Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2017 October 18.

Published in final edited form as:

J Am Stat Assoc. 2016 ; 111(515): 951–966. doi:10.1080/01621459.2016.1140050.

Analyzing Single-Molecule Protein Transportation Experiments via Hierarchical Hidden Markov Models

Yang Chen,
Ph.D. candidate, Department of Statistics, Harvard University, Cambridge, MA 02138

Kuang Shen,
Pfizer fellow of the Life Sciences Research Foundation, Whitehead Institute for Biomedical Research, Cambridge, MA 02142

Shu-Ou Shan, and
Professor, Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125

S. C. Kou
Professor, Department of Statistics, Harvard University, Cambridge, MA 02138

Abstract

To maintain proper cellular functions, over 50% of proteins encoded in the genome need to be transported to cellular membranes. The molecular mechanism behind such a process, often referred to as protein targeting, is not well understood. Single-molecule experiments are designed to unveil the detailed mechanisms and reveal the functions of different molecular machineries involved in the process. The experimental data consist of hundreds of stochastic time traces from the fluorescence recordings of the experimental system. We introduce a Bayesian hierarchical model on top of hidden Markov models (HMMs) to analyze these data and use the statistical results to answer the biological questions. In addition to resolving the biological puzzles and delineating the regulating roles of different molecular complexes, our statistical results enable us to propose a more detailed mechanism for the late stages of the protein targeting process.

Keywords

Protein targeting; conformational change; FRET; hierarchical model; HMM (hidden Markov model); MCMC (Markov Chain Monte Carlo); model checking

1 Introduction

In cells, proteins often need to be transported to appropriate destinations inside or outside of a cell in order to maintain proper cellular functions (Rapoport, 2007). In fact, over 50% of all proteins encoded in the genome need to be properly localized from the site of their synthesis (Lodish et al., 2000; Rapoport, 1991). Co-translational protein targeting is such a process in which proteins still being synthesized on the ribosome (called ribosome nascent-

Correspondence to: Shu-Ou Shan; S. C. Kou.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

chain complex or RNC) are transported to the membrane. This is achieved by the collaboration of a signal recognition particle (SRP) in the cytoplasm and its receptor (SR) located on the endoplasmic reticulum (ER) membrane. It is known that the co-translational protein targeting process consists of four basic steps (Zhang et al., 2009b; Nyathi et al., 2013), as schematically illustrated in Figure 1. First, SRP recognizes and binds the signal sequence on the RNC. Second, SRP forms a complex with SR on the membrane, bringing the RNC-SRP complex to the membrane surface (here, an RNC-SRP-SR ternary complex is formed near the membrane). Third, the RNC is released from the SRP-SR complex and docks on the protein conducting channel, known as the translocon. Fourth, SRP and SR dissociate (through GTP-hydrolysis) to enter a new round of protein targeting; at the same time, the nascent polypeptide chain goes through the translocon on the membrane.

While the four steps give the big picture, the detailed molecular mechanisms of the protein targeting process remained unclear (Shen et al., 2012). One particularly puzzling question arises from the earlier observation that SRP and the translocon bind the same sites on the ribosome and the signal sequence; thus, the bindings of the targeting and translocation machineries to RNC are mutually exclusive. How do these two machineries exchange on the RNC, and how do they accomplish this without losing the RNC (which aborts the pathway)? Recent biochemical, structural, and single-molecule work (Zhang et al., 2008; Shen and Shan, 2010; Ataide et al., 2011; Voigts-Hoffmann et al., 2013; Nyathi et al., 2013; Akopian et al., 2013b) offered valuable clues to this question. These works showed that the SRP-SR complex can undergo a large-scale structural change and visit an alternative state in which the proteins in the SRP-SR complex are moved away from their initial binding site on the ribosome (see Figure 4 below); this provides a potential mechanism to enable a step-wise exchange with the translocon.

To provide direct evidence for this mechanism and resolve its molecular details, single-molecule experiments on the prokaryotic SRP system were conducted by the Shan group. Single-molecule experiments are one of the major experimental breakthroughs in chemistry and biophysics in the last two decades: using advanced tools in optics, imaging, fluorescence tagging, biomolecule labeling, etc., researchers are able to study biological processes on a molecule-by-molecule basis (Moerner, 2002; Nie and Zare, 1997; Tamarat et al., 2000; Weiss, 2000; Xie and Trautman, 1998; Xie and Lu, 1999; Qian and Kou, 2014). Under single-molecule experiments, transient excursions of molecules to alternative structures can be directly visualized, rather than lost in the statistical averaging of bulk experiments.

The single-molecule experiments under our study employ an experimental technique, FRET (Föster resonance energy transfer) (Roy et al., 2008), which uses resonance energy transfer as a molecular ruler to track the dynamic movement of a molecule in distinct conformational states, providing information on the pathway, kinetics and equilibrium of the structural transitions of molecules. The experimental data consist of hundreds of FRET trajectories, three of which are shown in Figure 2. Each FRET trajectory is a time series $(y_1, y_2, \ldots)$. These experimental FRET trajectories provide crucial information on the structural dynamics for us to resolve the questions regarding the underlying mechanism of protein targeting. We will describe the experimental details as well as the molecular structures in Section 2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

From the hundreds of traces collected, we can clearly see a low FRET state and a high FRET state in each trace, with one or more possible intermediate states. Several critical questions arise regarding the correct interpretation of the data.

1. Molecular behavior is inherently stochastic. Ensembles of molecules that are chemically identical will vary in their behavior at the single-molecule level (in a manner predicted by the Boltzmann distribution). Thus, individual single molecule traces are inherently heterogeneous. In addition, due to the experimental limitations, such as uneven laser illumination, each FRET trajectory has its own FRET values and length. Moreover, it is possible that some observed molecules are partially damaged during sample preparation or application. Therefore, we want to carefully examine the homogeneity/heterogeneity of the data set: Does the collection of FRET trajectories represent chemically homogeneous molecules or molecular complexes? If not, is the heterogeneity biologically relevant?

