamma distribution:

$$p(\mu_k, \lambda_k) = \sqrt{\frac{u_\beta^k \lambda_k}{2\pi}} e^{-\frac{1}{2} u_\beta^k \lambda_k (\mu_k - u_\mu^k)^2} \frac{1}{\Gamma(u_v^k/2)} (2u_W^k)^{-u_v^k/2} \lambda_k^{(u_v^k/2)-1} e^{-\frac{\lambda_k}{2u_W^k}}. \tag{6.3}$$

The terms $\vec{u}_\pi$, $\vec{u}_a$, $\vec{u}_\beta$, $\vec{u}_\mu$, $\vec{u}_v$, and $\vec{u}_W$ are called the *hyperparameters* for the probability distributions over $\vec{\theta}$.

6.4.2 Hyperparameter settings

Hyperparameters for vbFRET were set so as to give distributions consistent with experimental data and to influence the inference as weakly as possible: $u_\pi^k = 1$, $u_a^{jk} = 1$, $u_\beta^k = 0.25$, $u_m^k = 0.5$, $u_v^k = 5$ and $u_W^k = 50$, for all values of k. Qualitatively, these hyperparameter priors correspond to probability distributions over the hidden states such that it is most probable that the hidden states are equally likely to be occupied and equally likely to be transitioned to. Quantitatively, they yield $< \mu_k >= 0.5$ and typical $\sigma \approx 0.08$, consistent with experimental observation. $(1/\sqrt{\text{mode}(\lambda_k)} = 1/\sqrt{150} \approx 0.08\ \forall k)$.

6.4.3 Sensitivity to hyperparameter settings

One standard approach (McCulloch and Rossi 1991; Kass and Raftery 1995) to sensitivity analysis is to halve and double hyperparameters and recompute the evidence for different models. The sensitivity of ME inference on hyperparameter settings was investigated on both experimental and synthetic data. First, the two and three state traces from Fig. 5.2 and Fig. 6.7 were reanalyzed with all the hyperparameters set to one half their default values and twice their default values (Figs. 6.3, 6.4, 6.5, 6.6). One hyperparameter, the prior on the mean of each Gaussian, was not changed during this analysis, since its value is set to 0.5 based on a symmetry argument.

The results show a relative insensitivity to the hyperparameter values over the settings considered. The largest difference in inference accuracy between the

different settings was for the noisy, slow-transitioning traces shown in Fig. 6.6, when the hyperparameters were doubled. Interestingly, these traces are harder to resolve than the two state traces but not as difficult to resolve as the noisy, fast-transitioning three state traces. A possible explanation for this behavior is that the two state trace results are insensitive to hyperparameter settings because the data are easy enough to resolve and the noisy, fast-transitioning three state traces are insensitive to hyper parameter settings because they are too hard to resolve. The noisy, slow-transition states are on the border of being resolvable, so using a prior that more closely matches the true parameters of the model yields more accurate results. Additionally, the three state, slow-transition data has the highest probability of having a sparsely populated state (i.e. one that is only present for a few time steps in a trace). When σ is large, these sparsely populated states become harder to identify as distinct states, which may explain why $p(|\hat{\mathbf{Z}}| = |\mathbf{Z}_0|)$ decreases more than $p(\hat{\mathbf{Z}} = \mathbf{Z}_0)$, sensitivity or specificity .

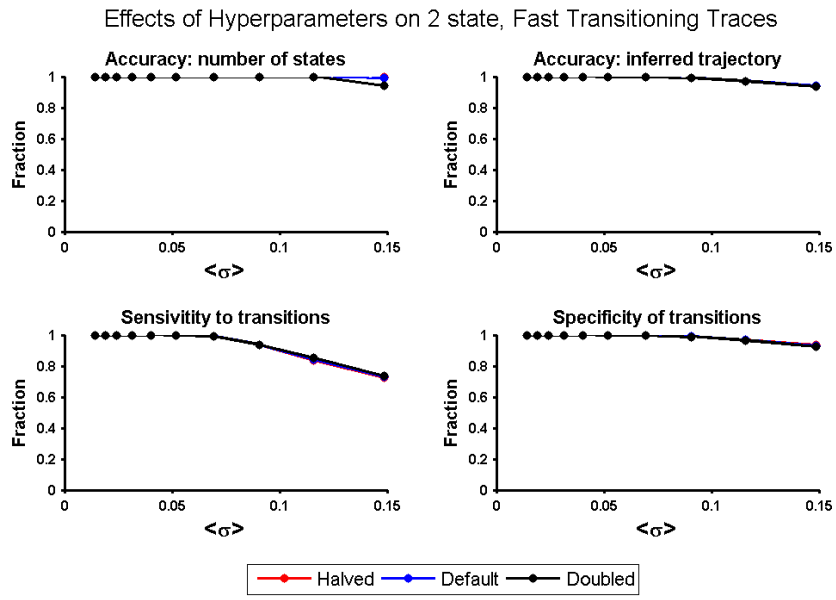

Figure 6.3: Effects of hyperparameter settings on fast-transitioning, two state traces.

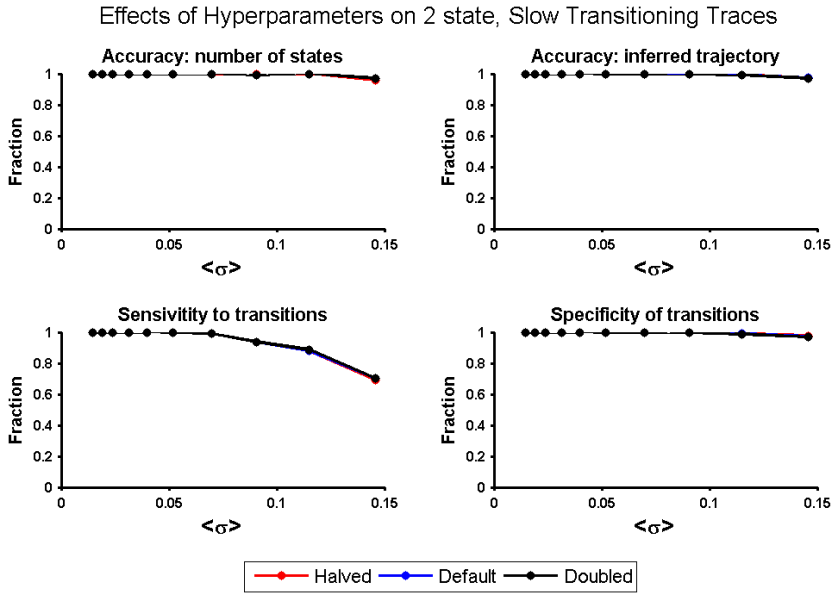

Figure 6.4: Effects of hyperparameter settings on slow-transitioning, two state traces.

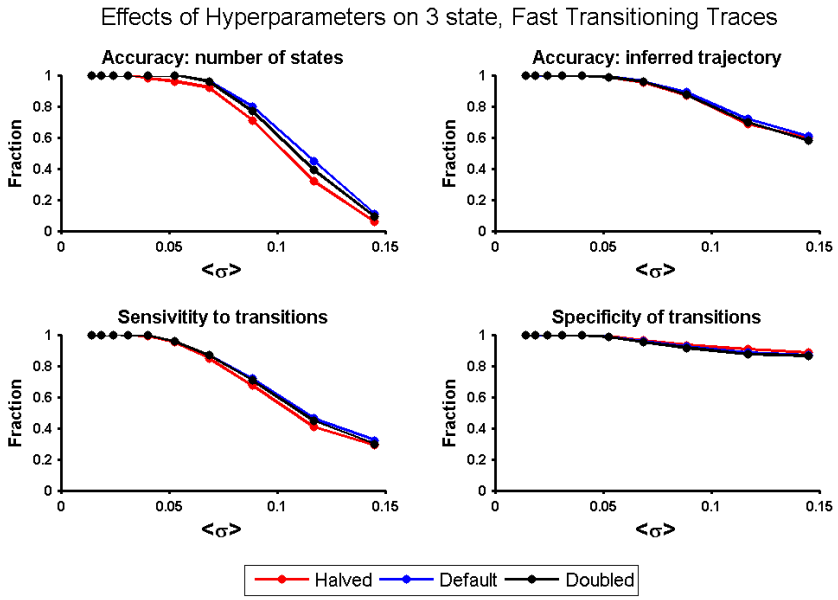

Figure 6.5: Effects of hyperparameter settings on fast-transitioning, three state traces.

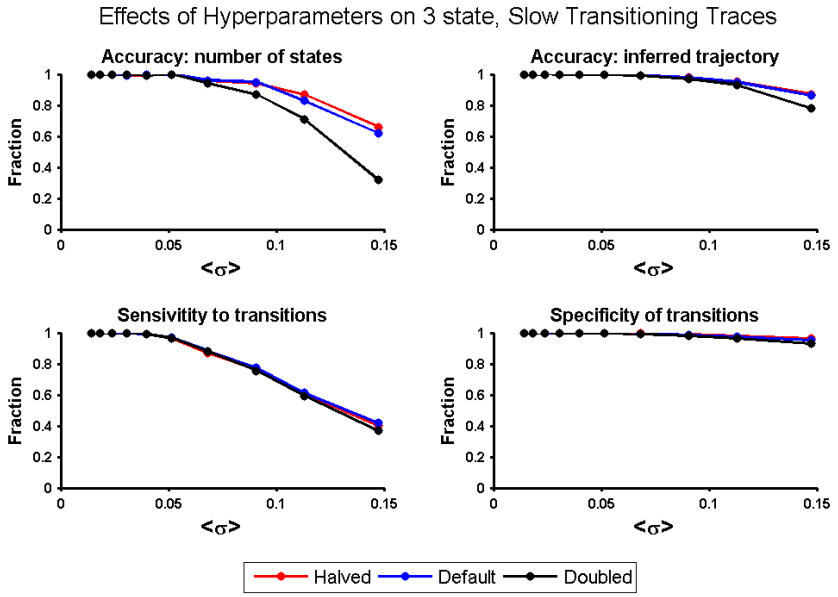

Figure 6.6: Effects of hyperparameter settings on slow-transitioning, three state traces.

To investigate further the effects of the hyperparameter settings on ME inference, the experimental data from Table 5.6 were reanalyzed using a more strongly diagonal transition matrix prior (Table 6.2). In this second prior, the diagonal terms of the transition matrix were set to 1 and the off-diagonal terms were set to 0.05, loosely corresponding to a prior belief that the ribosome was 10x more likely to remain in its current state than transition to a new one. For all of the data, the transition rates calculated with both hyperparameter settings are within error of each other for all transition rates.

Table 6.2: Effect of hyperparameters on transition rate inference.

Data set*	Settings	k_{close}	k_{open}
PMN_{Phe}[†]	Default	0.66 ± 0.05	1.0 ± 0.2
	Diagonal	0.66 ± 0.04	1.0 ± 0.2
PMN_{fMet}[‡]	Default	0.53 ± 0.08	1.7 ± 0.3
	Diagonal	0.52 ± 0.09	1.7 ± 0.1
$\text{PMN}_{\text{fMet+EFG}}$	Default	3.1 ± 0.6	1.3 ± 0.2
$(1\ \mu M)$[§]	Diagonal	2.8 ± 0.5	1.3 ± 0.1
$\text{PMN}_{\text{fMet+EFG}}$	Default	2.6 ± 0.6	1.5 ± 0.1
$(0.5\ \mu M)$[§]	Diagonal	2.6 ± 0.5	1.4 ± 0.1

* Rates reported here are the average and standard deviation from three or four independent data sets. Rates were not corrected for photobleaching of the fluorophores.

† PMN_{Phe} was prepared by adding the antibiotic puromycin to a post-translocation complex carrying deacylated-$\text{tRNA}^{\text{fMet}}$ at the E site and fMet-Phe-tRNA^{Phe} at the P site, and thus contains a deacylated-tRNA^{Phe} at the P site.

‡ PMN_{fMet} was prepared by adding the antibiotic puromycin to an initiation complex carrying fMet-$\text{tRNA}^{\text{fMet}}$ a the P site, and thus contains a deacylated-$\text{tRNA}^{\text{fMet}}$ at the P site.

§ 1.0 μM and 0.5 μM EF-G in the presence of 1 mM GDPNP (a non-hydrolyzable GTP analog) were added to PMN_{fMet}, respectively.

6.5 Synthetic validation – 2 and 4 state traces

Synthetic data for 2 FRET state traces (fast- and slow-transitioning, smFRET state means at 0.3 and 0.7 FRET) and 4 FRET state traces (fast-transitioning only, smFRET state means at 0.21, 0.41, 0.61 and 0.81 FRET) were generated and analyzed exactly as the traces in Fig. 5.2. The results are qualitatively similar to those in Fig. 5.2. Inference accuracy begins to decrease at a lower noise level as more FRET states are added to the traces. This should not be surprising, though, since the states are more closely spaced as the number of states increases, and therefore should be harder to resolve. Results for $K > 4$ state traces follow the same trend as those for $K = 2, 3, 4$ (data not shown).

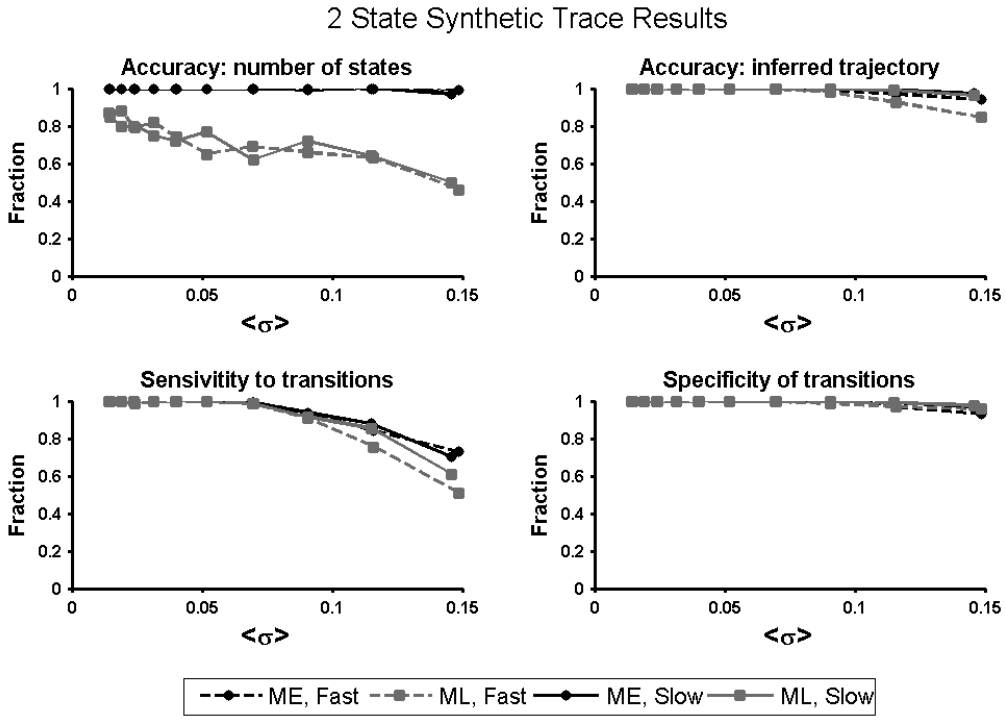

Figure 6.7: Synthetic results for two state traces.

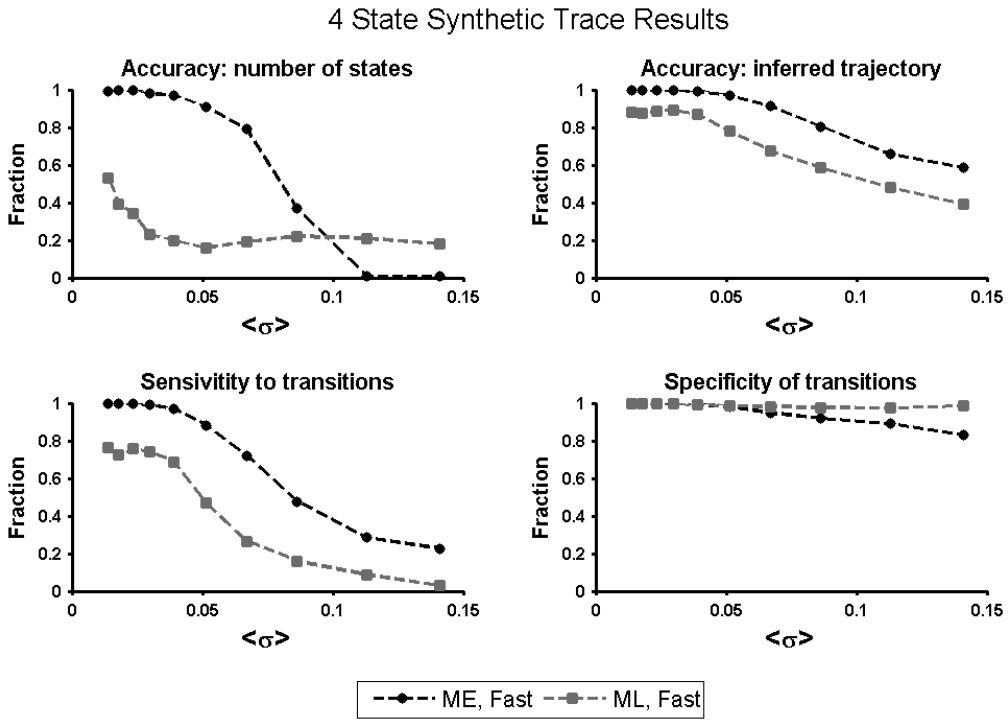

Figure 6.8: Synthetic results for four state traces.

6.6 Proof of variational relation

We provide a proof of the variational relation in Eq. 5.6. We start with the desired quantity, the evidence $p(\mathbf{D}|\mathrm{K})$, and multiply by one,

$$\ln p(\mathbf{D}|\mathrm{K}) \quad = \quad \left[\sum_{z} \int d\vec{\theta} q(\mathbf{Z}, \vec{\theta})\right] \ln p(\mathbf{D}|\mathrm{K}), \tag{6.4}$$

valid for any normalized probability distribution $q(\mathbf{Z}, \vec{\theta})$. We then use the definition of conditional probability to write

$$p(\mathbf{D}, \mathbf{Z}, \vec{\theta}|\mathrm{K}) = p(\mathbf{Z}, \vec{\theta}|\mathbf{D}, \mathrm{K}) p(\vec{\theta}|\mathrm{K}). \tag{6.5}$$

We use this to rewrite the argument of the logarithm and multiply by one yet again:

$$\begin{aligned}
\ln p(\mathbf{D}|\mathrm{K}) &= \sum_z \int d\vec{\theta} q(\mathbf{Z},\vec{\theta}) \ln \left[\frac{p(\mathbf{D},\mathbf{Z},\vec{\theta}|\mathrm{K})}{p(\mathbf{Z},\vec{\theta}|\mathbf{D},\mathrm{K})}\right] && (6.6)\\
&= \sum_z \int d\vec{\theta} q(\mathbf{Z},\vec{\theta}) \ln \left[\frac{p(\mathbf{D},\mathbf{Z},\vec{\theta}|\mathrm{K}) q(\mathbf{Z},\vec{\theta})}{q(\mathbf{Z},\vec{\theta}) p(\mathbf{Z},\vec{\theta}|\mathbf{D},\mathrm{K})}\right] && (6.7)\\
&= \sum_z \int d\vec{\theta} q(\mathbf{Z},\vec{\theta}) \ln \left[\frac{p(\mathbf{D},\mathbf{Z},\vec{\theta}|\mathrm{K})}{q(\mathbf{Z},\vec{\theta})}\right] \\
&\quad + \sum_z \int d\vec{\theta} q(\mathbf{Z},\vec{\theta}) \ln \left[\frac{q(\mathbf{Z},\vec{\theta})}{p(\mathbf{Z},\vec{\theta}|\mathbf{D},\mathrm{K})}\right], && (6.8)
\end{aligned}$$

where in the last line we have separated logarithm to decompose the integral into two parts. We recognize the rightmost term as the Kullback-Leibler divergence between $q(\mathbf{Z},\vec{\theta})$ and $p(\mathbf{Z},\vec{\theta}|\mathbf{D},\mathrm{K})$,

$$D_{KL}(q(\mathbf{Z},\vec{\theta})||p(\mathbf{Z},\vec{\theta}|\mathbf{D},\mathrm{K})) = \sum_z \int d\vec{\theta} q(\mathbf{Z},\vec{\theta}) \ln \left[\frac{q(\mathbf{Z},\vec{\theta})}{p(\mathbf{Z},\vec{\theta}|\mathbf{D},\mathrm{K})}\right] \quad (6.9)$$

and define the remaining term as the *free energy*,

$$F[q(\mathbf{Z},\vec{\theta})] = -\sum_z \int d\vec{\theta} q(\mathbf{Z},\vec{\theta}) \ln \left[\frac{p(\mathbf{D},\mathbf{Z},\vec{\theta}|\mathrm{K})}{q(\mathbf{Z},\vec{\theta})}\right], \quad (6.10)$$

which results in the variational relation presented in Eq. 5.6,

$$\ln p(\mathbf{D}|\mathrm{K}) = -F[q(\mathbf{Z},\vec{\theta})] + D_{KL}(q(\mathbf{Z},\vec{\theta})||p(\mathbf{Z},\vec{\theta}|\mathbf{D},\mathrm{K})). \quad (6.11)$$

This completes the proof of the variational relation and offers several insights.

The first is that the free energy is strictly bounded by the log-evidence, as the Kullback-Leibler (KL) divergence is a non-negative quantity, proven through an application of Jensen's inequality (an extension of the definition of convexity). Thus we have reduced the problem of approximating the evidence to that of finding the distribution $q(\mathbf{Z},\vec{\theta})$ which is "closest" to the true (and intractable) posterior

$p(\mathbf{Z}, \vec{\theta}|\mathbf{D}, \mathrm{K})$ in the KL sense. As per Eq. 6.11, we see that this is equivalent to minimizing the free energy $F[q(\mathbf{Z}, \vec{\theta})]$ as a functional of $q(\mathbf{Z}, \vec{\theta})$. This observation motivates the VBEM algorithm, in which a specific factorization for $q(\mathbf{Z}, \vec{\theta})$ is chosen as to make calculation of $F[q(\mathbf{Z}, \vec{\theta})]$ tractable (here, that $q(\mathbf{Z}, \vec{\theta}) = q(\mathbf{Z})q(\vec{\theta})$), and iterative coordinate ascent is performed to find a local minimum.

In addition, we provide motivation for the term "free energy", rewriting Eq. 6.10 by decomposing the logarithm:

$$F[q(\mathbf{Z}, \vec{\theta})] = -\sum_z \int d\vec{\theta} q(\mathbf{Z}, \vec{\theta}) \ln \left[\frac{p(\mathbf{D}, \mathbf{Z}, \vec{\theta}|\mathrm{K})}{q(\mathbf{Z}, \vec{\theta})} \right] \tag{6.12}$$

$$= -\sum_z \int d\vec{\theta} q(\mathbf{Z}, \vec{\theta}) \ln p(\mathbf{D}, \mathbf{Z}, \vec{\theta}|\mathrm{K}) + \sum_z \int d\vec{\theta} q(\mathbf{Z}, \vec{\theta}) \ln q(\mathbf{Z}, \vec{\theta}). \tag{6.13}$$

Recognizing the negative log-probability in the first term as an energy (as in the Boltzmann distribution) and the second term as the information entropy (Shannon 1948a;b) of $q(\mathbf{Z}, \vec{\theta})$, i.e.

$$E \equiv -\ln p(\mathbf{D}, \mathbf{Z}, \vec{\theta}|\mathrm{K}) \tag{6.14}$$

$$S \equiv -\sum_z \int d\vec{\theta} q(\mathbf{Z}, \vec{\theta}) \ln q(\mathbf{Z}, \vec{\theta}). \tag{6.15}$$

Thus we can rewrite 6.10 as

$$F = < E > -TS, \tag{6.16}$$

(with unit "temperature" T) where the angled brackets denote expectation under the variational distribution $q(\mathbf{Z}, \vec{\theta})$. This familiar form from statistical physics offers the following interpretation: in approximating the evidence (and posterior),

we seek to minimize the free energy by finding a distribution $q(\mathbf{Z}, \vec{\theta})$ that balances minimizing the energy and maximizing entropy.

We encourage the reader to enjoy the texts (MacKay 2003) and (Bishop 2006) for more pedagogical discussions of variational methods.

Chapter 7

hFRET

7.1 Abstract

Single-molecule fluorescence resonance energy transfer (smFRET) data presents the opportunity to learn the number of conformational states explored by a molecule as well as the rate constants governing transitions between these states. While good methods exist to analyze individual time series, how best to aggregate results from the many time series collected in an experiment to learn a high-confidence consensus model remains a challenge. We here present a statistical method which analyzes an entire corpus of time trajectories, using hierarchical modeling, with inference performed via an optimization method we term "ensemble variational Bayes". We demonstrate superior inference accuracy over methods which analyze individual time series and show that it is possible to detect and model sub-populations of traces within a data set which possess identically valued smFRET states, but differ in the transition rates between these states.

7.2 Introduction

Single-molecule fluorescence resonance energy transfer (smFRET) provides an indirect probe to explore structure and dynamics at the molecular level. The efficiency of non-radiative energy transfer from a fluorescing donor molecule to a fluorescing acceptor molecule is highly distant dependent in the range of $\sim 1-10$ nm, making it possible to use the relative fluorescent intensities of the donor/acceptor to report on their nanoscale movements. When attached to an individual protein, nucleic acid or biomolecule, the donor/acceptor can be used to report on the dynamics of the molecular complex. Diverse processes such as the crossing of Holliday junctions (Hohng et al. 2004), the conformational dynamics of individual proteins in vivo (Sakon and Weninger 2010) and the marching of motor proteins on microtubules (Mori et al. 2007) have been studied via smFRET.

Typically, the donor and acceptor fluorescent intensities (I_D and I_A, respectively) observed are converted into a 1D summary static, the "FRET ratio", given by FRET $= I_A/(I_D + I_A)$. This dimensionless ratio, traditionally quoted in "units" of FRET, is analyzed as a function of time in smFRET traces. In many smFRET studies, the molecule of interest transitions between a series of locally stable conformations (*i.e.* states). From the smFRET time series, it is possible to infer (1) the number of states the molecule occupies and (2) the transition rates between these states.

To avoid the tedium and subjectivity of manual analysis, several smFRET analysis software packages have been developed. QuB (Qin et al. 1997; 2000),

originally developed for ion channel analysis, has been adapted for smFRET analysis. HaMMy (McKinney et al. 2006) and, most recently, our vbFRET software (Bronson et al. 2009) were developed specifically to analyze smFRET. All of these programs model the smFRET time series as a hidden Markov model (HMM) with Gaussian observables, which treats time discretely and assumes that at every time step:

1. The system is in one of K conformations (states). The identity of the conformation is hidden from the observer. (The hidden state at time t will be denoted by z_t.)

2. There is an observable, d_t. The $p(d_t)$ is a Gaussian with parameters $\{\mu_k, \sigma_k\}$ which are a function of z_t.

3. After producing d_t the system either transitions to a new state or remains in its current state. The probability of transitioning is also a function of z_t.

These assumptions are appropriate for most smFRET experiments since: the observed FRET signal is a function of the hidden conformation of the molecule and has roughly Gaussian noise; the probability of adopting a new molecular conformation is a function of the molecule's current state; the molecule transitions between a finite number of locally stable conformations; and the CCD camera commonly used in smFRET studies naturally time bins the data.

Both QuB and HaMMy solve the HMM using the principle of maximum likelihood (ML) and the expectation maximization algorithm (EM). ML seeks to find the parameters, $\vec{\theta}_*$, which maximize the probability of the data ($\mathbf{D}$) given the

parameters ($\vec{\theta}$) and model (K) ($p(\mathbf{D}|\vec{\theta}, \mathrm{K})$, also termed the *likelihood*):

$$\vec{\theta}_* = \underset{\vec{\theta}}{\operatorname{argmax}}\, p(\mathbf{D}|\vec{\theta}, \mathrm{K}), \tag{7.1}$$

where K is a HMM with Gaussian observables and K states. Although often effective, ML suffers from two pathologies: (1) ML has a strong propensity to overfit data (*i.e.* find more states than are supported by the observed data) and (2) ML can find divergent solutions to the HMM (i.e. the algorithm can converge to solutions where a FRET state has zero variance and infinite likelihood, rendering the analysis worthless) which must be detected and avoided using some form of user defined algorithm.

The vbFRET software package solves the HMM using the principle of maximum evidence (ME) and the variational Bayesian expectation maximization algorithm (VBEM). ME can be thought of as ML for model selection. It seeks to find K_*, the model complexity which maximizes the evidence, $p(\mathbf{D}|\mathrm{K})$:

$$\mathrm{K}_* = \underset{\mathrm{K}}{\operatorname{argmax}}\, p(\mathbf{D}|\mathrm{K}) \tag{7.2}$$

Unlike the likelihood, the evidence peaks for the model with the highest probability of having the correct number of states. In addition, calculation of $p(\mathbf{D}|\mathrm{K})$ via VBEM also returns the posterior parameter distribution ($p(\vec{\theta}|\mathbf{D}, \mathrm{K})$), the parameters learned from the data. From the evidence, it is possible to determine how many states are in the data. From the posterior, it is possible to learn transition rates between states. Because an entire distribution over $\vec{\theta}$ is learned rather than a point estimate for $\vec{\theta}$, ME cannot converge to divergent solutions to the HMM. Note $\vec{\theta}_*$ is easily learned from this posterior by finding $\max_{\vec{\theta}} p(\vec{\theta}|\mathbf{D}, \mathrm{K})$. This $\vec{\theta}_*$,

however, will not have divergent parameter values (Bishop 2006). For some data sets, the use of ME rather than ML substantially improves data inference (Bronson et al. 2009). A less cursory discussion of ME and ML can be found in (MacKay 2003; Bishop 2006)and, with emphasis on smFRET data, (Bronson et al. 2009).

All of these analysis programs suffer from one major shortcoming, however: they can only analyze individual time series. While it is trivial for a human to look at two similar traces with slightly different photophysical parameters (*e.g.*, slightly different FRET means or variances), and recognize that they are reporting on the same process, these programs cannot. All of these models assume the data are identically distributed, both within a trace and between traces, so they can only be used to analyze multiple traces at once when the model parameters are identical from one trace to another. All real experiments have some trace to trace variation of their model parameters, so this *ensemble inference* (concurrent inference of multiple time series) is impossible with existing methods.

Undoubtedly more information exists in the ensemble of traces than in any individual trace (if for no other reason, far more data is present in the ensemble), so confining inference to individual traces necessarily lowers the potential ability of these programs to perform inference. For well resolved data it may not matter which way the analysis is performed. For marginal data, or data where some information is present only in the ensemble, it would be a substantial advantage to analyze the ensemble all at once. In particular, the presence of sub-populations of traces within a data set with similar smFRET states but different transition rates between those states would only be detectable by analyzing the ensemble.

This situation might arise if an inhibitor were present in the experiment in sub-saturating concentrations. Some molecules would be inhibitor bound and some would not, creating "fast transitioning" and "slow transiting" traces.

When inference is performed on individual traces, the results must be combined in a post-processing step to learn a single, high-confidence consensus model describing the transition rates between states. Current inference methods learn a transition matrix (A) for each time series, which contains the transition probability from every state to every state in the time series. The j^{th} row and k^{th} column of A ($a_{j,k}$) holds the probability of transitioning from state j to state k (*i.e.*, $a_{j,k} = p(z_t = k|z_{t-1} = j)$). Consequently all the information about the rate constants for the trace are contained in A as well. (The interconversion between A and rate constants is straight forward and described in Sec. 7.4.2.) Combining the individual A_ns into a consensus A for the data set is non-trivial and rarely done. Instead, rate constants are typically learned via dwell-time analysis, a multi-step procedure which requires subjective post processing, is sensitive to outlying data points and systematically overestimates transition rates.