2. How many states are there in these FRET trajectories? Previous analysis utilized an arbitrary number of states for HMM (Shen et al., 2012). However, there is no statistical analysis to legitimate that number. A careful analysis is needed to unravel the existence of intermediate state(s) from the noisy experimental data; this information is critical, as it reflects possible pathways through which the SRP-SR undergoes its structural transitions.

3. Are these intermediates on-pathway or off-pathway? In other words, during the transition from the low FRET state to the high FRET state, must or may not the trajectory go through one or more intermediate state(s)? Clarifying the transition pathway will differentiate between different mechanisms. In one model, often termed trial-and-error, the intermediate states are “mistakes” made by the complex as it searches for alternative structures. This model predicts that the molecules must return from the intermediate back to the low FRET state before transitioning to the high FRET state. In an alternative model, the active-searching model, the intermediate FRET state(s) represent on-pathway intermediate(s) through which the SRP-SR complex attains the high FRET state. This model predicts that most of the successful low-to-high or high-to-low FRET transitions occur via the intermediate state(s).

4. During the protein targeting process, RNC and translocon regulate the conformation of the SRP-SR complex. This was also observed in the single-molecule experiments. Addition of RNC or translocon changes the equilibrium and kinetics via which the SRP-SR complex transits between the different FRET states, as reflected by altered frequency and durations of these transitions. However, as individual single-molecule traces are stochastic due to a combination of inherent and experimental limitations (as explained in question 1), it is not possible to accurately extract kinetic and equilibrium information from individual trajectories. Rigorous statistical analysis using the information from all trajectories is required to extract this information and understand

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

whether the RNC and translocon change the conformational space of the SRP-SR complex, and if so, how.

With these questions posed, we employ a hidden Markov model (HMM), modeling each trajectory (y_1, y_2, …) as originated from a hidden Markov chain. The parameters governing the hidden Markov chain, such as the number of distinct states and the transition probabilities, capture the molecular conformations and dynamics of the underlying biological processes.

We note that the analysis of *individual* FRET trajectories based on HMMs has been considered in the biophysical community (Rabiner, 1989; Eddy, 1996; Liu et al., 2010). Software packages *HaMMy* (McKinney et al., 2006) and *SMART* (Greenfeld et al., 2012) give the maximum likelihood estimators of parameters for a *single* trajectory using the EM/Baum-Welch algorithm (Baum and Petrie, 1966; Baum et al., 1970; Dempster et al., 1977). Variational Bayes method is also suggested in the FRET data analysis, which incorporates prior information about the range of parameter values into the model fitting (Bronson et al., 2009). Empirical Bayes methods (van de Meent et al., 2014) and bootstrap methods (König et al., 2013) have also been proposed for the analysis of FRET data.

The information from individual FRET trajectories is rather limited, mainly due to the low signal-to-noise ratio and the limited observation time of each individual molecule (before its pho-tobleaching). Consequently, the inference based on single FRET trajectories is highly variable and unreliable in the sense that even for FRET trajectories recorded under the same experimental condition, heterogeneities of estimated parameters and the estimated number of hidden states across trajectories are apparent. Experimentalists address this issue by performing hundreds of replicate experiments. Quantifying cross-sample variability has recently drawn attention among the biophysics community (König et al., 2013; van de Meent et al., 2014). How to pool information from these replicate experimental trajectories as well as to account for their heterogeneity is the key statistical question.

Two statistical questions naturally arise in our analysis of the FRET trajectories: (1) the determination of the total number of hidden states and (2) a robust and reliable estimation of model parameters by pooling information from “seemingly” heterogeneous FRET trajectories obtained from the same experimental condition.

The first quesiton, which is a preliminary step of building models to pool information from multiple trajectories, has been widely studied in the statistics and chemistry literature (Finesso, 1990; Leroux, 1992; Ryden, 1995; Blanco and Walter, 2010; Bulla et al., 2010). We adopt a population approach based on the Bayesian information criterion, which estimates the number of hidden states by the majority rule (e.g., if the majority of the FRET trajectories under the same experimental condition shows three states, then the method selects three as the number of hidden states). This approach actually has been recommended in the chemistry literature (Watkins and Yang, 2005) and is described in Section 3, which also discusses our fitting of HMM to individual FRET trajectories.

Second, we propose a hierarchical model on top of the HMMs to combine information from multiple trajectories. The hierarchical model embodies the biological intuition that the same

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

dynamics underlies all the experimental replicates, but each replicate is a noisy realization of the common process due to intrinsic/experimental fluctuation and noise. The hierarchical HMM enables us to not only robustly estimate the parameters from the common dynamics but also fit the individual trajectories better than if fitted individually. Section 4 describes in detail our hierarchical HMM and how we use it to combine information from individual trajectories. Simulation studies demonstrating that the hierarchical model can work effectively under low signal-to-noise ratio, which is very difficult to analyze if one only fits individual trajectories.

From an applied angle, our statistical analysis of the experimental FRET data leads to a resolution of several questions about the protein targeting process that are described above. The model fitting and biological implications are discussed in Section 5, at the end of which (Section 5.4) we are able to provide a detailed molecular mechanism of the co-translational protein targeting process. Model assessment is conducted in Section 6. We conclude this article in Section 7 with a summary. The appendix contains the technical details of our computation and Monte Carlo sampling.

2 Single-molecule experiments on co-translational protein targeting

2.1 Single-molecule FRET experiments

The single-molecule experiments use the FRET technique to study the protein targeting process. FRET tracks in real time the distance and orientation between two microscopic tags, a donor fluorophore and an acceptor fluorophore, placed in a molecular complex (Roy et al., 2008). It is often the case that the experimentalists cannot directly observe the structural change of a bio-molecule. The FRET recording, on the other hand, measures the distance changes of the two tags on the bio-molecule and thus reveals the structural changes during a biological process.