Here we describe a method for smFRET inference which can perform ensemble inference to learn a single consensus model from an entire data set. The method is based on *hierarchical modeling*, so we term it "hFRET". It uses an optimization similar to variational Bayes, which we term "ensemble variational Bayes". Using synthetic data, we show the statistical superiority of hFRET for data inference over both ML and ME and show that rate constants extracted directly from an ensemble inferred transition matrix more accurate than rate constants learned from dwell

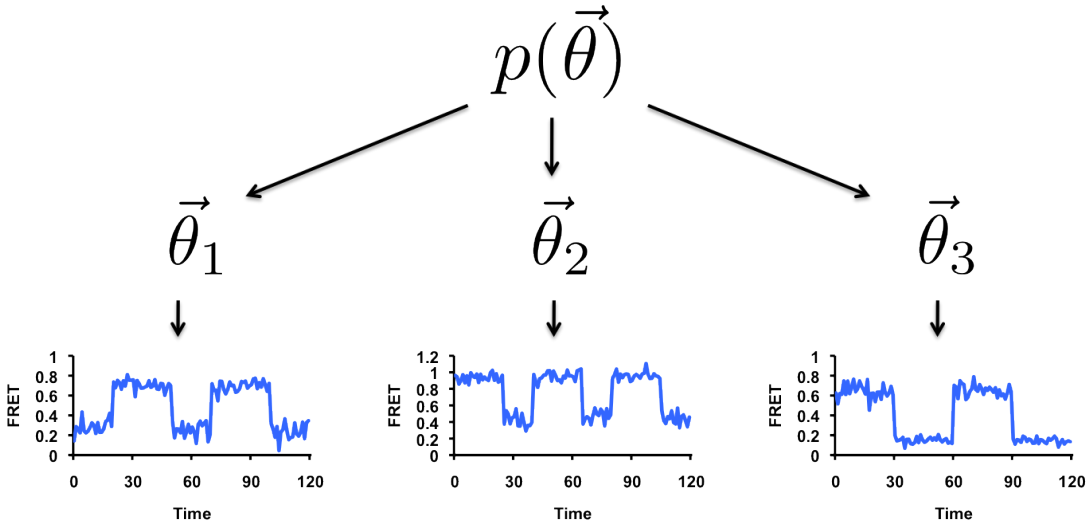

Figure 7.1: Cartoon model showing the relationship among individual smFRET traces. Each individual trace has a specific set of model parameters, $\vec{\theta}_n$, drawn from a distribution of parameters, $p(\vec{\theta}_N)$. The distribution $p(\vec{\theta}_N)$ describes the full ensemble of parameters for the data set, from which any individual trace's parameters can be drawn, meaning that $p(\vec{\theta}_N)$ is a complete description of the experimental system. To our knowledge, all previous inference methods only learn the individual $\vec{\theta}_n$. The method introduced here learns $p(\vec{\theta}_N)$ as well.

time analysis. Finally we show that hFRET can be used to learn when data sets contain sub-populations of traces within the data, using both synthetic data and experimental smFRET data taken from the ribosome. This is, to our knowledge, the first example of a biophysical time series inference method which is capable of performing ensemble inference on real data.

7.3 The model

In order to perform ensemble analysis, hFRET assumes the following model for the data.

1. The data set comprises an ensemble of traces reporting on the same biophysical process.

2. Each trace is modeled as a HMM with Gaussian observables. The parameters for each trace ($\vec{\theta}_n$) describe the mean ($\vec{\mu}$) and variance ($\vec{\sigma^2}$) of each state, transition probabilities to other states (A) and the probability that the time series began in each state ($\vec{\pi}$).

3. Within a given trace, $\vec{\theta}_n$ is fixed (*i.e.*, the means and variances of states are fixed within a trace).

4. The values in $\vec{\theta}_n$ vary slightly from trace to trace.

5. The trace to trace variance of $\vec{\theta}_n$ can be described by an "ensemble probability distribution", $p(\vec{\theta}_N)$.

Departing from the earlier notation of this thesis, the dependence on model complexity K will now be omitted from equations for clarity of presentation. There are several points worth mentioning about this framework. First, the ensemble does not have to comprise only one process. It could also comprise a small number of distinct processes, although this is a more challenging inference problem. (The model must then learn which traces report on which processes in addition to modeling the processes). Second, this framework can easily be adapted to data which is not described by an HMM by changing the model type in #2. Third, assumption #3 needs to be carefully considered for each data set. The hidden states of the system model the locally stable conformations of the molecule. Every

time the system returns to a locally stable conformation, it might be a slightly different conformation resulting in a slightly different smFRET signal. For many data sets these fluctuations are negligible, but not for all data sets. It should be noted that this is a problem for all smFRET analysis methods. Small shifts of a smFRET states mean intensity within a trace are currently corrected via manually set thresholds in a post-processing smoothing procedure prior to dwell-time analysis. As discussed later, hFRET obviates the need for much of this subjective post-processing, which makes the inference more objective and reliable, but also more susceptible to data which deviate from the model. Fourth, it is assumption #5 which makes this model hierarchical; by definition, a hierarchical model is one where parameters are drawn from specified probability distributions (Gelman and Hill 2006, p.2). The $p(\vec{\theta}_N)$ used here comprises a Gaussian-Gamma distribution for each $\{\mu_k, \sigma_k\}$ and a Dirichlet distribution for each row of A.

In this model, the ensemble distribution $p(\vec{\theta}_N)$ provides the fullest possible description of the whole data set. Knowledge of $p(\vec{\theta}_N)$ also facilitates inference of individual traces since, according to Bayes' rule Sec. 4.3.1, the posterior of an individual trace is then given by

$$p(\vec{\theta}_n|\mathbf{D}_n) = \frac{p(\mathbf{D}_n|\vec{\theta}_n)p(\vec{\theta}_N)}{p(\mathbf{D}_n)}, \tag{7.3}$$

where $\mathbf{D}_n$ denotes the data of the n^{th} trace. The use of $p(\vec{\theta}_N)$ for the prior in Eq. 7.3 follows from postulate #5 of the hFRET model. Exact calculation of $p(\vec{\theta}_n|\mathbf{D}_n)$ is impossible for the HMM, but an accurate estimate can be efficiently calculated via

VBEM (Bronson et al. 2009). From $p(\vec{\theta}_n|\mathbf{D}_n)$, $\vec{\theta}_n$ can be estimated using

$$\vec{\theta}_n = \max_{\vec{\theta}_n} p(\vec{\theta}_n|\mathbf{D}_n). \tag{7.4}$$

If $\vec{\theta}_n$ for each trace ($\{\vec{\theta}_n\}_N$) were known, a maximum likelihood estimate of $p(\vec{\theta}_N)$ could be obtained by using $\{\vec{\theta}_n\}_N$ to calculate a mean and variance over each parameter in $\vec{\theta}$, and setting $p(\vec{\theta}_N)$ to have those sufficient statistics. In the limit of an infinite number of traces, this estimate would be exact.

The dependency of $p(\vec{\theta}_N)$ on the set of $p(\vec{\theta}_n|\mathbf{D}_n)$ and the dependency of $p(\vec{\theta}_n|\mathbf{D}_n)$ on $p(\vec{\theta}_N)$ suggests an iterative algorithm, which we term *ensemble variational Bayes*: guess $p(\vec{\theta}_N)$, use it to find $p(\vec{\theta}_n|\mathbf{D}_n)$ for each trace via VBEM, get $\{\vec{\theta}_n\}_N$ using Eq. 7.4, use $\{\vec{\theta}_n\}_N$ to calculate the sufficient statistics needed to learn a new $p(\vec{\theta}_N)$. Unlike EM and VBEM, convergence is not guaranteed for this iterative procedure so a criterion is needed for when to stop this iterative process. Additionally, it is necessary to have a criterion for evaluating the quality of $p(\vec{\theta}_N)$ so that it is possible to chose among competing models (*i.e.* different values of K or different $p(\vec{\theta}_N)$s learned from different initializations). We assert that such a criterion should be that *the best* $p(\vec{\theta}_N)$ *is the one which maximizes the model's evidence*:

$$p(\mathbf{D}) = \sum_N p(\mathbf{D}_n) = \sum_N \int d\vec{\theta} p(\mathbf{D}_n|\vec{\theta}_n) p(\vec{\theta}_N). \tag{7.5}$$

Based on this criterion, the above procedure should iterate until the model's evidence either peaks or converges.

The algorithm used by hFRET has many similarities to ME. The ensemble distribution, $p(\vec{\theta}_N)$, is algebraically equivalent to a prior, which is why it is possible

to calculate $p(\vec{\theta}_n|\mathbf{D}_n)$ via VBEM. The $p(\vec{\theta}_N)$ is technically not a prior, however, since it is learned rather than asserted. This is a fundamental difference between ME and the hFRET algorithm. The posterior learned for each trace, $p(\vec{\theta}_n|\mathbf{D}_n)$, still allows one to infer model parameters and generate idealized traces. The evidence calculated here can still be used for model selection and the optimization still naturally suppresses unnecessarily populated states (*e.g.*, if a specific trace in a K = 3 data set only has two states populated, the optimization will leave one state unpopulated in the inference of that trace). This feature is especially important since the majority of data sets do not have every smFRET state populated in each trace. One caveat is that it is possible to overfit using the hFRET algorithm by choosing a pathological prior (i.e. if $p(\vec{\theta}_N)$ includes a state k such that $p(\vec{\theta}_{n,k})$ is a delta function for a Gaussian of zero variance and mean centered directly on a data point). Such initializations are extremely unrealistic and, in our experience, difficult to achieve even when done intentionally (data not shown).

7.4 Validating the model

Synthetic data was used to test the performance of hFRET. Unlike real data, the true hidden states of synthetic data are known, so its is possible to measure inference accuracy. Generation of synthetic data was performed as previously described (Bronson et al. 2009). Since the noise in the donor/acceptor fluorescence intensities of real smFRET data is more likely Gaussian than the noise of the 1D FRET transform, synthetic traces were generated from a 2D Gaussian model and FRET

transformed. This procedure creates more realistic synthetic data and avoids testing the inference algorithm's performance on data generated by the exact emissions model (one in which the scalar FRET signal is taken to be normally distributed in each state).

7.4.1 Increasingly noisy data

First, a synthetic data set previously analyzed by ME and ML was analyzed by hFRET to compare its performance (Bronson et al. 2009). In this data set all traces have K = 3 states with means centered at $\mu_z = (0.25, 0.5, 0.75)$. Every state can transition to every other state with equal probability. For half the traces the mean lifetime of a state is 15 ± 5 time steps (slow transitioning) and for half the data the mean lifetime of a state is 4 ± 2 time steps. All smFRET states in a trace have one of ten different noise levels, ranging from $\sigma = 0.02 - 0.15 \pm 10\%$ (unrealistically noisless to unrealistically noisy, given the separation of states). Trace length, T, varied from $50 \leq T \leq 500$ time steps, drawn randomly from a uniform distribution. For each of the 10 noise levels and two transition speeds, 100 traces were generated (2000 traces in total). All traces were analyzed individually by ME and ML. For hFRET, each combination of σ and transition rate was treated as a "data set" and analyzed collectively (20 data sets total).

Before comparing the performance of hFRET to ME and ML, the algorithm's ability to perform model selection and converge to a $p(\vec{\theta}_N)$ was assessed. Each data set was fit with $K = 1 - 5$ states and ten rounds of ensemble inference was use to learn $p(\vec{\theta}_N)$. For all 20 data sets evidence was largest for the K = 3 model and all

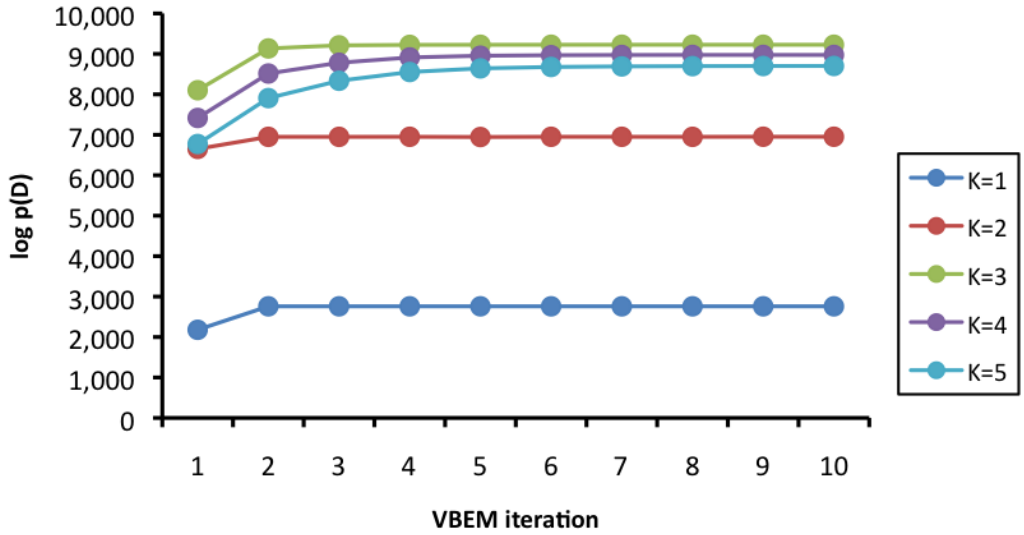

Figure 7.2: Convergence of $\log p(\mathbf{D})$ during ensemble variational Bayesian inference. Inference was performed for K $= 1 - 5$. The evidence as a function VBEM iteration is plotted. The evidence for K $= K_{true} = 3$ is the largest, illustrating the value of $\log p(\mathbf{D})$ for model selection. For all values of K the evidence quickly approaches an asymptotic value.

showed convergence of their evidence before the 10^{th} round of ensemble inference (convergence was typically observed after 2 or 3 iterations). To illustrate this result the log(evidence) versus ensemble inference iteration number for one data set, fast transitioning traces with $\sigma = 0.09$, is show in Fig. 7.2.

As before the Viterbi path (idealized trace) was learned for each trace and its accuracy was assessed via four probabilities: (1) accuracy in number of states $p(|\hat{\mathbf{Z}}| = |\mathbf{Z}_0|)$: the probability in any trace of inferring the correct number of states (where $|\mathbf{Z}_0|$ is the number of states in the model generating the data and $|\hat{\mathbf{Z}}|$ is the number of populated states in the idealized trace); (2) accuracy in state identity $p(\hat{\mathbf{Z}} = \mathbf{Z}_0)$: the probability in any trace at any time of inferring the correct state;

(3) sensitivity to true transitions: the probability in any trace at any time that the inferred trace $\hat{\mathbf{Z}}$ exhibits a transition, given that $\mathbf{Z}_0$ does; and (4) specificity of inferred transitions: the probability in any trace at any time that the true trace $\mathbf{Z}_0$ does not exhibit a transition, given that $\hat{\mathbf{Z}}$ does not.

For all probabilities except specificity of transitions, hFRET performed as well or better than ME and ML. (On difficult inference problems a high specificity score is often a sign that no inference is being performed, since a trace with no transitions always has 100% specificity. Inspection of the inference here showed high specificity of ME&ML could be explained by underfitting of transitions.) The biggest improvements of hFRET over ME/ML were on fast transitioning data, where inference is harder. For the fast transitioning $\sigma = 0.15$ data the accuracy of inferred trajectory for hFRET, ME and ML were 0.77, 0.61 and 0.37, respectively, and the sensitivity to transitions for hFRET, ME and ML were 0.50, 0.32 and 0.03, respectively. The largest performance improvement from inference via hFRET was for accuracy in the number of states, where hFRET, ME and ML were 0.99, 0.45 and 0.25 on the $\sigma = 0.15$ data. The large improvement of hFRET on this last metric should not be surprising. Once the program determines that the data set contains three states, it is substantially easier to determine that individual traces contain three states as well.

7.4.2 Learning a transition matrix

All this above analysis requires the traces to be idealized by the inference method. While idealized traces can be useful visually and are necessary for dwell-time anal-

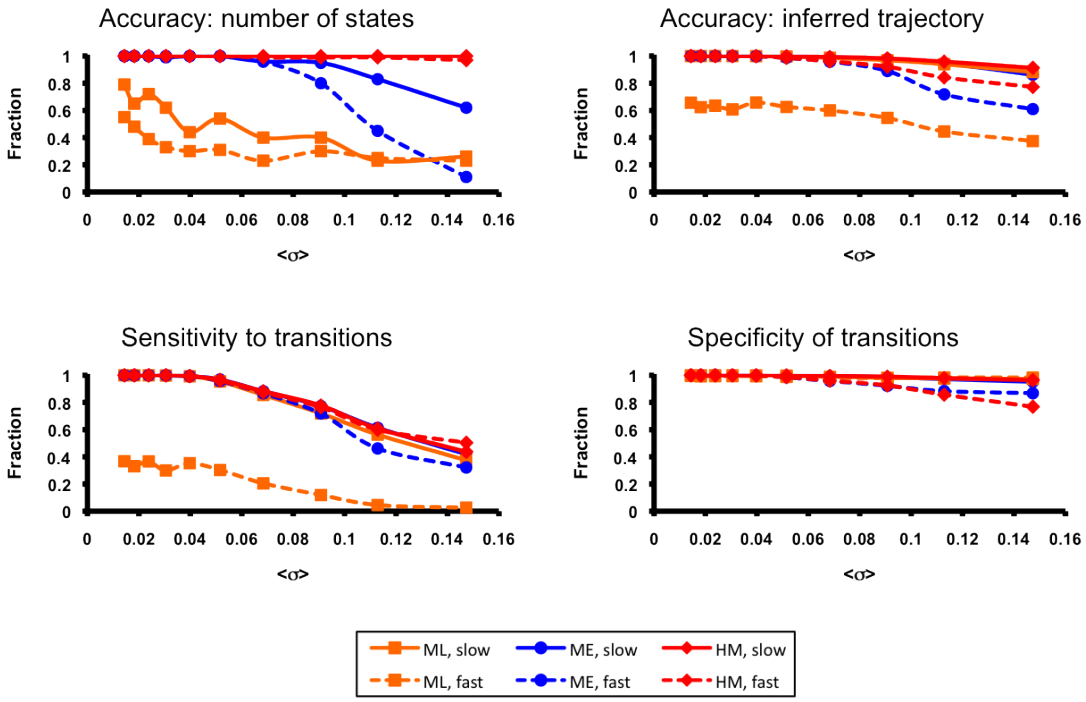

Figure 7.3: Comparison of hFRET inference (red) to ME (blue) and ML (orange) inference as a function of increasing hidden state noise. Fast transitioning (dashed line) and slow transitioning (solid line) synthetic, 3-state data (described in the text and in (Bronson et al. 2009)) were analyzed by the methods. The idealized trajectories inferred were compared to the true trace trajectories and used to assess accuracy using four metrics: (1) accuracy in number of states (probability of inferring the correct number of states in a trace), (2) accuracy in idealized trajectory (probability at each time step in a trajectory of inferring the correct state) (3) sensitivity to transitions (probability of inferring a transition given that one occurs) and (4) specificity of transitions (probability of inferring no transition has occurred given that none occurs).

ysis, the ultimate goal of analysis is to learn rate constants. Rate constants can be learned from dwell-time analysis or extracted directly from a transition matrix. The entry of the j^{th} row and k^{th} column of A ($a_{j,k}$) holds the probability of transitioning from state j to state k. The diagonal of A holds the probability of remaining in the systems current state. For a state k with rate constant r_k, the probability

of remaining in state k for t or more seconds is $1 - \exp(-r_k t)$. According to the transition matrix, the probability remaining in state k for t or more seconds is given by $1 - (a_{k,k})^t$. Equating the two converts $a_{k,k}$ to r_k:

$$r_k = -\ln a_{k,k}. \tag{7.6}$$

Estimates of rate constants extracted from a transition matrix should be more accurate and robust than estimates learned from dwell-time analysis. In dwell-time analysis, data preceding the first transition and following the last transition are ignored, so not all the data available is for inference (Fei et al. 2009). In addition, this procedure effectively truncates every trace after its last transition. Virtually all traces will have non-transitioning data, which is not factored into the analysis, systematically inflating transition rates. This is especially problematic for slow transitioning data where zero or one transitions are present in some traces and entire traces can be omitted from the analysis process. Dwell-time analysis is also susceptible to overfitting data containing a few (non-representative) long lifetime events. Finally, the modeler must assume the transition rate between states conforms to single our double exponential decay.

Each entry of the transition matrix $a_{j,k}$ is calculated by summing the probability over each pair of time steps that the system was in state j and transitioned to state k, and then normalizing the rows of the transition matrix:

$$a^{\dagger}_{j,k} = \sum_{n=1}^{N} \sum_{t=1}^{T_n - 1} p(z_{t+1} = k | z_t = j) \tag{7.7}$$

$$a_{j,k} = \frac{a^{\dagger}_{j,k}}{\sum_{k=1}^{K} a^{\dagger}_{j,k}} \tag{7.8}$$

where $a^{\dagger}_{j,k}$ is the unnormalized precursor to $a_{j,k}$, n is an index over traces and t is an index over time steps. This procedure uses all the available data, can handle any number of transitions per trace and is less sensitive to non-representative long lifetime events. As shown in Sec. 7.4.3, evidence can often be used to distinguish between single and double exponential decay processes.

The performance of dwell-time analysis versus transition matrix inference was assessed on a set of 8 data sets with these characteristics:

$$A_0 = \begin{pmatrix} 0.88 & 0.06 & 0.06 \\ 0.08 & 0.80 & 0.12 \\ 0.19 & 0.17 & 0.64 \end{pmatrix}$$

$\mu_z = (0.25, 0.50, 0.75) \pm\ 0.1$

$\sigma = (0.090.100.09) \pm\ (0.01, 0.02, 0.02)$

Each data set had a total of 20,000 time steps of data, but spread over traces of length $\{500, 250, 150, 100, 75, 50, 25, 10\}$ (i.e. the inference problem was made more difficult by holding all things constant but trace length). The data were analyzed by ME and ML as well. Only dwell-time analysis was performed with these methods, since they do not lend themselves to learning a transition matrix from the data. Accuracy was assessed by comparing the Kullback-Leibler divergence between A_0 and A inferred $D_{KL}(\mathrm{A}_0||\mathrm{A}_{\mathrm{inf}})$. The Kullback-Leibler divergence is a common dissimilarity metric in information theory. The more similar $\mathrm{A}_{\mathrm{inf}}$ and A_0, the smaller $D_{KL}(\mathrm{A}_0||\mathrm{A}_{\mathrm{inf}})$. Accuracy of rate constants learned by dwell-time analysis was assessed by converting the rate constants into a transition matrix[1]

[1]Each r_k can be converted to an $a_{k,k}$ using Eq. 7.6. The off diagonal terms are then found by setting each $a_{j,k}$ proportional to the number of transitions from j to k such that the probability

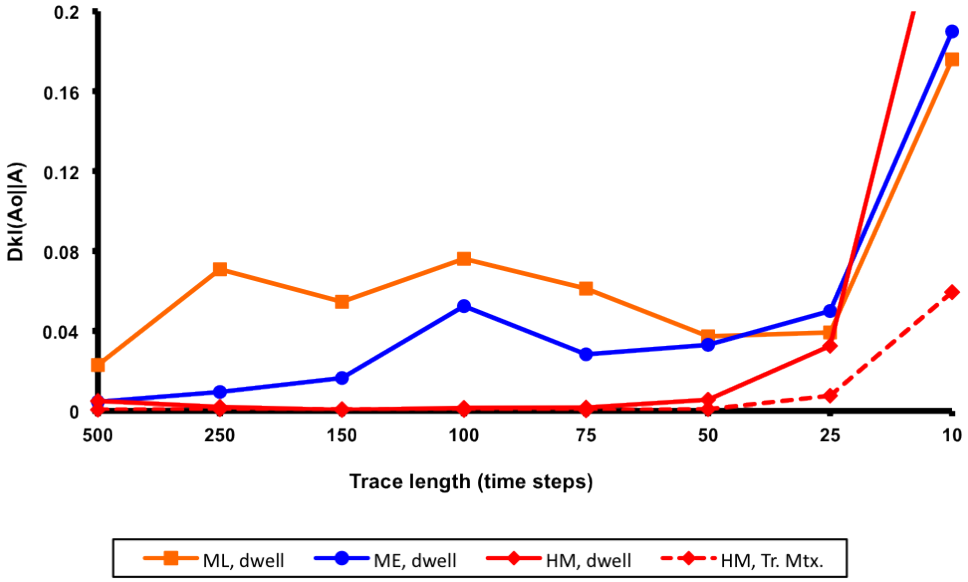

Figure 7.4: The D_{KL} between A_0 and A_{inf} for data sets of increasingly short trace length (all other parameters held constant). The D_{KL} between A_0 and A_{inf} learned by hFRET (red, dashed), learned by dwell-time analysis based on hFRET inference (red, solid), dwell-time analysis based on ME inference (blue) and dwell-time analysis based on ML inference (orange) are shown for each data set.

and then computing $D_{KL}(A_0||A_{inf})$. As seen in Fig. 7.4, transition matrix inference outperforms dwell-time analysis, especially on shorter traces. Inference accuracy for these data sets, as judged by the four probabilities discussed in Sec. 7.4.1, is shown in Sec. 7.7.4.

7.4.3 Learning a mixture of models

Arguably the most exciting advantage of ensemble learning is the ability to detect the presence of sub-populations within a data set whose smFRET states share the same photophysical parameters and differ only in transition rates between states.

distribution is normalized.

Such data sets can arise in many situations, such as protein dynamics studies where some proteins are bound to an inhibitor and some are not. The presence of the inhibitor might not change the smFRET states, but could significantly alter the transition rates between these states. In RNA folding studies, it is common to have different molecules fold via different pathways (Zhuang et al. 2000). Here too, the smFRET states observed are the same from trace to trace, but the kinetic pathways can differ.

A data set where every trace is governed by the same kinetics and a data set containing sub-populations with different kinetics (i.e. fast transitioning and slow transitioning traces) are described by different models. Consider a two state system with smFRET states at $\{\mu_1, \mu_2\}$ and transition probabilities $\{a_{12}, a_{21}\}$. The transition matrix for this system is:

$$A_0 = \begin{pmatrix} 1\text{-}a_{12} & a_{12} \\ a_{21} & 1\text{-}a_{21} \end{pmatrix}.$$

Now consider a system that has the same two smFRET states, but half the traces are fast transitioning, with transition probabilities $\{a_{12}, a_{21}\}$ and half are slow transitioning, with transition probabilities $\{a_{34}, a_{43}\}$. This data set is actually a four state system, with smFRET state means equal to $\{\mu_1, \mu_2, \mu_1, \mu_2\}$ and a transition matrix

$$A_0 = \begin{pmatrix} 1\text{-}a_{12} & a_{12} & 0 & 0 \\ a_{21} & 1\text{-}a_{21} & 0 & 0 \\ 0 & 0 & 1\text{-}a_{34} & a_{34} \\ 0 & 0 & a_{43} & 1\text{-}a_{43} \end{pmatrix}$$

where the block off diagonal zeros indicate that each trace can either transition between states one and two or it can transition between states three and four, but no transitions between the slow transitioning states and the fast transitioning states are allowed. (If the data switched between fast and slow transitioning states within a single trace then theses terms would be non-zero.) The HMM variable $\vec{\pi}$, which stores the probability that the time series began in the k^{th} state, can be used to learn the fraction of slow transitioning and fast transitioning traces in the data. Normally $\vec{\pi}$ has no biophysical significance.

Because these two scenarios are described by different models it is, in principle, possible to used the models' evidence to discriminate between the two scenarios. A set of data set was designed to test hFRET's ability to perform this selection in practice. Each data set comprised 100 traces, each containing 4 states with $\mu_z = (0.3, 0.7, 0.3, 0.7)$ and $\sigma = 0.075 \pm 0.0075$ for each state. Trace length, T, was allowed to be either $\{100, 200, 300, 400, 500, 600\}$ for each trace in the data set. The transition rates of each data set were governed by

$$A_0 = \begin{pmatrix} \text{a} & \text{1-a} & 0 & 0 \\ 0.05 & 0.95 & 0 & 0 \\ 0 & 0 & 0.95 & 0.05 \\ 0 & 0 & 0.05 & 0.95 \end{pmatrix}.$$

The value of a was allowed to range from 0.5 to 0.95 in increments of 0.05 for each value of T (a total of 60 data sets). All traces only have two smFRET states, one at 0.3 FRET (f_{low}) and one at 0.7 FRET (f_{high}). When $a = 0.5$, the data set consists of two distinct types of traces: traces which quickly transition from

$f_{\text{low}} \rightarrow f_{\text{high}}$ and slowly from $f_{\text{high}} \rightarrow f_{\text{low}}$, and traces which transition slowly from both $f_{\text{high}} \rightarrow f_{\text{low}}$ and $f_{\text{low}} \rightarrow f_{\text{high}}$. As a gradually increases, the transition from $f_{\text{high}} \rightarrow f_{\text{low}}$ becomes slower in the fast transitioning data. By the time $a = 0.95$, the data set is indistinguishable from one containing only one type of trace. The difficulty of inferring the presence of fast and slow sub-populations should be a function of both a and trace length.

Each data set was analyzed by hFRET under a two state model and a four state model (initializations used are described in Sec. 7.7.1.1). The model with the higher evidence for each data set is shown in Fig. 7.5. This figure is analogous to a phase diagram, showing in what regions of $\{a, T\}$ space hFRET can find the fast and slow transitioning sub-populations and in what regions the evidence suggest only type of trace in the data. Encouragingly, evidence suggests mixtures of sub-populations for most data sets. As expected, only for large values of a and small values of T does evidence favor the two state model.

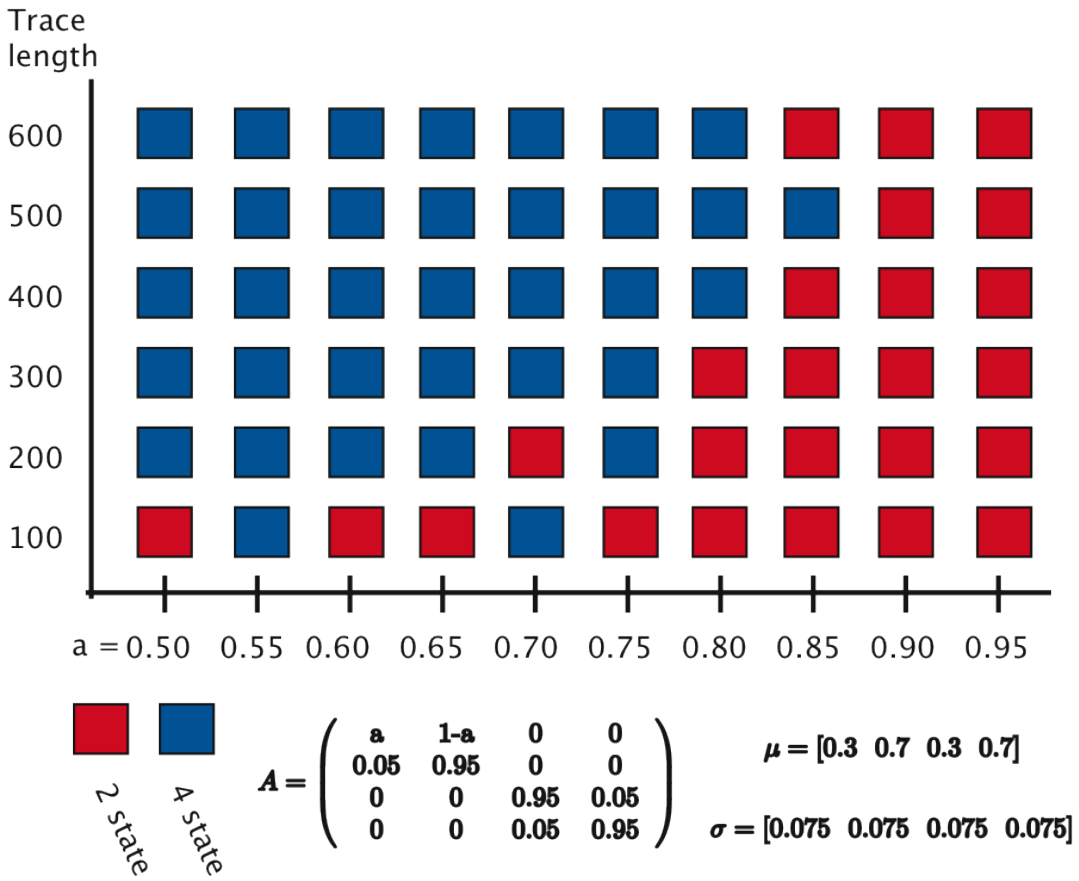

Figure 7.5: Ability of hFRET to detect the presence of fast and slow transitioning time series in a single data set. Each data set contains 100 traces with transition matrix A. In each data set 50 traces start in states 1 or 2 and 50 start in states 3 or 4. This is the transition matrix of a system with two sub-populations of traces in them (because transitions are only allowed between either states 1 and 2 or states 3 and 4, but not between the two groups). The smaller the value of a, the larger the difference between the traces containing states 1 and 2 and the traces containing states 3 and 4. By the time $a = 0.95$, the data are indistinguishable from a single two state system. Each data set was fit as both a two and four state system. The fit with the higher evidence is shown in this figure. Red denotes a two state model had the highest evidence. Blue denotes a four state model had the highest evidence. As seen from the results, the ability to infer the correct composition of the data depends on both the trace length and the value of a.