Each experimental FRET trajectory is a time series (y_1, y_2, …), obtained at every 30 millisecond (ms) in our case. $y_i \in [0, 1]$ is calculated as y_i = acceptor fluorescence / (donor fluorescence + acceptor fluorescence). A high FRET value y_i implies that the two tags, the donor and acceptor, are close to each other, while a low FRET value means the donor and acceptor are far apart. A sample FRET trajectory is shown in Figure 3. On the top panel, the red curve is the acceptor fluorescence and the green curve is the donor fluorescence. The black curve in the lower panel shows the FRET values, i.e., the ratio of acceptor fluorescence over the total fluorescence.

2.2 FRET on bacterial SRP system

In this subsection, we give the necessary background on the molecular structure of our experimental system and how FRET reveals information about protein targeting.

Single-molecule FRET technique was used to study the bacterial SRP system. The bacterial SRP is comprised of two subunits: an RNA segment (the SRP RNA) and an Ffh protein. Ffh contains two domains connected by a flexible linker: the M-domain binds tightly to the SRP RNA near its capped (tetraloop) end and recognizes the signal sequence on the nascent protein; the NG-domain interacts with the SRP receptor, termed FtsY in bacteria, and binds

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

a ribosomal protein at the “exit site” where the nascent protein emerges from the ribosome. We will use Ffh-M and Ffh-NG to denote the M- and NG- domains of Ffh (Akopian et al., 2013b; Halic et al., 2004; Keenan et al., 2001; Zhang et al., 2008). The SRP RNA has an elongated structure: it stretches over 100 Å (angstrom) from one end (the capped end) to the other end (the distal end). Figure 4 illustrates ° the *E.coli* SRP and SR.

When the SRP-SR complex is formed, Ffh-NG binds FtsY (step 2 in Figure 1). In a single-molecule experiment, we placed a FRET donor at Ffh-NG or FtsY and a FRET acceptor at the distal end of RNA. The resulting FRET trajectory tracks the movement of the FtsY-[Ffh-NG] complex along the RNA in real time: a low FRET value implies the FtsY-[Ffh-NG] complex is far from the RNA distal end, whereas a high FRET value implies the FtsY-[Ffh-NG] complex is close to the RNA distal end. See C and D of Figure 4 for illustration (where the FRET donor is the green star and the FRET acceptor is the red star). The FRET tracking provides direct information on the structural change of SRP-SR complex critical for the biological process. It is known that the FtsY-[Ffh-NG] complex initially assembles at the RNA capped end (the low FRET state of Figure 4(C)), where it excludes the translocon from binding RNC. When this complex moves to the RNA distal end (the high FRET state of Figure 4(D)), the ribosome is vacated to allow translocon binding, and disassembly of the FtsY-[Ffh-NG] complex is triggered (Shen and Shan, 2010; Ataide et al., 2011). Therefore, from the FRET trajectory, we know when the SRP-SR complex is positioned for assembly or disassembly, and when ribosome-translocon contacts are enabled.

To study how the RNC and translocon regulate the structural change on the SRP-SR complex, two more sets of single-molecule FRET experiments were done: one with RNC, SRP and SR, the other with all four components: translocon, RNC, SRP and SR. Together, these experiments reveal the functional role of RNC and translocon in the protein targeting process. Table 1 summarizes the four sets of data labeled *Ffh-Data, FtsY-Data, RNC-Data* and *Translocon-Data* obtained from these experiments, and Table 2 summarizes the lengths of the trajectories in each data set. We will analyze and discuss these data starting from Section 3.

2.3 More experimental details

This subsection gives the experimental details. A statistics oriented reader can skip it and directly go to the statistical analysis in Section 3.

2.3.1 Sample preparations—Single cysteine mutants of Ffh and FtsY were expressed and purified in bacterial cells and were subsequently labeled with Cy3-maleimide by the thiol side chain. Labeling reaction was carried out in 50 mM KHEPES (pH 7.0), 300 mM NaCl, 2 mM EDTA, 10% glycerol at room temperature for 2 hours. Free dyes were removed by a gel filtration column. Labeled SRP RNA was prepared by annealing a Quasar670-labeled DNA splint with a T7-transcribed RNA. All the labeled protein or RNA was tested using a well-established GTP hydrolysis assay, and showed no functional difference with wildtype protein or RNA.

2.3.2 Single molecule instrument—All the experiments were carried out on a home-built objective-type TIRF microscope based on an Olympus IX-81 model. Green (532nm)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and red (638nm) lasers were aligned and focused on the sample in a 100 × oil immersed objective. Cy3 and Quasar670 signals were split by a dichroic mirror and were simultaneously imaged using an Ixon 897 camera through DV2 Dualview. Data points were recorded at 30 milliseconds time resolution.

2.3.3 Single molecule assay—Before conducting experiments, all protein samples were ultracentrifuged at 100,000 rpm in a TLA100 rotor for an hour to remove possible aggregates. PEGylated slides and coverslips were assembled into a flowing chamber, in which fluorescent molecules were attached through biotin-neutravidin interaction.

SRP complexes were assembled in SRP buffer and diluted to 50 picomolar in imaging buffer with oxygen scavenging system (saturated Trolox solution containing 50 mM potassium-HEPES (pH 7.5), 150 mM KOAc, 2 mM $Mg(OAc)_2$, 2 mM DTT, 0.01% Nikkol, 0.4% glucose and 1% Gloxy), flowed onto the sample chamber and incubated for 5 minutes before imaging. Movies were recorded at 30 milliseconds time intervals for up to 3 minutes until most fluorescent molecules were photobleached.

2.3.4 Data aquisition—Single molecule data were initially processed by scripts written in IDL and Matlab. Fluorescent peaks in the images were identified and traced throughout the movie. Fluorescent trajectories that showed a single donor bleaching event, which implied single-molecule attachment, and no photoblinking event, were hand-picked for subsequent data analysis. The background was subtracted using the residual fluorescent intensities in both channels, after the fluorophore has been photobleached.