7.4.4 Finding sub-populations in real data

	R_-	R_+	R_{mix}
Transition Matrix (A)	$\begin{pmatrix} 0.953 & 0.047 \\ 0.028 & 0.972 \end{pmatrix}$	$\begin{pmatrix} 0.908 & 0.092 \\ 0.183 & 0.818 \end{pmatrix}$	$\begin{pmatrix} 0.955 & 0.044 & 0.001 & 0.000 \\ 0.024 & 0.975 & 0.001 & 0.000 \\ 0.000 & 0.000 & 0.905 & 0.095 \\ 0.000 & 0.002 & 0.260 & 0.738 \end{pmatrix}$
Means ($\vec{\mu}$)	[0.36 0.57]	[0.34 0.53]	[0.36 0.56 0.35 0.54]
Std. Dev. ($\vec{\sigma}$)	[0.048 0.051]	[0.056 0.059]	[0.048 0.050 0.055 0.059]
Number of traces	115	97	R_- 127, R_+ 85

Table 7.1: Table comparing the parameters inferred from smFRET data sets reporting on the ratcheting motion of the L1 stalk of the ribosome during translation (left) in the absence of EF-G (center) in the presence of saturating EF-G (right) in the presence of sub-saturating EF-G. The data set in the right column consists of all the traces analyzed in the left and center columns combined. The most probable means, standard deviations and transition matrices for the states in the data are shown. Just as in Fig. 7.5, both inference with 2-state and 4-state models were performed on the data. The results of the inference with the highest evidence are shown. In addition, the number of traces in the R_- and R_+ data sets are shown and the number of traces classified as originating from R_- and R_+ in the R_{mix} data set is shown. If inference were perfect, the R_{mix} transition matrix would be block diagonal and comprise the two 2x2 transition matrices of R_- and R_+ (with A_{R_-} corresponding to the upper left of A_{mix}, and A_{R_+} corresponding to the lower right), and the means and standard deviations of the states in R_{mix} would exactly match those in R_- and R_+. The agreement between $\vec{\theta}_{R_-}$, $\vec{\theta}_{R_+}$ and $\vec{\theta}_{R_{mix}}$ is good, although not exact, demonstrating the potential for hFRET to learn sub-populations within experimental data sets.

The true test of an inference program, however, is its performance on experimental data. In general, it is not possible to assess accuracy on experimental data, since the true states and rates of an experimental data set can never be known. In the case of detecting sub-populations, accuracy can be assessed by taking two separate data sets and comparing the inference results for the data analyzed separately

and analyzed as a single data set.

Previous work has shown that during protein translation the L1 stalk domain of the ribosome transitions between an open and closed conformation, correlating with tRNA movements between the classical and hybrid ribosome-bound configurations (Fei et al. 2008). The transition rate between open (k_{open}) and closed (k_{close}) conformations is a function of whether or not elongation factor G (EF-G) is bound to the ribosome. Ribosomes bound by EF-G (R_+) transition between k_{close} and k_{open} faster than ribosomes not bound by EF-G (R_-). The L1 stalk has been labeled with smFRET probes, and its ratcheting motion has been observed in the presence and absence of EF-G (Fei et al. 2009). The experimental setup is depicted in Fig. 7.7. The open and closed conformations give rise to distinct smFRET signals. Binding of EF-G to the ribosome only minimally effects these smFRET states, making this system an ideal candidate to test sub-population inference.

A data set containing translocating R_- and a data set containing translocating R_+ were analyzed by hFRET. The data were then pooled to form one large data set (R_{mix}) and inference was performed again. Evidence was used to choose the best model for each data set. The results are shown in Table 7.1. hFRET was able to detect the presence of the sub-populations in R_{mix}. States 1&2 of R_{mix} correspond to R_-. States 3&4 of R_{mix} correspond R_+. If inference were perfect, the upper left 2x2 block of A_{mix} (A_{UL}) would be identical to A_- and the lower right 2x2 (A_{LR}) block of A_{mix} would be identical to A_+. The agreement is extremely good, but not perfect, with $D_{KL}(A_-||A_{\text{UL}}) = 2.32 \times 10^{-4}$ and $D_{KL}(A_+||A_{\text{LR}}) = 5.69 \times 10^{-3}$. The largest transition rate difference between the individual and com-

bined inference was for the $f_{\text{high}} \rightarrow f_{\text{low}}$ transition of R_+ (0.183 in the individual and 0.260 for the combined inference). In addition, 12 traces were incorrectly assigned to the R_- data set, suggesting that some of the slower transitioning EF-G traces were incorrectly assigned as no EF-G data.

7.5 Discussion

hFRET is unique among smFRET inference programs in that it learns an ensemble probability distribution $p(\vec{\theta}_N)$ which governs all traces in a data set, whereas other inference methods only learn the parameters of individual traces. Once the ensemble distribution is learned, inference of individual traces is substantially more accurate as well, as demonstrated by the results in Fig. 7.3 and Fig. 7.4. The improved inference accuracy of hFRET is due to the algebraic equivalence of $\vec{\theta}_N$ to a prior. Inference of a trace's posterior using a learned "prior" which accurately describes the probability of the model parameters will be more accurate than inference using a preset prior which less accurately describes the probability of the model's parameters. Additionally, hFRET uses more information about the data set while performing inference on a trace than other existing methods do; when other inference methods analyze an individual trace, they do not "know" anything about the rest of the data set, but hFRET does.

An important goal of hFRET is to remove post-processing and subjective user input from the data analysis process. The trace to trace ambiguity of states in other inference methods (i.e. "state 1" might refer to the high FRET state in trace

1 and the low FRET state in trace 2) does not occur for hFRET because "state 1" will refer to the high FRET state for the entire data set in hFRET. The transition matrix learned by hFRET does not require idealized traces to be thresholded. As shown in Fig. 7.4, the transition rates learned from this transition matrix are more accurate than those learned in dwell-time analysis.

Moreover, when dwell-time analysis is performed on experimental data there there is often data which are hard to categorize. Imagine, for example, a data set with smFRET states centered at 0.4 and 0.6 and 0.9 FRET, all of which can vary by ± 0.1 FRET from trace to trace. A transition from a state idealized at 0.5 FRET to 0.9 FRET is observed. Did the 0.4 FRET state shift up in this trace or did the 0.6 FRET state shift down? The standard post-processing techniques cannot say, and the data would simply be thrown out. When hFRET performs inference on such a trace, it will account for the existence of the 0.4 and 0.6 FRET states and assign the observed 0.5 FRET data to the correct state based on the presence of other states in the data (e.g. if there is also a 0.4 FRET state present in the data, then the ambiguous data probably belongs to the 0.6 FRET state and vice versa).

There are several additional features of the ensemble learning algorithm that should be noted. First, while constructing the transition matrix, hFRET never actually "fits" the data. It calculates the probability that each observed data point belongs to each hidden state and uses those probabilities to construct the transition matrix. For most data points, the probability of belonging to one specific state is close to 1, in which case this distinction is largely irrelevant. For ambiguous data there is a difference though. For example, if a molecule transitions between two

smFRET states exactly halfway through a time binned step, the observed data point will be half way between the two smFRET states. In dwell time analysis, the data point must be idealized to one smFRET state or the other. hFRET would assign the data point to each state with a 50% probability, more accurately modeling the data.

Second, even if all states are not populated in every trace of a data set, hFRET can still be used for effective inference. When a state is absent from a specific trace, the VBEM algorithm used in each iteration of ensemble variational Bayes will leave the state unpopulated rather than use it to overfit another state. One important exception to this is that data containing both unpopulated states and photophysical shifts (i.e. the mean value of a smFRET state moves) within the same trace. In these cases, the unpopulated state will be used to fit part of the shifted smFRET state (because, statistically speaking, there are two distinct states with different mean smFRET values). These traces can substantially influence the $p(\vec{\theta}_N)$ learned and should not be used in hFRET inference.

Third, evidence based model selection is an excellent method to objectively choose between competing models in many situations. It is computationally efficient and uses the entire data set for inference. The evidence calculated by hFRET can be used both to choose the number of smFRET states in the data and to detect the presence of sub-populations in the data differing only in kinetic parameters. There are two caveats that should be considered when using evidence based model selection. The first is that if the data is poorly described by the model, then the evidence may monotonically increase with increasing model complexity. In the case

of time series inference, this most commonly occurs when a HMM is used to model a process where transitions between states occur on the same time scales as lifetimes in states. The HMM assumes the system jumps instantly from one state to another. This assumption is reasonable for most processes observed by smFRET. It may not be an appropriate assumption in all smFRET data however, and it is certainly not a reasonable assumption for many other single-molecule time series techniques. The second caveat is that evidence will favor the simpler of two competing models when either the simpler model is correct or the more complex model is correct but there is insufficient data to support it. If hFRET fails to detect the presence of sub-populations in the data, it may reflect an insufficient amount of data to support their existence rather than a lack of their existence. These two caveats are true for all evidence based model selection techniques, not just hFRET.

7.6 Conclusions

The results of these synthetic and experimental data analyses demonstrate the value of learning from the entire ensemble of traces in a data set over learning from just the individual traces. hFRET outperformed inference by ME and ML on synthetic data of increasing smFRET state noise and decreasing trace length. The ability to efficiently and straightforwardly learn a transition matrix directly from the data is a substantial advantage of hFRET. Not only does the transition matrix learned yield more accurate rate constants, but it can also be accurately learned from data with few (i.e. 0 or 1) transitions per trace. These sparsely transitioning traces are

problematic for dwell-time analysis.

Arguably the most exciting advantage of ensemble learning is the ability to detect sub-populations of data within a data set. As demonstrated on both synthetic data and experimental data taken from a smFRET study of the ribosome, even sub-populations of data which possess identical smFRET states, and only differ in the transition rates between these states can be identified. Rate constants for the sub populations can be accurately inferred.

Finally the approach to ensemble learning proposed here need not be limited to smFRET, or even HMM analysis. Any time series data suitable for inference by ME/VBEM, such as stepping data or tethered particle motion data, can also benefit from inference via ensemble variational Bayes. Even non-time series data, such as Gaussian mixture modeling across multiple data sets, can be analyzed via this approach.

7.7 Supplementary materials

7.7.1 Methods

7.7.1.1 hFRET algorithm

The ensemble learning algorithm used by hFRET is performed as follows:

1. Guess $p(\vec{\theta}_N)$.

2. Use $p(\vec{\theta}_N)$ as a prior and perform VBEM inference on individual traces, as previously described (Bronson et al. 2009), to learn $p(\vec{\theta}_n|\mathbf{D}_n)$.

3. Estimate $\vec{\theta}_{n*} = \max_{\vec{\theta}_n} p(\vec{\theta}_n|\mathbf{D}_n)$ for each trace.

4. Use the means and variances of each parameter in $\vec{\theta}_{n*}$ to set the sufficient statistics for a new $p(\vec{\theta}_N)$.

5. Repeat 2-4 until the evidence peaks or converges.

The solution converged to by the ensemble learning algorithm depends on the initial choice of $p(\vec{\theta}_N)$. To ensure the best possible choice for $p(\vec{\theta}_N)$, multiple initial guesses should be used. The $p(\vec{\theta}_N)$ guess for each model under consideration which results in the highest evidence score after one round of VBEM should be used for the ensemble learning algorithm.

In all inference performed in this work, up to 10 rounds of ensemble learning were allowed. For each data set in Fig. 7.3 and Fig. 7.4, inference with $\mathrm{K} = 1 - 5$ smFRET states was attempted. For each value of K, 10 initial guesses for $p(\vec{\theta}_N)$ were used. Hyperparameters for the initial guesses were set as follows:

$u_\pi^k = 1 \; \forall \; k$
$u_a^{jk} = 1 \; \forall \; j, k$
$u_\beta^k = 1 \; \forall \; k$
$u_W^k = 50 \; \forall \; k$
$u_v^k = 5 \; \forall \; k$

u_μ^k were evenly distributed between 0 and 1 for the first guess and randomly distributed for the remaining guesses.

Different initializations of $p(\vec{\theta}_N)$ are required to identify sub-populations of fast and slow transitioning data because fast and slow transitioning states cannot consistently be assigned to the same state in different traces if $u_a^{jk} = 1 \; \forall \; j, k$. A

different set of initial guesses was used for inference on data sets containing sub-populations of data, designed specifically to look for sub-populations of data with different rate constants. A total of 8 initial $p(\vec{\theta}_N)$ guesses were used:

$u_{\pi}^{k} = 1 \; \forall \; k$
$u_{\beta}^{k} = 1 \; \forall \; k$
$u_{W}^{k} = 50 \; \forall \; k$
$u_{v}^{k} = 5 \; \forall \; k$
$\vec{u}_{mu} = \{0.3, 0.7\}$ or $\{0.3, 0.7, 0.3, 0.7\}$

$$u_{A1} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \; u_{A2} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \; u_{A3} = \begin{pmatrix} 10 & 1 & 0 & 0 \\ 1 & 10 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

$$u_{A4} = \begin{pmatrix} 1 & 1 & 0.1 & 0 \\ 0.1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \; u_{A5} = \begin{pmatrix} 10 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \; u_{A6} = \begin{pmatrix} 1 & 0.1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

$$u_{A7} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 10 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \; u_{A8} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0.1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

The $\vec{\theta}_{n*}$ values were used to set the hyperparameters as follows:

For a Dirichlet distribution of the form

$$p(\vec{\pi}) = \frac{\Gamma(\widehat{u_\pi})}{\prod_{k=1}^{K}\Gamma(u_\pi^k)}\prod_{k=1}^{K}\pi_k^{u_\pi^k-1}, \tag{7.9}$$

where $\widehat{u_\pi} = \sum_{k=1}^{K} u_\pi^k$, the variance of each u_π^k is

$$\text{var}[\pi_k] = \frac{u_\pi^k(\widehat{u_\pi} - u_\pi^k)}{\widehat{u_\pi}^2(\widehat{u_\pi}+1)} \tag{7.10}$$

and the mean of each u_π^k is

$$\mathbb{E}[\pi_k] = \frac{u_\pi^k}{\widehat{u_\pi}} \tag{7.11}$$

Once $\text{var}[\pi_k]$ and $\mathbb{E}[\pi_k]$ are estimated from the data then Eq. 7.10 and Eq. 7.11 can be rearranged to yield

$$\widehat{u_\pi} = \left[\frac{\mathbb{E}[\pi_k](1-\mathbb{E}[\pi_k])}{\text{var}[\pi_k]}\right] - 1 \tag{7.12}$$

$$u_\pi^k = \widehat{u_\pi}\ \mathbb{E}[\pi_k]. \tag{7.13}$$

For a Gamma function with parameters a and b describing λ_k,

$$\text{Gam}(\lambda_k|a,b) = \frac{1}{\Gamma(a)}b^a\lambda_k^{a-1}e^{-b\lambda_k}, \tag{7.14}$$

the variance of λ_k is

$$\text{var}[\lambda_k] = \frac{a}{b} \tag{7.15}$$

$$\mathbb{E}[\lambda_k] = \frac{a}{b^2}. \tag{7.16}$$

Rearranging to solve for $\text{var}[\lambda_k]$ and $\mathbb{E}[\lambda_k]$ yields

$$b = \frac{\mathbb{E}[\lambda_k]}{\text{var}[\lambda_k]} \tag{7.17}$$

$$a = b\ \mathbb{E}[\lambda_k]. \tag{7.18}$$

For a Gaussian with parameters $\{u_\mu^k, (u_\beta^k \lambda_k)^{-1}\}$ describing

$$\mu_k = \sqrt{\frac{u_\beta^k \lambda_k}{2\pi}} \exp\left(-\frac{u_\beta^k \lambda_k}{2}(\mu_k - u_\mu^k)^2\right) \tag{7.19}$$

u_β^k should be set such that $u_\beta^k \lambda_k$ is var$[\mu_k]$ and u_μ^k should be chosen to be $\mathbb{E}[\mu_k]$.

7.7.2 ME & ML inference

Inference via ME and ML was performed as previously described (Bronson et al. 2009).

7.7.3 Data

Synthetic data was generated as previously described (Bronson et al. 2009). Ribosome data was taken from (Fei et al. 2009). Only experimental traces longer than 100 time steps were analyzed.

7.7.4 Additional Figures

.