3 Preliminary analysis of individual trajectories

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ be an observed experimental FRET trajectory. We model it as a hidden Markov model (HMM):

$$y_i|(z_i=k)\sim N(\mu_k, \sigma_k^2), \quad (1)$$

where $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)$ are the hidden Markov states, evolving according to a K-state Markov chain. Although, rigorously speaking, the FRET value y_i is between 0 and 1, the Gaussian assumption is widely used and accepted in the single-molecule FRET literature in that with moderate observational noise Gaussian distribution is a good approximation (Dahan et al., 1999; McKinney et al., 2006; Liu et al., 2010). The distinct states of z_i, K in total, model the different conformations of a biological complex. A conformation is a specific 3D structure of a protein or a protein complex. For example, the low- and high-FRET states in C and D of Figure 4 correspond to two distinct conformations of the SRP-SR complex. Let $\boldsymbol{P} = (P_{ij})$ be the $K \times K$ transition matrix of $\boldsymbol{z}$; it represents the conformational kinetics of a complex. For each FRET trajectory, the parameters are $\boldsymbol{\theta}=(\boldsymbol{P}, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$, where μ_k and σ_k^2 are the mean and variance of the FRET value at state k; $k = 1,\cdots, K$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ be the probabilities that the first hidden state z_1 is in state $1,\cdots, K$. The joint likelihood of observations $\boldsymbol{y}_{1:N}$ and the hidden states $\boldsymbol{z}_{1:N}$ is

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$p(\boldsymbol{y}_{1:N}, \boldsymbol{z}_{1:N}|\boldsymbol{\theta}) = \pi_{z_1} \prod_{n=2}^{N} p(z_n|z_{n-1}, P) \prod_{n=1}^{N} p(y_n|z_n, \boldsymbol{\mu}, \boldsymbol{\sigma}^2).$$

Please note that for notational ease, we use $\boldsymbol{y}_{m:n}$ to denote the vector $(y_m, y_{m+1}, \ldots, y_n)$ for $m < n$ throughout this article. The marginal likelihood $L(\boldsymbol{\theta}|\boldsymbol{y}_{1:N}) = \int p(\boldsymbol{y}_{1:N}, \boldsymbol{z}_{1:N}|\boldsymbol{\theta}) d\boldsymbol{z}_{1:N}$ is given by integrating out $\boldsymbol{z}_{1:N}$ in the joint likelihood.

3.1 Infer the parameters with a given number of total states

For each FRET trajectory, for a given K, we can use the Baum-Welch algorithm (Baum and Petrie, 1966; Baum et al., 1970), or equivalently, the EM algorithm (Dempster et al., 1977), to calculate the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$. The Baum-Welch/EM algorithm, in addition, can yield the marginal likelihood evaluated at the MLE, $L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{1:N})$. Appendix A gives the details of our implementation of the algorithm, which uses the forward-backward algorithm.

Alternatively, taking a Bayesian perspective, we can use the Gibbs sampler (Geman and Geman, 1984) together with data augmentation (Tanner and Wong, 1987) to jointly draw posterior samples of the parameters and the hidden states. This gives the posterior distribution (instead of point estimates) of the parameters. Appendix B gives the details of our implementation of the Gibbs sampler with data augmentation.

3.2 Detecting the number of hidden states

At the molecular level, the total number of states K corresponds to the number of conformations accessible to the complex in the experimental duration. The two conformations in C and D of Figure 4 have already been identified in previous studies, and one of our aims is to detect if there are more conformations involved in the protein targeting process (Shen et al., 2012). Statistically, we want to find the K that can "best" explain the variability of the observed FRET trajectories. As an exploratory analysis, we fit each FRET trajectory with the Baum-Welch/EM algorithm for $K = 1, 2, 3, \ldots$ and find that when $K \geq 6$, the hidden states become highly non-identifiable in that the difference of the means of neighboring hidden states are less than 10% of their corresponding standard deviations, which are not experimentally meaningful; and the variance parameters converge to zero, the boundary of the parameter space. Thus, the candidates are $K = 1, 2, 3, 4, 5$ for our data.

Determining K for each trajectory is a model selection problem. Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978) are two popular model selection methods. It is well observed in the literature that AIC has a tendency to overestimate the number of mixture components (Windham and Cutler, 1992; Hawkins et al., 2001; Frühwirth-Schnatter, 2006), which we also observe in our simulations. Thus, we focus on using the BIC in our study, which is known to be consistent (as the sample size goes to infinity) for mixture models (McLachlan and Peel, 2005; Frühwirth-Schnatter, 2006; Biernacki et al., 1998; Leroux, 1992). Though the consistency of BIC for Gaussian HMMs has not been completely established (Cappe et al., 2005; Finesso, 1990; Ryden, 1995), it has been shown through simulations that BIC empirically tends to select the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

correct model when the sample size is large but could give highly variable results when the sample size is small or moderate (Celeux and Durand, 2008; Ryden, 1995; MacKAY, 2002; Watkins and Yang, 2005; Frühwirth-Schnatter, 2006; Keribin, 2000). In the context of FRET trajectories, the variability of BIC for HMMs has also been observed (van de Meent et al., 2014; Blanco and Walter, 2010; Keller et al., 2014). The general recommendation in the statistics literature and in the FRET literature for the state-selection of HMM is to use BIC as a first step of preliminary analysis and then assess the selection result based on scientific and experimental insight (McKinney et al., 2006; Greenfeld et al., 2012; Bulla et al., 2010; Keller et al., 2014; Celeux and Durand, 2008). We adopt this recommendation.