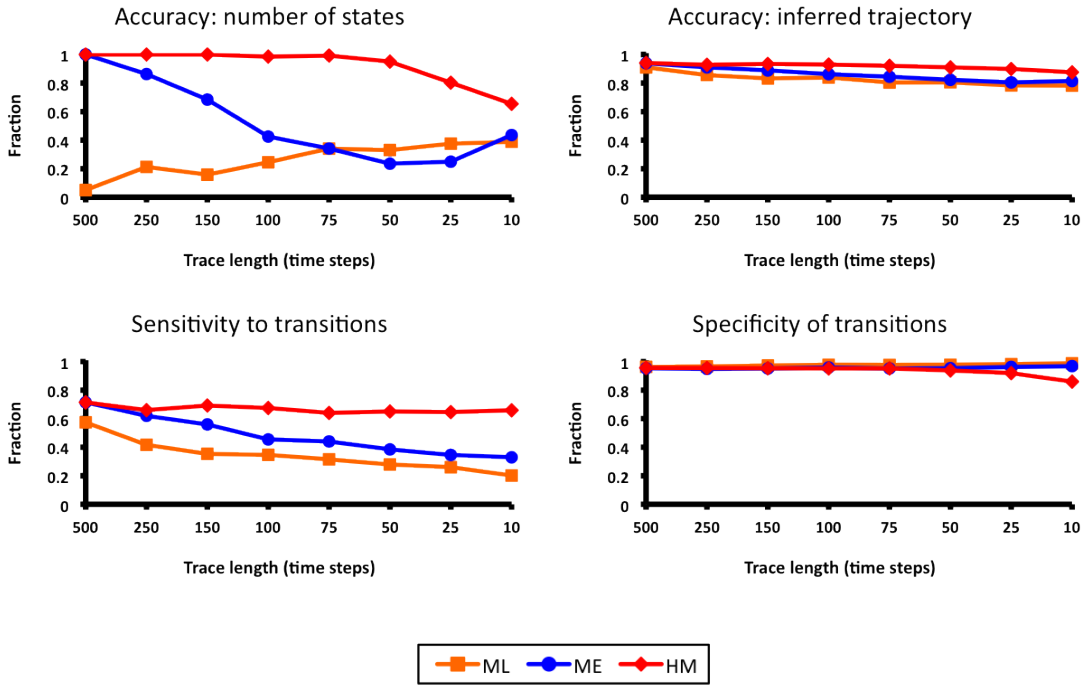

Figure 7.6: Accuracy, as measured by the four probabilities described in Sec. 7.4.1, for the data analyzed in Sec. 7.4.2. Inference results for hFRET, ME and ML are shown in red, blue and orange, respectively.

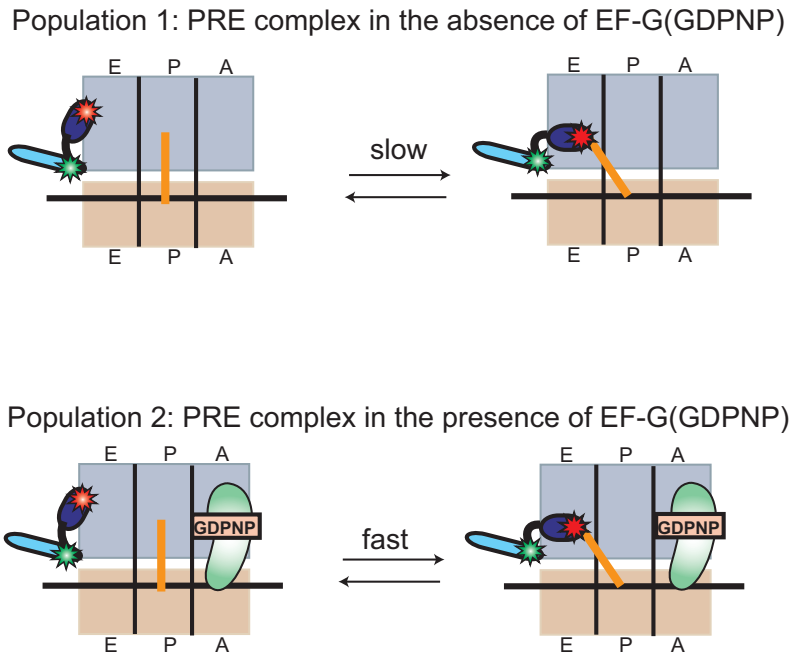

Figure 7.7: Cartoon depicting the experimental setup for the data analyzed in Sec. 7.4.4. The small and large ribosomal subunits are shown in tan and lavender, respectively, with the L1 stalk depicted in dark blue, and ribosomal protein L9 in cyan. The aminoacyl-, peptidyl- and deacylated-tRNA binding sites are labeled as A, P and E, respectively, and the P-site tRNA is depicted as a brown line. The smFRET probes Cy3 and Cy5 are depicted as green and red stars, respectively. In this complex, the ribosome fluctuates between the classical (left) and hybrid (right) configurations causing a smFRET detectable ratcheting motion of the L1 stalk. Transition rates between the classical and hybrid configurations change in the presence and absence of EF-G.

Chapter 8

Future Work

The work presented here describes several advances in the field of smFRET inference. The introduction of the ME criteria for model selection, described in Ch. 5 and Ch. 6, is the first principled approach to model selection employed in a smFRET inference software package. The use of ME appears to provide the additional benefit of improved accuracy over previously used ML methods, especially for fast transitioning data.

It is not obvious why ME inference should be more accurate than ML inference when the number of states in the data is known. My personal suspicion is that pathological solutions with infinite likelihood (Sec. 4.4.1) are interfering with ML inference. On a theoretical level, the presence of divergent solutions negates the entire principle of ML since it is possible to achieve infinite likelihood with a model that poorly describes the data. On a practical level, convergence to divergent solutions must be detected by the ML software and corrected by resetting model parameters to assign more data to collapsing states. One would expect that if the

transition matrix allows states to rapidly transition, it would be more probable that a single data point would be assigned to its own state and result in a pathological solution. Consequently, faster transitioning data should be more problematic for ML. For the same reason, initializing the transition matrix to have fast transitioning states could be problematic. This may be why the popular ML software HaMMy requires each state to have a 90% chance of not transitioning in the initial guess for the transition matrix (*i.e.* HaMMy is initialized so $a_{k,k} = 0.9 \; \forall \; \mathrm{K}$).

The use of hierarchical modeling, described in Ch. 7, further improves inference accuracy and allows the detection of photophysically identical, but kinetically distinct, sub-populations within a data set. My hope is that this approach will ultimately replace both ME and ML for smFRET inference. Unfortunately there are many data sets which cannot currently be analyzed this way. When data sets have traces which have both (1) unpopulated states and (2) states where the mean smFRET intensity shifts mid time series within the same trace, the algorithm uses the state which should be unpopulated to overfit the shifted state. An example is shown in Fig. 8.1. If enough of these traces exist in the data, they will skew the sufficient statistics learned during ensemble variational Bayes, $p(\vec{\theta}_N)$ will not describe the data well and the resulting inference can be nonsensical — *e.g.*, the idealized traces learned look, by visual inspection, wrong or the order of the states will switch from trace to trace (*i.e.* state 1 might be the high FRET state in trace 1 and the low FRET state in trace 2). It should be emphasized that simply having traces with unpopulated states is *not* problematic. The issue here is that a smFRET state with two mean intensity values deviates from model of the data used

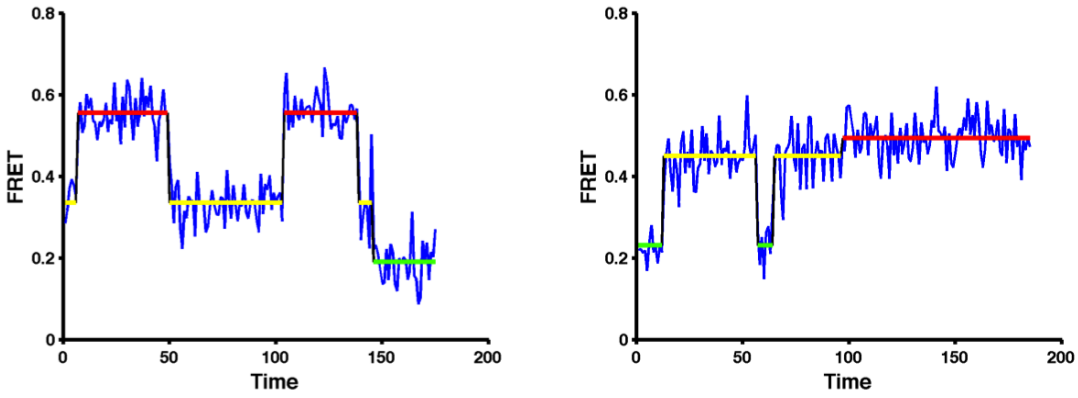

Figure 8.1: Traces demonstrating the current limitations of hFRET. The majority of the traces in this experimental data set, taken from (Fei et al. 2009), contain three smFRET states (a representative trace is shown on the left). The data are shown in blue. The idealized f_{low}, f_{mid} and f_{high} states are shown as green, yellow and red, respectively, and connected by black lines. A small number of traces only contain two of the three smFRET states and have the mean intensity of a smFRET state shift in the middle of the trace (a sample is shown on the right). hFRET uses the unpopulated state in these traces to overfit the shifted state. This skews the sufficient statistics used to set $p(\vec{\theta}_N)$, resulting in poor inference for all traces and unrealistic rate constants learned from the data.

by hFRET.

There are several possible fixes for this issue. A criteria could be devised to detect these traces and ignore them during inference. For example, any trace having two states with smFRET means less than 0.1 FRET could be ignored during ensemble Variational Bayes. Should this fix prove effective it would still be undesirable, however, since it requires an unprincipled correction the algorithm. An important goal in the work presented in this thesis is to avoid such fixes and rely on principled graphical modeling for inference. A more promising venue would be to adjust the model itself to better describe the data. For example, the emissions model could be changed to describe each state with a beta distribution. Although

a beta distribution still would not account for a smFRET state with multiple mean smFRET intensity values, it might generally describe smFRET data better, since both the beta distribution and FRET transformed data can only have values between 0 and 1. The distribution might be less sensitive to shifts in mean smFRET state intensities. Unfortunately, $\text{Beta}(\theta|a,b)$ is an exponential distribution with respect to θ, but not with respect to a and b. There is no known prior over a and b which allows for inference via VBEM. The emissions model could also assign each smFRET state a mixture of Gaussians. This would better model our belief that each smFRET state is actually an ensemble of locally stable conformations, each of which has a closely centered but possibly distinct mean smFRET intensity. Choosing the number of Gaussians for each smFRET state and ensuring that each mixture only fits data within a single state could prove challenging.

Finally, although briefly discussed in Sec. 6.1, the question of whether analyzing the 2D raw smFRET donor/acceptor data or the 1D smFRET transformation is still largely unanswered. The question is hard to answer using data since synthetic data can easily be constructed to favor 1D or 2D inference and the true hidden states of real data can never be known. The issue is whether the donor/acceptor intensity signals provide real information about the molecule, or only the smFRET transfer efficiency is indicative of intermolecular distances. If the former is true, 2D inference should be more accurate, since information is lost in the 1D transform. If the latter is true then 2D inference would simply add artifacts to the data every time the donor/acceptor change absolute intensity without altering the FRET ratio. The true answer might require a very detailed picture of the physics of the

FRET phenomenon, the biophysics of the target molecule's dynamics and the experimental set up. A possible experiment to address this issue would be to record the same data at multiple camera shutter speeds. The longer the exposure time, the better resolution the data. For example the same experiment could be recorded with a {200 ms, 100 ms, 50 ms, 25 ms, 10 ms} exposure time. Most data sets are well resolved at 200 ms and unresolvable at 10 ms. Since the observed system is the same regardless of the camera shutter speed, the 200 ms inference could be used to learn the "correct" rate constants and the relative performance of 1D and 2D inference could be compared on the faster shutter speed data sets. This experiment would require slow enough transitioning data to observe transitions at all shutter speeds. It would only answer questions about camera noise, which may or may not be the most significant source of noise in the data.

In conclusion, while much exciting progress has been made, there are still many interesting questions left to explore in the field of smFRET inference. I hope this work, and the open source software I have written, will allow the smFRET community to employ more accurate and statistically rigorous data analysis and that the modeling approach used here can inspire biophysicists analyzing other forms of time series data as well.

Part III

Biblography & Appendix

Bibliography

Agirrezabala, X., Lei, J., Brunelle, J. L., Ortiz-Meoz, R. F., Green, R., and Frank, J. 2008. Visualization of the hybrid state of trna binding promoted by spontaneous ratcheting of the ribosome. *Mol Cell*, 32(2):190–7.

Aharoni, A., Griffiths, A. D., and Tawfik, D. S. 2005. High-throughput screens and selections of enzyme-encoding genes. *Curr Opin Chem Biol*, 9:210–216.

Amann, E., Brosius, J., and Ptashne, M. 1983. Vectors bearing a hybrid trp-lac promoter useful for regulated expression of cloned genes in Escherichia coli. *Gene*, 25:167–178.

Anderson, J. C., Clarke, E. J., Arkin, A. P., and Voigt, C. A. 2006. Environmentally controlled invasion of cancer cells by engineered bacteria. *J. Mol. Biol.*, 355:619–627.

Andrec, M., Levy, R. M., and Talaga, D. S. 2003. Direct Determination of Kinetic Rates from Single-Molecule Photon Arrival Trajectories Using Hidden Markov Models. *J Phys Chem A*, 107:7454–7464.

Baccanari, D. P., Daluge, S., and King, R. W. 1982. Inhibition of dihydrofolate reductase: effect of reduced nicotinamide adenine dinucleotide phosphate on the selectivity and affinity of diaminobenzylpyrimidines. *Biochemistry*, 21:5068–5075.

Baker, D., Group, B., Church, G., Collins, J., Endy, D., Jacobson, J., Keasling, J., Modrich, P., Smolke, C., and Weiss, R. 2006. Engineering life: Building a fab for biology. *Scientific American*, 294(6):44–51.

Baker, K., Bleczinski, C., Lin, H., Salazar-Jimenez, G., Sengupta, D., Krane, S., and Cornish, V. W. 2002. Chemical complementation: a reaction-independent genetic assay for enzyme catalysis. *Proc. Natl. Acad. Sci. U.S.A.*, 99:16537–16542.

Baker, K., Sengupta, D., Salazar-Jimenez, G., and Cornish, V. W. 2003. An optimized dexamethasone-methotrexate yeast 3-hybrid system for high-throughput screening of small molecule-protein interactions. *Anal. Biochem.*, 315:134–137.

Barrick, D., Villanueba, K., Childs, J., Kalil, R., Schneider, T. D., Lawrence, C. E., Gold, L., and Stormo, G. D. 1994. Quantitative analysis of ribosome binding sites in E.coli. *Nucleic Acids Res.*, 22:1287–1295.

Beekwilder, J., Wolswinkel, R., Jonker, H., Hall, R., de Vos, C. H., and Bovy, A. 2006. Production of resveratrol in recombinant microorganisms. *Appl. Environ. Microbiol.*, 72:5670–5672.

Benkovic, S. J., Fierke, C. A., and Naylor, A. M. 1988. Insights into enzyme function from studies on mutants of dihydrofolate reductase. *Science*, 239:1105–1110.

Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K., Huebert, D. J., McMahon, S., Karlsson, E. K., Kulbokas, E. J., Gingeras, T. R., Schreiber, S. L., and Lander, E. S. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120:169–181.

Binz, H. K., Amstutz, P., and Pluckthun, A. 2005. Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.*, 23:1257–1268.

Bishop, C. 2006. *Pattern Recognition and Machine Learning*. Oxford University Press, Oxford Oxfordshire.

Blanchard, S. C., Gonzalez, R. L., Kim, H. D., Chu, S., and Puglisi, J. D. 2004a. tRNA selection and kinetic proofreading in translation. *Nat. Struct. Mol. Biol.*, 11:1008–1014.

Blanchard, S. C., Kim, H. D., Gonzalez, R. L., J., Puglisi, J. D., and Chu, S. 2004b. trna dynamics on the ribosome during translation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(35):12893–8.

Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. 2006. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U.S.A.*, 103:5869–5874.

Bloom, J. D., Meyer, M. M., Meinhold, P., Otey, C. R., MacMillan, D., and Arnold, F. H. 2005. Evolving strategies for enzyme engineering. *Curr. Opin. Struct. Biol.*, 15:447–452.

Brakmann, S. and Johnsson, K. 2002. *Directed Molecular Evolution of Proteins: or How to Improve Enzymes for Biocatalysis*. Wiley-VCH.

Bronson, J. E., Fei, J., Hofman, J. M., Gonzalez, R. L., and Wiggins, C. H. 2009. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.*, 97:3196–3205.

Calos, M. P. 1978. DNA sequence for a low-level promoter of the lac repressor gene and an 'up' promoter mutation. *Nature*, 274:762–765.

Camps, M., Naukkarinen, J., Johnson, B. P., and Loeb, L. A. 2003. Targeted gene evolution in Escherichia coli using a highly error-prone DNA polymerase I. *Proc. Natl. Acad. Sci. U.S.A.*, 100:9727–9732.

Chen, M. T. and Weiss, R. 2005. Artificial cell-cell communication in yeast Saccharomyces cerevisiae using signaling elements from Arabidopsis thaliana. *Nat. Biotechnol.*, 23:1551–1555.

Cormack, B. P., Valdivia, R. H., and Falkow, S. 1996. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene*, 173:33–38.

Cornish, P. V., Ermolenko, D. N., Noller, H. F., and Ha, T. 2008. Spontaneous intersubunit rotation in single ribosomes. *Molecular Cell*, 30(5):578–588.