In our case of a K-state HMM, the BIC statistic, denoted by BIC_K, is

$$BIC_K = -2\log L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{1:N}) + \log N \times (K^2 + 2K - 1),$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ and K^2+2K-1 is the total number of parameters: K^2-K for the transition matrix, $2K$ for the mean and variance parameters, $K-1$ for the initial distribution of the first hidden state. Minimizing BIC_K over K gives the BIC selection of K for each trajectory. There are two potential issues with the computation of the BIC statistics: (i) the Baum-Welch/EM algorithm converges to local maximum (Baum et al., 1970; Dempster et al., 1977), and (ii) the likelihood function is unbounded at the boundary of the parameter space for Gaussian mixture models (Chen and Li, 2009). These problems make the choice of initial points of the Baum-Welch/EM algorithm critical (Frühwirth-Schnatter, 2006). We treat them by starting the Baum-Welch/EM algorithm from more than 500 randomly generated initial points: the initial values of the mean parameters $\boldsymbol{\mu}$ are uniformly generated from [0, 1], the initial values of each row of the transition matrix P and the distribution $\boldsymbol{\pi}$ of the first hidden state are independently generated from the Dirichlet distribution with concentration parameters all equal to 1, and the initial values of the standard deviations $\boldsymbol{\sigma}$ are independently generated from uniform distribution on [0.01, 0.3]; these distributions are employed based on the scientific knowledge of the plausible ranges of the parameters. For each of the 500+ initial values, we run the Baum-Welch/EM algorithm until convergence. The minimum of the BIC statistic over the 500+ algorithm outputs is taken as the value of the BIC for model selection. Table 3 tallies the BIC selection of K for the experimental FRET trajectories. Note that we put the *Ffh*- and *Fts Y-Data* together in the first row as they are both designed to study the SRP-SR interaction by itself.

Based on the mode, we select $K = 3$ for the *Ffh*-, *Fts Y*- and *Translocon-Data* and $K = 1$ for *RNC-Data*. Using the estimation mode to select K reflects “majority rule”, i.e., using the consensus to capture the behavior in majority of the experimental replicates. We note that this approach has in fact been proposed in the chemistry literature: Watkins and Yang (2005) showed through simulation and real data studies that it gives a highly robust estimate of K. Note that although we cannot totally rule out the possibility of 4 or more hidden states for some trajectories, we have enough evidence that 3 is the minimum number of K, which the majority of trajectories support. We will see later (in Section 4.2) that $K = 3$ is well supported by the fitting of all the trajectories.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4 Modeling FRET trajectories with hierarchical hidden Markov model

The analysis of individual FRET trajectories reveals that they could have significantly different $\boldsymbol{\theta}$. For instance, a likelihood-ratio test on the three trajectories in Figure 2, which are from the *Ffh-Data*, gives a *p*-value smaller than 0.01, soundly rejecting the hypothesis that the three trajectories share the same $\boldsymbol{\theta}$.

Biologically, the trajectories from replicate experiments under the same condition should reflect the common underlying process. Hence, our goal is to account for the heterogeneity among the experimental trajectories and at the same time to pool information from the trajectories under the same experimental condition. We propose a hierarchical HMM. Suppose $\{\boldsymbol{y}^{(l)}, \boldsymbol{z}^{(l)}\}$ are the observations and hidden states for trajectory l. We assume that the same transition matrix $\boldsymbol{P}$ is shared by all trajectories; for trajectory l, the means $(\mu_1^{(l)}, \ldots, \mu_K^{(l)})$ come from a higher level distribution $\mu_i^{(l)} \sim \mathcal{N}(\mu_{0i}, \eta_{0i}^2)$ with (vector) hyperparameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\eta}_0^2$, and the variances $((\sigma_1^2)^{(l)}, \ldots (\sigma_K^2)^{(l)})$ come from scaled inverse-χ^2 distributions with (vector) hyperparameters $(\boldsymbol{\nu}, \boldsymbol{s}^2)$, where $\boldsymbol{\nu}$ denotes the degrees of freedom and $\boldsymbol{s}^2$ are the scale parameters. The intuition behind this hierarchical HMM is that (i) the transition matrix $\boldsymbol{P}$ represents the conformational kinetics, which is intrinsic to the molecule; it thus should be the same across the trajectories. (ii) The experimental replicates are subject to equipment noise, thermal fluctuation and random variations in experimental samples; the hierarchical structure on $\boldsymbol{\mu}^{(l)}$ and $(\boldsymbol{\sigma}^2)^{(l)}$ reflects it – each trajectory can be considered as a noisy version of the underlying truth. Figure 5 diagrams our hierarchical HMM.

We note that the real experimental trajectories have different lengths: some are quite short. Within a short experimental time window it is possible that not every conformation shows up —some fast transitions and rare states might be missed in short trajectories. To accommodate this we incorporate a set of indicators into our hierarchical HMM: $\boldsymbol{I}^{(l)}$ indicates which states are present in trajectory l. For example, if the maximum number of states is $K = 3$, $\boldsymbol{I}^{(l)}$ can take four values $\boldsymbol{I}^{(l)} = \{1, 2, 3\}$, $\boldsymbol{I}^{(l)} = \{1, 2\}$, $\boldsymbol{I}^{(l)} = \{1, 3\}$ or $\boldsymbol{I}^{(l)} = \{2, 3\}$, corresponding to the states present in trajectory l. Note that we exclude the singletons (such as {1}, {2} or {3}) in the set of possible states, since we know from the preliminary analysis of individual trajectories that there are at least two states in each trajectory of the *Ffh-Data*, *FtsY-Data* and *Translocon-Data*.

Let $N_{i,j}^{(l)}$ be the number of transitions from state i to j in trajectory l; $N_{i,j}^{(l)}=0$ if either state i or j does not appear in trajectory l. The likelihood for trajectory l is

$$p(\boldsymbol{y}^{(l)}, \boldsymbol{z}^{(l)} | \boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}, \boldsymbol{P}, \boldsymbol{I}^{(l)}) = \prod_{i,j=1}^{K} \left(\frac{P_{ij}}{\sum_{k \in I^{(l)}} P_{ik}} \right)^{N_{i,j}^{(l)}} \cdot \prod_{n=1}^{N_l} \mathcal{N}(y_n^{(l)}; \mu_{z_n^{(l)}}^{(l)}, \sigma_{z_n^{(l)}}^{(l)}),$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $\left(\frac{P_{ij}}{\sum_{k\in I^{(l)}} P_{ik}}\right)_{i,j\in I^{(l)}}$ is the re-normalized transition matrix for trajectory l according to which states are present in $I^{(l)}$, and N_l is the length of trajectory l. The likelihood function of all the trajectories (under the same experimental condition) under our hierarchical HMM is

$$\prod_l p(\boldsymbol{y}^{(l)}, \boldsymbol{z}^{(l)} | \boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}, \boldsymbol{P}, I^{(l)}) p(\boldsymbol{\mu}^{(l)} | \mu_0, \eta_0^2) p((\boldsymbol{\sigma}^{(l)})^2 | \nu, s^2).$$