Cornish, P. V., Ermolenko, D. N., Staple, D. W., Hoang, L., Hickerson, R. P., Noller, H. F., and Ha, T. 2009. Following movement of the l1 stalk between three functional states in single ribosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(8):2571–6.

Dahan, M., Deniz, A. A., Ha, T. J., Chemla, D. S., Schultz, P. G., and Weiss, S. 1999. Ratiometric measurement and identification of single diffusing molecules. *Chemical Physics*, 247(1):85–106.

del Solar, G., Giraldo, R., Ruiz-Echevarria, M. J., Espinosa, M., and Diaz-Orejas, R. 1998. Replication and control of circular bacterial plasmids. *Microbiol. Mol. Biol. Rev.*, 62:434–464.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society Series B-MEthodological*, 39(1):1–38.

Deniz, A. A., Laurence, T. A., Beligere, G. S., Dahan, M., Martin, A. B., Chemla, D. S., Dawson, P. E., Schultz, P. G., and Weiss, S. 2000. Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. U.S.A.*, 97:5179–5184.

Dias, N. and Stein, C. A. 2002. Antisense oligonucleotides: basic concepts and mechanisms. *Mol. Cancer Ther.*, 1:347–355.

Doublie, S., Tabor, S., Long, A. M., Richardson, C. C., and Ellenberger, T. 1998. Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 A resolution. *Nature*, 391:251–258.

Dower, W. J., Miller, J. F., and Ragsdale, C. W. 1988. High efficiency transformation of E. coli by high voltage electroporation. *Nucleic Acids Res.*, 16:6127–6145.

Drummond, D. A., Iverson, B. L., Georgiou, G., and Arnold, F. H. 2005. Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. *J. Mol. Biol.*, 350:806–816.

Elowitz, M. B. and Leibler, S. 2000. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335–338.

Fabret, C., Poncet, S., Danielsen, S., Borchert, T. V., Ehrlich, S. D., and Janniere, L. 2000. Efficient gene targeted random mutagenesis in genetically stable Escherichia coli strains. *Nucleic Acids Res.*, 28:E95.

Fei, J., Bronson, J. E., Hofman, J. M., Srinivas, R. L., Wiggins, C. H., and Gonzalez, R. L. 2009. Allosteric collaboration between elongation factor G and the ribosomal L1 stalk directs tRNA movements during translation. *Proc. Natl. Acad. Sci. U.S.A.*, 106:15702–15707.

Fei, J., Kosuri, P., MacDougall, D. D., and Gonzalez, R. L. 2008. Coupling of ribosomal l1 stalk and trna dynamics during translation elongation. *Molecular Cell*, 30(3):348–359.

Fields, S. and Song, O. 1989. A novel genetic system to detect protein-protein interactions. *Nature*, 340:245–246.

Fischer, H. and Hinkle, D. C. 1980. Bacteriophage T7 DNA replication in vitro. Stimulation of DNA synthesis by T7 RNA polymerase. *J. Biol. Chem.*, 255:7956–7964.

Förster, T. 1948. Zwischenmolekulare Energiewanderung Und Fluoreszenz. *Annalen Der Physik*, 2(1-2):55–75.

Foster, P. L. 2005. Stress responses and genetic variation in bacteria. *Mutat. Res.*, 569:3–11.

Fuller, C. W. and Richardson, C. C. 1985. Initiation of DNA replication at the primary origin of bacteriophage T7 by purified proteins. Site and direction of initial DNA synthesis. *J. Biol. Chem.*, 260:3185–3196.

Gallagher, S. S., Miller, L. W., and Cornish, V. W. 2007. An orthogonal dexamethasone-trimethoprim yeast three-hybrid system. *Anal. Biochem.*, 363:160–162.

Gardner, T. S., Cantor, C. R., and Collins, J. J. 2000. Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403:339–342.

Gauvain, J.-L. and Lee, C.-H. 1994. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on speech and audio processing*, 2(2):291–298.

Gelman, A. and Hill, J. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Giver, L., Gershenson, A., Freskgard, P. O., and Arnold, F. H. 1998. Directed evolution of a thermostable esterase. *Proc. Natl. Acad. Sci. U.S.A.*, 95:12809–12813.

Goldberg, S. D., Iannuccilli, W., Nguyen, T., Ju, J., and Cornish, V. W. 2003. Identification of residues critical for catalysis in a class C beta-lactamase by combinatorial scanning mutagenesis. *Protein Sci.*, 12:1633–1645.

Guet, C. C., Elowitz, M. B., Hsing, W., and Leibler, S. 2002. Combinatorial synthesis of genetic networks. *Science*, 296:1466–1470.

Ha, T., Enderle, T., Ogletree, D. F., Chemla, D. S., Selvin, P. R., and Weiss, S. 1996. 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.*, 93:6264–6268.

Hasty, J., McMillen, D., and Collins, J. J. 2002. Engineered gene circuits. *Nature*, 420:224–230.

Hohng, S., Joo, C., and Ha, T. 2004. Single-molecule three-color FRET. *Biophys. J.*, 87:1328–1337.

Jares-Erijman, E. A. and Jovin, T. M. 2003. Fret imaging. *Nature Biotechnology*, 21(11):1387–1395.

Ji, S., Krishnapuram, B., and Carin, L. 2006. Variational Bayes for continuous hidden Markov models and its application to active learning. *IEEE Trans Pattern Anal Mach Intell*, 28:522–532.

Joo, C., Balci, H., Ishitsuka, Y., Buranachai, C., and Ha, T. 2008. Advances in single-molecule fluorescence methods for molecular biology. *Annu. Rev. Biochem.*, 77:51–76.

Julian, P., Konevega, A. L., Scheres, S. H., Lazaro, M., Gil, D., Wintermeyer, W., Rodnina, M. V., and Valle, M. 2008. Structure of ratcheted ribosomes with trnas in hybrid states. *Proceedings of the National Academy of Sciences of the United States of America*, 105(44):16924–7.

Kaern, M., Blake, W. J., and Collins, J. J. 2003. The engineering of gene regulatory networks. *Annu Rev Biomed Eng*, 5:179–206.

Kass, R. and Raftery, A. 1995. Bayes factors. *Journal of the American Statistical Association*, 90(430).

Kemp, C. and Tenenbaum, J. B. 2008. The discovery of structural form. *Proc. Natl. Acad. Sci. U.S.A.*, 105:10687–10692.

Kim, H. D., Puglisi, J. D., and Chu, S. 2007. Fluctuations of transfer rnas between classical and hybrid states. *Biophys J*, 93(10):3575–82.

Kirby, J. and Keasling, J. D. 2009. Biosynthesis of plant isoprenoids: perspectives for microbial engineering. *Annu Rev Plant Biol*, 60:335–355.

Koster, D., Wiggins, C., and Dekker, N. 2006. Multiple events on single molecules: Unbiased estimation in single-molecule biophysics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(6):1750–1755.

Kruger, D. H. and Schroeder, C. 1981. Bacteriophage T3 and bacteriophage T7 virus-host cell interactions. *Microbiol. Rev.*, 45:9–51.

Kuwayama, H., Obara, S., Morio, T., Katoh, M., Urushihara, H., and Tanaka, Y. 2002. PCR-mediated generation of a gene disruption construct without the use of DNA ligase and plasmid vectors. *Nucleic Acids Res.*, 30:E2.

Leung, D. W., Chen, E., and Goeddel, D. V. 1989. A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique*, 1:11–15.

Li, S. C., Squires, C. L., and Squires, C. 1984. Antitermination of E. coli rRNA transcription is caused by a control region segment containing lambda nut-like sequences. *Cell*, 38:851–860.

Licitra, E. J. and Liu, J. O. 1996. A three-hybrid system for detecting small ligand-protein receptor interactions. *Proc. Natl. Acad. Sci. U.S.A.*, 93:12817–12821.

Lin, H., Abida, W., Sauer, R., and Cornish, V. 2000. Dexamethasone-methotrexate: An efficient chemical inducer of protein dimerization in vivo. *Journal of the American Chemical Society*, 122(17):4247–4248.

MacKay, D. J. 2003. *Information theory, inference, and learning algorithms*. Cambridge University Press.

McCulloch, R. and Rossi, P. E. 1991. A bayesian approach to testing the arbitrage pricing theory. *Journal of Econometrics*, 49(1-2):141 – 168.

McKinney, S. A., Joo, C., and Ha, T. 2006. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.*, 91:1941–1951.

Moazed, D. and Noller, H. F. 1989. Intermediate states in the movement of transfer rna in the ribosome. *Nature*, 342(6246):142–8.

Moffitt, J. R., Chemla, Y. R., Aathavan, K., Grimes, S., Jardine, P. J., Anderson, D. L., and Bustamante, C. 2009. Intersubunit coordination in a homomeric ring atpase. *Nature*, 457(7228):446–U2.

Mori, T., Vale, R. D., and Tomishige, M. 2007. How kinesin waits between steps. *Nature*, 450:750–754.

Munro, J. B., Altman, R. B., O'Connor, N., and Blanchard, S. C. 2007. Identification of two distinct hybrid state intermediates on the ribosome. *Molecular Cell*, 25(4):505–517.

Myong, S., Rasnik, I., Joo, C., Lohman, T. M., and Ha, T. 2005. Repetitive shuttling of a motor protein on dna. *Nature*, 437(7063):1321–5.

Neal, R. 1993. Probabilistic inference using Markov chain Monte Carlo methods. *Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto*.

Neylon, C. 2004. Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic Acids Res.*, 32:1448–1459.

Nguyen, A. W. and Daugherty, P. S. 2003. Production of randomly mutated plasmid libraries using mutator strains. *Methods Mol. Biol.*, 231:39–44.

Ollis, D. L., Kline, C., and Steitz, T. A. 1985. Domain of E. coli DNA polymerase I showing sequence homology to T7 DNA polymerase. *Nature*, 313:818–819.

Park, H. S., Nam, S. H., Lee, J. K., Yoon, C. N., Mannervik, B., Benkovic, S. J., and Kim, H. S. 2006. Design and evolution of new catalytic activity with an existing protein scaffold. *Science*, 311:535–538.

Patel, S. S., Rosenberg, A. H., Studier, F. W., and Johnson, K. A. 1992. Large scale purification and biochemical characterization of T7 primase/helicase proteins. Evidence for homodimer and heterodimer formation. *J. Biol. Chem.*, 267:15013–15021.

Perumal, S. K., Yue, H., Hu, Z., Spiering, M. M., and Benkovic, S. J. 2009. Single-molecule studies of DNA replisome function. *Biochim. Biophys. Acta*.

Pfleger, B. F., Pitera, D. J., Smolke, C. D., and Keasling, J. D. 2006. Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol.*, 24:1027–1032.

Qin, F., Auerbach, A., and Sachs, F. 1997. Maximum likelihood estimation of aggregated markov processes. *Proceedings of the Royal Society of London Series B-Biological Sciences*, 264(1380):375–383.

Qin, F., Auerbach, A., and Sachs, F. 2000. A direct optimization approach to hidden markov modeling for single channel kinetics. *Biophysical Journal*, 79(4):1915–1927.

Rabiner, L. R. 1989. A tutorial on hidden markov-models and selected applications in speech recognition. *Proceedings of the Ieee*, 77(2):257–286.

Rabkin, S. D. and Richardson, C. C. 1988. Initiation of DNA replication at cloned origins of bacteriophage T7. *J. Mol. Biol.*, 204:903–916.

Rice, G. C., Goeddel, D. V., Cachianes, G., Woronicz, J., Chen, E. Y., Williams, S. R., and Leung, D. W. 1992. Random PCR mutagenesis screening of secreted proteins by direct expression in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.*, 89:5467–5471.

Ro, D. K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., Ho, K. A., Eachus, R. A., Ham, T. S., Kirby, J., Chang, M. C., Withers, S. T., Shiba, Y., Sarpong, R., and Keasling, J. D. 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440:940–943.

Romero, P. A. and Arnold, F. H. 2009. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, 10:866–876.

Rose, R. E. 1988. The nucleotide sequence of pACYC177. *Nucleic Acids Res.*, 16:356.

Rosenberg, A. H., Patel, S. S., Johnson, K. A., and Studier, F. W. 1992. Cloning and expression of gene 4 of bacteriophage T7 and creation and analysis of T7 mutants lacking the 4A primase/helicase or the 4B helicase. *J. Biol. Chem.*, 267:15005–15012.

Ross, S. 2008. *First Course in Probability, a*. Oxford University Press, Oxford Oxfordshire.

Roy, R., Hohng, S., and Ha, T. 2008. A practical guide to single-molecule fret. *Nature Methods*, 5(6):507–516.

Roy, R., Kozlov, A. G., Lohman, T. M., and Ha, T. 2009. SSB protein diffusion on single-stranded DNA stimulates RecA filament formation. *Nature*, 461:1092–1097.

Sakon, J. J. and Weninger, K. R. 2010. Detecting the conformation of individual proteins in live cells. *Nat. Methods*, 7:203–205.

Schuler, B. and Eaton, W. A. 2008. Protein folding studied by single-molecule fret. *Current Opinion in Structural Biology*, 18(1):16–26.

Schuler, B., Lipman, E. A., Steinbach, P. J., Kumke, M., and Eaton, W. A. 2005. Polyproline and the "spectroscopic ruler" revisited with single-molecule fluorescence. *Proc. Natl. Acad. Sci. U.S.A.*, 102:2754–2759.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Setty, Y., Mayo, A. E., Surette, M. G., and Alon, U. 2003. Detailed map of a cis-regulatory input function. *Proc. Natl. Acad. Sci. U.S.A.*, 100:7702–7707.

Shannon, C. E. 1948a. A mathematical theory of communication (Part 1). *Bell System Technical Journal*, 27(1):379–423.

Shannon, C. E. 1948b. A mathematical theory of communication (Part 2). *Bell System Technical Journal*, 27(2):632–656.

Shevchuk, N. A., Bryksin, A. V., Nusinovich, Y. A., Cabello, F. C., Sutherland, M., and Ladisch, S. 2004. Construction of long DNA molecules using long PCR-based fusion of several fragments simultaneously. *Nucleic Acids Res.*, 32:e19.

Spencer, D. M., Wandless, T. J., Schreiber, S. L., and Crabtree, G. R. 1993. Controlling signal transduction with synthetic ligands. *Science*, 262:1019–1024.

Stemmer, W. P. 1994. Rapid evolution of a protein in vitro by DNA shuffling. *Nature*, 370:389–391.

Sternberg, S. H., Fei, J., Prywes, N., McGrath, K. A., and Gonzalez, R. L. 2009. Translation factors direct intrinsic ribosome dynamics during translation termination and ribosome recycling. *Nat. Struct. Mol. Biol.*, 16:861–868.

Strathern, D. C. A. D. J. B. J. N. 2005. *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual, 2005 Edition*. Cold Spring Harbor Laboratory Press.

Stryer, L. and Haugland, R. P. 1967. Energy transfer: a spectroscopic ruler. *Proc. Natl. Acad. Sci. U.S.A.*, 58:719–726.

Studier, F. W. and Moffatt, B. A. 1986. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.*, 189:113–130.

Tabor, S. and Richardson, C. C. 1990. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. Effect of pyrophosphorolysis and metal ions. *J. Biol. Chem.*, 265:8322–8328.

Tan, E., Wilson, T. J., Nahas, M. K., Clegg, R. M., Lilley, D. M., and Ha, T. 2003. A four-way junction accelerates hairpin ribozyme folding via a discrete intermediate. *Proc. Natl. Acad. Sci. U.S.A.*, 100:9308–9313.

Van Dongen, S. 2006. Prior specification in Bayesian statistics: three cautionary tales. *J. Theor. Biol.*, 242:90–100.

Viterbi, A. J. 1967. Error Bounds For Convolutional Codes And An Asymptotically Optimum Decoding Algorithm. *IEEE Transactions On Information Theory*, 13(2):260+.

Walters, W. P. and Namchuk, M. 2003. Designing screens: how to make your hits a hit. *Nat Rev Drug Discov*, 2:259–266.

Wiita, A. P., Perez-Jimenez, R., Walther, K. A., Grater, F., Berne, B. J., Holmgren, A., Sanchez-Ruiz, J. M., and Fernandez, J. M. 2007. Probing the chemistry of thioredoxin catalysis with force. *Nature*, 450(7166):124–7.

Wilson, D. S., Szostak, J. W., and Szostak, J. W. 1999. In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.*, 68:611–647.

Yildiz, A., Tomishige, M., Vale, R. D., and Selvin, P. R. 2004. Kinesin walks hand-over-hand. *Science*, 303(5658):676–8.