4.1 Estimation under the hierarchical HMM

To obtain the posterior distribution of the parameters in this model, we use MCMC (Liu, 2001) algorithms. The priors are specified as follows. Each row of the transition matrix $\boldsymbol{P}$ has a flat prior (i.e., a Dirichlet distribution with all parameters equal to 1), which is a proper prior. The global parameters μ_0, η_0^2 have flat priors. The categorical variable $I^{(l)}$ also has flat priors, with equal probability of falling into each category. Similar to the Bayesian data augmentation (Tanner and Wong, 1987) procedure for fitting a single trajectory in Appendix B, we augment the parameter space ($\boldsymbol{P}$;μ_0, η_0^2; {$\boldsymbol{\mu}^{(l)}$, $\boldsymbol{\sigma}^{(l)}$; $I^{(l)}$}) with the hidden states {$\boldsymbol{z}^{(l)}$} and sample from the conditional distributions of these two parts iteratively until convergence. The parameters ($\boldsymbol{P}$;μ_0, η_0^2; {$\boldsymbol{\mu}^{(l)}$, $\boldsymbol{\sigma}^{(l)}$; $I^{(l)}$}) are updated one at a time from the conditional distributions using Metropolis-Hastings (for $\boldsymbol{P}$) or Gibbs (for μ_0, η_0^2; {$\boldsymbol{\mu}^{(l)}$, $\boldsymbol{\sigma}^{(l)}$; $I^{(l)}$}). Conditioning on the parameters ($\boldsymbol{P}$, {$\boldsymbol{\mu}^{(l)}$, $\boldsymbol{\sigma}^{(l)}$; $I^{(l)}$}), the hidden states {$\boldsymbol{z}^{(l)}$} are updated sequentially for l = 1, 2 …. The details of the sampling procedure are given in Appendix C.

Figure 6 shows the fitting of our hierarchical HMM with K = 3 to two representative FRET trajectories: one long trajectory from the *Ffh-Data* and one short trajectory from the *Translocon-Data*. The grey curves on the top two panels are the observed experimental FRET values. The solid black lines are the fitted values $\{\hat{\mu}_{\hat{z}_n}\}_{n=1}^N$, where $\hat{\mu}$ and $\hat{\boldsymbol{z}}_n$ denotes the posterior modes from our MCMC sampling. The lower panel plots the histograms of y_i, the FRET values, of the two FRET trajectories. The black curves overlaid on the histograms are the fittings from our hierarchical HMM, using the posterior mode.

4.2 Assessing the number of hidden states with the hierarchical HMM

The posterior distribution of the indicator $I^{(l)}$ gives the probability that a given trajectory l contains a specific collection of states. This posterior distribution thus provides a hierarchical-HMM-based method of model selection: we can allocate the number of hidden states for each trajectory based on the posterior mode of $|I^{(l)}|$, the size of $I^{(l)}$. By combining multiple trajectories and allowing the sharing of information, we potentially obtain more stable model selection results — borrowing information from other trajectories helps identify rarely occurred hidden states for some trajectories.

Table 4 tallies the hierarchical-HMM based assignment of the number of hidden states for the experimental FRET trajectories. We apply the hierarchical HMM separately with K = 3, where the maximum number of states is three, and with K = 4, where the maximum number

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of states is four. Table 4 shows that no matter we set three or four states as the maximum to begin with, the majority of the trajectories are assigned three states. The allocation of states based on the hierarchical HMM, therefore, corroborates our selection of three total states for the *Ffh*-, *FtsY*-and *Translocon-Data*, indicating the robustness of the selection.

4.3 Hierarchical fitting versus individual fitting

It is worth pointing out that by pooling the information from the multiple trajectories, we obtain more robust and reliable estimates. Figure 7 shows what happens if we only fit the individual trajectory by itself. The left panel shows the fitting of the 2-state, 3-state and 4-state HMMs to the long trajectory of Figure 6(A) alone; the right panel shows the fitting to the short trajectory of Figure 6(B) by itself. The individual fitting is seen to be unstable in that it is quite difficult to judge which fitting is better. The hierarchical model, in contrast, allows the information to be pooled from all the trajectories, resulting in stable estimates.

To further compare the fitting under the hierarchical model versus the fitting on individual trajectories and to test the limit of the hierarchical model fitting, we conduct a sequence of simulations. The mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ is generated according to $\mu_1 \sim \mathcal{N}(0.1, 0.1^2)$, $\mu_2 \sim \mathcal{N}(0.4, 0.1^2)$, $\mu_3 \sim \mathcal{N}(0.7, 0.1^2)$. The standard deviation vector $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$ is taken to be $\sigma_1 = \sigma_2 = \sigma_3$. Trajectories each with length $N = 1000$ are generated from a three-state HMM with transition matrix with diagonal elements equal to 0.9 and off-diagonal elements equal to 0.05. For each value of $\sigma_1 = \sigma_2 = \sigma_3 \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8\}$, we repeat the data generation 100 times, so we have 16 sets of simulated data, each set containing 100 trajectories with length 1000.

For each of the 16 sets of simulated data, we apply the hierarchical fitting as well as the individual fitting. Intuitively, as the hierarchical HMM pools information from multiple trajectories, it is able to handle data with much lower signal-to-noise-ratio (SNR) than the fitting of HMM to individual trajectories. Figure 8 provides an illustration, showing the results for the case of $\sigma_1 = \sigma_2 = \sigma_3 = 0.65$. The left panel compares the estimation of the global means $\boldsymbol{\mu}_0 = (0.1, 0.4, 0.7)$. The right panel compares the estimation of the transition probabilities P_{11}, P_{22}, P_{33}. In each panel, the left half shows the posterior distribution under the hierarchical HMM, and the right half shows the aggregated posterior distribution based on fitting the 3-state HMM to individual trajectories. It is evident that individual fitting gives highly variable and biased estimates; in contrast, by pooling the information from the 100 trajectories together, the hierarchical fitting gives much more reliable and accurate estimates.

Formally, for each trajectory we can define SNR as $SNR = \min_k \{\frac{u_{k+1} - \mu_k}{\sigma_k}, \frac{\mu_{k+1} - \mu_k}{\sigma_{k+1}}\}$ (Greenfeld et al., 2012; Hawkins et al., 2001). For the 100 trajectories of Figure 8, the median SNR is 0.3. In contrast, we find from our 16 simulated data sets that for individual fitting to give meaningful result, the median SNR has to be as high as 2.0. As the standard deviation increases, the SNR decreases. Intuitively, as the SNR becomes smaller and smaller, eventually the hierarchical model fitting will start to break down. In our simulation, we observe that the breakdown happens at $\sigma_1 = \sigma_2 = \sigma_3 = 0.7$, where the median SNR is less than 0.3. This number is in sharp contrast with the SNR limit of around 2.0 for the individual trajectory fitting. For the experimental data, the median SNR is 1.47 for the *Ffh-Data*, 1.36

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

for the *FtsY-Data*, and 1.46 for the *Translocon-Data*; all three are below the SNR limit of around 2.0 for reliable individual-trajectory fitting.

5 Resolving the biological questions

Based on our analysis of the single-molecule FRET data, we will address in this section the unsolved questions regarding the detailed mechanism of the protein targeting process put forward in Section 1, delineating the roles of different components in the protein targeting process. We will consider first the conformation change of the SRP-SR complex without RNC or translocon, and then the effect of RNC and translocon in regulating the protein targeting process. Based on the results of our data analysis, we will propose a refined mechanism for co-translational protein targeting process, addressing the biological puzzles.

It is worth pointing out that the hierarchical structure enables us to include heterogeneous trajectories in a single model, capturing common characteristics while allowing for individual variabilities. Our analysis allows us to distinguish between two possibilities that could give rise to the heterogeneous FRET trajectories: (i) heterogeneity of sample, meaning that the SRP-SR complex can exist in distinct populations that have different structural and chemical properties, therefore exhibiting different kinetic and equilibrium behaviors; and (ii) intrinsic noise due to the stochastic nature and molecular reactions and limited time scale for sampling in single-molecule experiments. Our result supports that the heterogeneous trajectories are well explained by (ii).

5.1 Conformational change of the SRP-SR complex

The *Ffh-Data* and *FtsY-Data* are obtained from the single-molecule FRET experiments on the SRP-SR complex in the absence of RNC or translocon. The only difference between these two datasets is the placement of the FRET donor. For the *Ffh-Data* the FRET donor is placed at Ffh-NG, while for the *FtsY-Data* the FRET donor is placed at FtsY; see Figure 4 and Table 1. These data reveal the conformational fluctuation of the SRP-SR complex without RNC or translocon.

As we described in Sections 3.2 and 4.2, three FRET states are detected, corresponding to three conformations. For these three conformations, Table 5 lists the 95% posterior intervals of the global parameters μ_{0i} and η_{0i} for the data sets. The state with a low FRET value, $\mu_{0,1} \approx 0.1$, corresponds to the conformation where the FtsY-[Ffh-NG] complex is near the capped end of the RNA (see C of Figure 4). The state with a high FRET value, $\mu_{0,3} \approx 0.6 \sim 0.8$, corresponds to the conformation where the FtsY-[Ffh-NG] complex is near the distal end of the RNA (see D of Figure 4). It is noteworthy that in addition to these two major conformations, our analysis identifies a “middle” state with the FRET value $\mu_{0,2}$ around 0.3 to 0.4, suggesting a third conformation of the SRP-SR complex. This conformation might correspond to alternative modes of docking of the FtsY-[Ffh-NG] complex at the RNA distal end (in which FtsY-[Ffh-NG] is oriented differently relative to the RNA), given the relative large value of $\mu_{0,2}$, or an alternative binding site of the FtsY-[Ffh-NG] complex on the RNA (Shen et al., 2012). As we shall see shortly, this conformation could serve as an intermediate stage that mediates the large scale movement of the FtsY-[Ffh-NG] complex, which travels 100 Å from the RNA capped end to the distal end.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Figure 9 compares the distributions of the mean parameters for the *Ffh-Data* to those for the *FtsY-Data*. It is also interesting to note from both Table 5 and Figure 9 that the FRET value $\mu_{0,3}$ of the *FtsY-Data* is higher than that of the *Ffh-Data*. This implies that FtsY is closer to the distal end than Ffh-NG is when the FtsY-[Ffh-NG] complex docks at the distal end. It thus gives a fine picture of the relative positions of FtsY and Ffh-NG as shown in Figure 4. This is consistent with findings from the crystal structures of the SRP-SR complex (Ataide et al., 2011; Voigts-Hoffmann et al., 2013).

The conformational change that SRP-SR undergoes on the RNA is unusually large, spanning over 90 Å. How this large-scale movement occurs is an interesting question. It is possible that the complex travels along the RNA via “intermediate” stops. Alternatively, the complex could constantly sample alternative potential docking sites on the RNA until it finds the distal site. The transitions among different states capture the pathways and mechanisms by which the SRP-SR complex undergoes the large-scale conformation change. Table 6 shows our estimates of the transition probabilities $\{P_{ij}\}$ for the data sets. We note that the estimates of the transition probabilities from the *Ffh-Data* are similar to those from the *FtsY-Data*.

We next investigate the functional role of the middle state based on the posterior distributions of $\{P_{ij}\}$ for the *Ffh-Data*. First, we obtain the 95% credible interval of $d_i = 1/(1 - P_{ii})$, the mean dwell time at state i. The intervals are [0.966, 1.057] seconds for d_1, the low-FRET state; [0.228, 0.249] seconds for d_2, the middle state; and [0.465,0.507] seconds for d_3, the high-FRET state. The observation that both d_1 and d_3 are significantly larger than d_2 indicates that the SRP-SR complex spends less time at the middle state than at the low- or high-FRET state, which are more stable.