You, L., Cox, R. S., Weiss, R., and Arnold, F. H. 2004. Programmed population control by cell-cell communication and regulated killing. *Nature*, 428:868–871.

Zhuang, X., Kim, H., Pereira, M. J., Babcock, H. P., Walter, N. G., and Chu, S. 2002. Correlating structural dynamics and function in single ribozyme molecules. *Science*, 296:1473–1476.

Zhuang, X. W., Bartley, L. E., Babcock, H. P., Russell, R., Ha, T. J., Herschlag, D., and Chu, S. 2000. A single-molecule study of rna catalysis and folding. *Science*, 288(5473):2048–+.

Appendix A

T7 Primers and sequences

A.1 Replisome genes

Gene	Number	Size (kb)
T7 RNA polymerase	gp1	2.75
ssBP	gp2.5	0.70
Helicase/primase	gp4	1.70
T7 DNA polymerase	gp5	2.12

Table A.1: Genes of the T7 DNA replisome. The full T7 genome is listed in NCBI under GI:431187.

A.2 T7 origin of replication

The T7 origin of replication + 50 bp was PCRed using the following primers. Bold is used to denote the SfiI restriction site. Lower case letters are complementary to the T7 DNA.

Forward: GCATACGTCACATGT **GGCCCCCGGGGCC** aaggtaacttgaacctccg

Reverse: GCATACGTCGGTACC **GGCCGGCAGGGCC** ttagaagtcacgagcattacg

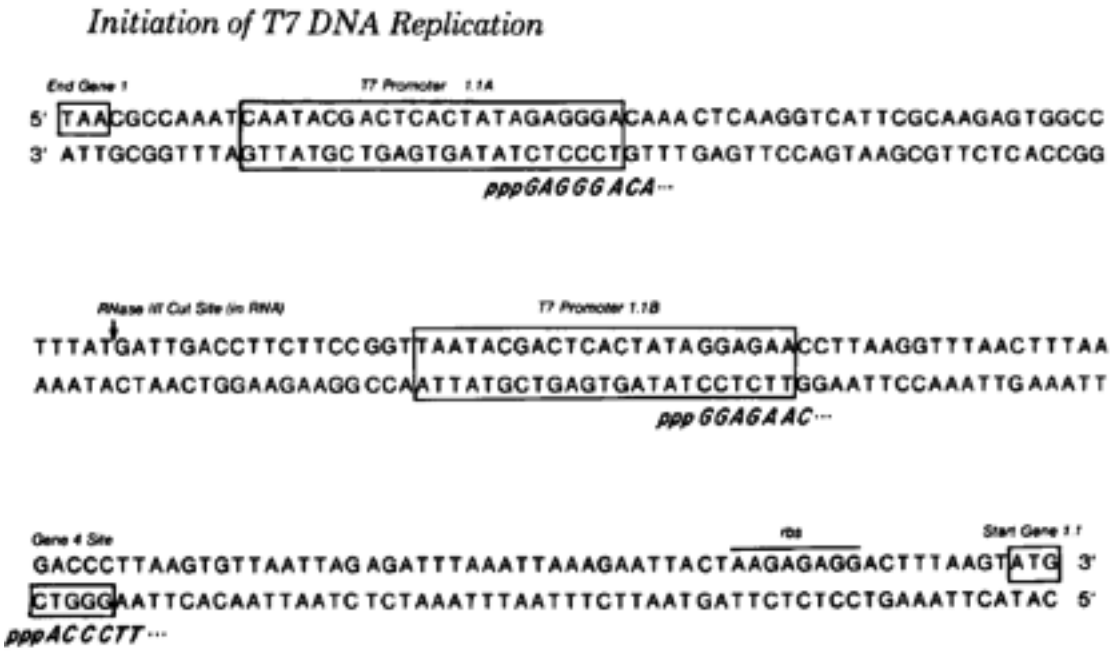

Figure A.1: The primary T7 origin of replication, take from Table 1 of (Fuller and Richardson 1985).

A.3 Replisome primers

Connection	Target Gene	Dir.	Type	Sequence	Name
1.Promoter - DNA PolT7	DNApolT7	F	SD lib	GCATACGTCGGCCCCCGGGGCCTAANNNNNNNNNNRRRRRRRNNNNNDTGatcgtttctgacatcga	T7-1: ProDNApolSD
1.Promoter - DNA PolT7	DNApolT7	F	Genome	GCATACGTCGGCCCCCGGGGCCTAANNNNNNNNNNaggagaaatcaatatgatcgtttct	T7-2: ProDNApolGen
2. Promoter – RNApolT7	Tail of DNA polT7	F	N/A	GCATACGTCGGCCCCCGGGGCCTAAcgatttgccactgatacag	T7-3: ProRNApol
3. DNApolT7 – RNApolT7	DNApolT7	R	Genome	GCATACGTC GCCTGCATGGCctgtatcagtggcaaatcg (RC of: cgatttgccactgatacag GCC**ATGCA**GGC + GCATACGTC)	T7-4: DNARNArev
3. DNApolT7 – RNApolT7	R7NApolT7	F	SD lib	GCATACGTCGCC**ATGCA**GGCNNNNNNNNNNRRRRRRRNNNNNDTGaacacgattaacatcgcta	T7-5: DNAtoRNASD
3. DNApolT7 – RNApolT7	RNApolT7	F	Genome	GCATACGTCGCC**ATGCA**GGCNNNNNNNNNgaagaggcactaaaacacgattaac	T7-6: DNAtoRNAgen
4. RNApolT7 – ssBP	RNApolT7	R	N/A	GCATACGTC GCCGTGCAGGCttacgcgaacgcgaagtccg (RC of: cggacttcgcgttcgcgtaa GCC**TGCA**CGGC + GCATACGTC)	T7-7: RNAssBPrev
4*. RNApol – vector	RNApolT7	R	N/A	GCATTGCTG**GGCCGGCAGGGCC**ttacgcgaacgcgaagtccg	T7-8: RNAend
4. RNApolT7 – ssBP	ssBP	F	SD	GCATACGTCGCC**TGCA**CGGCNNNNNNNNNRRRRRRRNNNNNDTGgctaagaagattttcacc	T7-9: RNAssBPSD
4. RNApolT7 – ssBP	ssBP	F	Genome	GCATACGTCGCC**TGCA**CGGCNNNNNNNNNggagattaacattatggctaagaag	T7-10 RNAssBPgen
5.ssBP – Heli/Primase	ssBP	R	N/A	GCATACGTC GCCGCACAGGCttagaagtctccgtcttcgt (RC of: acgaagacggagacttctaa GCCTGTGCGGC + GCATACGTC)	T7-11 ssBPHPrev
5*. ssPB-vector	ssBP	R	N/A	GCATTGCTG**GGCCGGCAGGGCC**ttagaagtctccgtcttcgt	T7-12 ssBPend
5. ssBP – Heli/Primase	Heli/Prim	F	SD	GCATACGTCGCCTGTGCGGCNNNNNNNNNNRRRRRRRNNNNNDTGgacaattcgcacgattc	T7-13 ssBPHPSD
5. ssBP – Heli/Primase	Heli/Prim	F	Genome	GCATACGTCGCCTGTGCGGCNNNNNNNNNNaggagggaattgcatggacaat	T7-14 ssBPHPgen
6. Heli/Primase – LacZ	Heli/Prim	R		GCATACGTC GCCGCTAAGGCtcagaagtcagtgtcgttg (RC of: caacgacactgacttctga GCC**TTAGC**GGC + GCATACGTC)	T7-15 hpLacZ
6.* Heli/Prim-vector	Heli/Prim	R	n/a	GCATTGCTG**GGCCGGCAGGGCC**tcagaagtcagtgtcgttg	T7-16 HPend
7. Heli/Prim-LacZ	LacZ	F	plasmid	GCATACGTCGCC**TTAG**CGGCacagaaacagctATGATAGATCCCGTCG	T7-17 hpLacZrev

Table A.2: Table of PCR primers used to PCR T7 replisome genes. Restriction sites are highlighted in dark green. Start codons are highlighted in bright green. Ribosome binding sites are highlighted in red. Regions complementary to T7 DNA are in lower case. N denotes any base (A, T, G or C) can occupy the spot. R denotes only A or G can occupy the spot. D denotes A, T or G can occupy the spot.

A.4 Additional figures

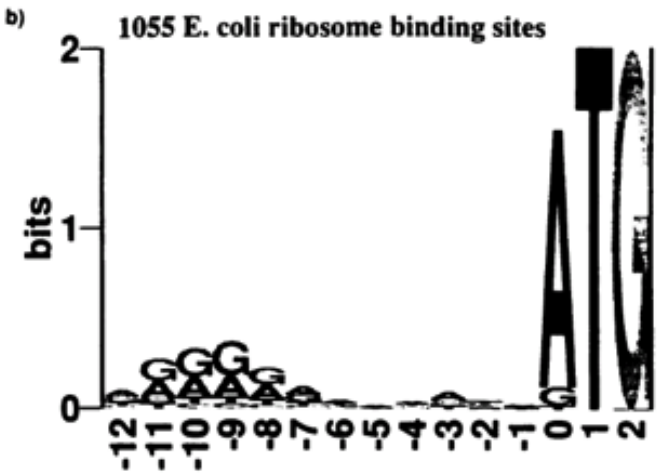

Figure A.2: Table showing the relative frequencies of DNA bases in the region from −12 to +2 (relative to the start codon) for 1055 genes in *E. coli*. Table taken from (Dower et al. 1988)

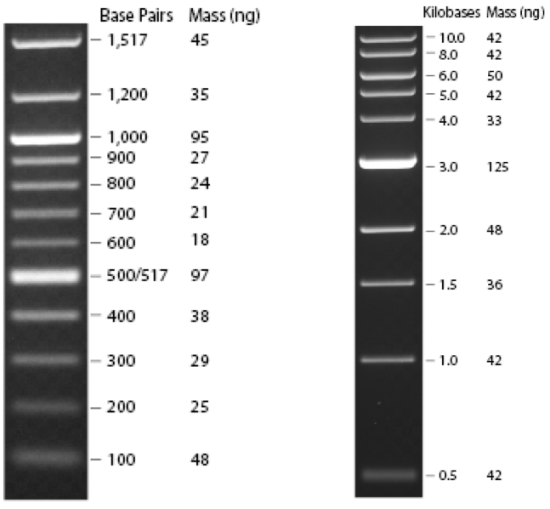

Figure A.3: The 1 kb (left) and 100 bp (right) DNA ladders. Figures taken from www.neb.com

Appendix B

Probability and statistics background

B.1 Probability rules

The notation

$$p(A|B) \tag{B.1}$$

is used to denote the probability of A *given* B. B may be either a random variable or a model parameter.

The sum rule of probability:

$$p(A) = \sum_{B} p(A, B) \tag{B.2}$$

The product rule of probability:

$$p(A, B) = p(A|B)p(B) = p(B|A)p(A) \tag{B.3}$$

A and B are independent variables if

$$p(A, B) = p(A)p(B) \tag{B.4}$$

For N independent and identically distributed (IID) data points:

$$p(d_1, d_2, \ldots, d_n) = \prod_{n=1}^{N} p(d_n) \tag{B.5}$$

$$\log p(d_1, d_2, \ldots, d_n) = \sum_{n=1}^{N} \log p(d_n) \tag{B.6}$$

B.2 Squared error and maximum likelihood

Minimization of squared loss is most commonly derived in the natural sciences by asserting that 'error', the difference between parameterized model prediction and experimental data, is additive, normally distributed, and independent for each example (here indexed by i):

$$d_i = f_\theta(x_i) + \xi_i; \quad \xi_i \sim \mathcal{N}(\xi|0, \sigma). \tag{B.7}$$

This notation emphasizes that the model f depends on parameters θ, and the $\sim$ indicates the distribution from which the error ξ_i on the i^{th} observation is drawn (i.e., the Gaussian or normal distribution and variance σ). Assuming

independent and identically distributed observations, the probability of all the N data $\mathbf{D} = \{d_i\}_{i=1}^{i=N}$ is then the *likelihood*

$$L = p(\mathbf{D}|\theta) = \Pi_{i=1}^{i=N} \mathcal{N}(y_i - f_\theta(x_i)|0, \sigma) = \frac{\mathrm{e}^{-\chi^2}}{(\sqrt{2\pi\sigma})^N} \tag{B.8}$$

with the usual $\chi^2 = \sum_{i=1}^{i=N} (d_i - f_\theta(x_i))^2/\sigma^2$ arising as a linear term in the logarithm of the likelihood ℓ:

$$\ell \equiv \ln L = -\chi^2 + \frac{N}{2} \ln 2\pi\sigma. \tag{B.9}$$

Minimization of χ^2, is thus derived from the more general principle of ML: the parameters θ_* chosen are those which are the most likely.

B.3 "BIC": an intuition-building heuristic

Often, explicit calculation of $p(\mathbf{D}|\mathrm{K})$ is computationally difficult, and one resorts to approximation. For example, if the likelihood $p(\mathbf{D}|\vec{\theta}, \mathrm{K})$ is sharply and uniquely peaked as a function of $\vec{\theta}$, meaning that there is one unique maximum, Schwartz (Schwarz 1978) suggested a pair of approximations: (i) Taylor expansion of $\ell(\vec{\theta})$ (from Eq. B.9) and Laplace approximation of the integral; and (ii) replacing the second derivative of $\ell(\vec{\theta})$ by its asymptotic behavior in the limit $\{K, N\} \to \infty$. The first approximation reads

$$p(\mathbf{D}|\mathrm{K}) = \int d^{\mathrm{K}}\vec{\theta} \mathrm{e}^{\ell(\vec{\theta})} p(\vec{\theta}|\mathrm{K}) \approx \mathrm{e}^{\ell_*} p(\vec{\theta}_*|\mathrm{K}) \frac{(2\pi)^{K/2}}{|H|^{1/2}} \tag{B.10}$$

where $\ell_* = \ell(\vec{\theta}_*)$ is the ML over all parameters $\vec{\theta}$, and the $K \times K$ matrix H, also termed the Hessian, is the matrix of derivatives (evaluated at $\vec{\theta}_*$)

$$H_{\alpha\beta} \equiv \frac{\partial^2 \ell(\tilde{\boldsymbol{\theta}})}{\partial\vec{\theta}_\alpha \partial\vec{\theta}_\beta}. \tag{B.11}$$

In the case of N independent data points the derivative of ℓ is a sum of N independent terms, and the determinant of the Hessian scales as N^K in the limit of infinite data N and infinitely many K equally-important parameters $\vec{\theta}_\alpha$. Under this pair of asymptotic approximations, then,

$$p(\mathbf{D}|\mathrm{K}) \approx \mathrm{e}^{\ell_*} p(\vec{\theta}_*|\mathrm{K}) \frac{(2\pi)^{K/2}}{|H|^{1/2}} \approx C(K,N)\mathrm{e}^{(\ell_* - (K/2)\ln N)}. \tag{B.12}$$

The exponent is sometimes referred to as the *Bayesian Information Criterion* or BIC; for clarity it is worth noting, though, that it does not depend on the prior (the most common meaning of the adjective 'Bayesian' in statistics) and that it is derived without any appeal to or use of information theory. The usage of such an algebraic expression alone, ignoring the possible dependence of terms lumped into $C(K,N)$ (i.e., treating $C(K,N)$ as a constant) is a simple[1], intuitive, and appealing approach to model selection. The increase in ℓ_* as K increases is penalized by the term $-(K/2)\ln N$, selecting the optimal model indexed by K_*, the maximizer of the BIC.

In the case of FRET data the likelihood is complicated by the presence of a hidden state z_i (the discrete conformational state of the molecule which gives rise to the observed FRET ratio), meaning that the evidence $p(\mathbf{D}|\mathrm{K})$ has the richer formulation (suppressing the cluttering superscripts K on the hidden and manifest variables z and $\vec{\theta}$, respectively)

$$p(\mathbf{D}|\mathrm{K}) = \sum_{\mathbf{Z}} \int d^{\mathrm{K}}\vec{\theta}\, p(\mathbf{D},\mathbf{Z}|\vec{\theta},\mathrm{K}) p(\mathbf{Z}|\vec{\theta},\mathrm{K}) p(\vec{\theta}|\mathrm{K}). \tag{B.13}$$

[1] Note that, although use of the BIC obviates determining many facets of one's model and its relation to the data, we still need to know the error bars σ, which appear in ℓ.

This rich structure renders completely inappropriate the assumptions of the BIC derivation above: among other problems, the hidden variables will be modeled by a Markovian dynamic, coupling each of the example data (and thus violating the assumption of N independent data); and the permutation symmetry of the labels on these violates the assumption that the likelihood is sharply and singly peaked – rather there are $K!$ such peaks from the possible relabelings of the states.