Second, it is known that biologically the SRP-SR complex initially assembles at the RNA capped end and the complex disassembles at the RNA distal end (Shen and Shan, 2010). Thus, a “complete transition” is the one that goes from the low-FRET state to the high-FRET state (see Figure 4). The observation that P_{13} is significantly smaller than P_{12} suggests that a direct transition from the low-FRET state to the high-FRET state is quite infrequent; rather, a “complete transition” more frequently proceeds through the middle state. In other words, without RNC or the translocon, the FtsY-[Ffh-NG] complex usually travels from the capped end to the distal end through an intermediate stage.

In fact, we can calculate the probability that a final passage from state 1 to state 3 goes through state 2 versus the probability that such a final passage does not go through state 2 as follows. For $i, j = 1, 2$, let us use $P_{i\to j}^{(k)}$ to denote the probability of transition from state i to state j in k steps without ever reaching state 3. Then the probability of going from state 1 to state 3 finally through state 2 is $\sum_{k=1}^{\infty} P_{1\to 2}^{(k)} P_{23}$ (i.e., taking any number of steps between state 1 and 2 and then finally reaching state 3 from state 2 in the last step). The probability of going from state 1 to state 3 not finally through state 2 is $P_{13} + \sum_{k=1}^{\infty} P_{1\to 1}^{(k)} P_{13}$. $P_{i\to j}^{(k)}$ satisfies the following recursive formulas, owing to the first-step analysis:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\begin{cases} P_{1\to2}^{(k+1)} = P_{11}P_{1\to2}^{(k)} + P_{12}P_{2\to2}^{(k)} \\ P_{2\to2}^{(k+1)} = P_{21}P_{1\to2}^{(k)} + P_{22}P_{2\to2}^{(k)} \end{cases} \quad \begin{cases} P_{1\to1}^{(k+1)} = P_{11}P_{1\to1}^{(k)} + P_{12}P_{2\to1}^{(k)} \\ P_{2\to1}^{(k+1)} = P_{21}P_{1\to1}^{(k)} + P_{22}P_{2\to1}^{(k)} \end{cases}$$

Summing over k on both sides of the equations yields

$$\begin{aligned} \sum_{k=1}^{\infty} P_{1\to2}^{(k)} P_{23} &= \frac{P_{12}P_{23}}{(1-P_{11})(1-P_{22})-P_{12}P_{21}} \\ P_{13} + \sum_{k=1}^{\infty} P_{1\to1}^{(k)} P_{13} &= \frac{(1-P_{22})P_{13}}{(1-P_{11})(1-P_{22})-P_{12}P_{21}} \end{aligned} \quad (2)$$

From these formulas and the posterior distributions of P_{ij}, we find that 91.2% of the transitions from state 1 to state 3 occurs finally through the intermediate state 2 for the *Ffh-Data*.

These observations and calculations reveal that (i) the movement of the FtsY-[Ffh-NG] complex from the RNA capped end to the distal end requires the middle state, which serves as an on-pathway intermediate to facilitate this largescale movement. (ii) The middle state is quite efficient in facilitating the search for the RNA distal site: once the SRP-SR complex reaches this state, over 50% of molecules move on successfully to the distal site (high-FRET state) (because $P_{23} > P_{21}$); this over 50% probability is much higher than that from the low-FRET state.

5.2 Effect of RNC

Once RNC is added to the SRP-SR complex, the experimental FRET trajectories, the *RNC-data*, show the presence of only *one* state with a low FRET value: the FRET values are well fitted by $y_i = const$ + Gaussian noise, see Table 5. Comparison of these results with those on SRP-SR alone (the *Ffh-Data* and *FtsY-Data*) show that the RNC has a pausing effect: it holds the SRP-SR complex near the capped end and prevents its movement to the RNA distal end (see C of Figure 4). This pausing effectively prevents premature dissociation of SRP and SR, which happens at the distal end of the SRP RNA and results in abortive reactions. We thus see that RNC plays an important regulating role in ensuring the efficiency of a successful protein targeting.

5.3 Role of Translocon

When the translocon is further added to the RNC-SRP-SR complex, single-molecule experiments on the translocon-RNC-SRP-SR complex yield the *Translocon-Data* in Table 1. As shown in Table 5, the high-FRET state ($\mu_{0,3} \approx 0.6$) is restored in the *Translocon-Data*, which is completely absent in the *RNC-Data*. Therefore, the translocon enables the FtsY-[Ffh-NG] complex to restore movement to the RNA distal end, where disassembly of SRP-SR (by GTP-hydrolysis) can be initiated.

We also observe that the transition probabilities of the *Translocon-Data*, shown in Table 6, differ significantly from those of the *Ffh-Data*. This rules out the model that the translocon simply awaits for and binds the RNC that has spontaneously dissociated from the SRP-SR

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

complex. If this were the case, the FRET trajectories in the presence of both RNC and translocon (the *Translocon-Data*) would exhibit nearly identical features as those for the SRP-SR complex (the *Ffh-Data*). Instead, these data strongly suggest that the translocon forms a quarternary complex together with RNC, SRP and SR, in which attainment of the distal conformation is favored.

We next consider the role of the middle state. Using formula (2) derived in Section 5.1, we find that *only* 40.7% of the transitions from the low FRET to high FRET state occur via the middle state as an intermediate for the *Translocon-Data*. This is in sharp contrast with the 91.2% probability for the *Ffh-Data*. This indicates that the translocon alters the pathway via which the FtsY-[Ffh-NG] complex searches for the RNA distal site, biasing them towards pathways in which transitions between low FRET and high FRET states occur directly. We note that it is possible that in the presence of translocon, the residence in the intermediate state could be too fast to be detected within the time resolution (30 ms) of the experiment.

To gain further insights into the regulatory role of the translocon, we asked whether and how it alters the kinetics by which the SRP-SR complex undergoes the structural change. To this end, we compare the dwell time of the FtsY-[Ffh-NG] complex at the high-FRET state, which is $d_3 = 1/(1 - P_{33})$, between the *Translocon-Data* and the *Ffh-Data*. The 95% posterior interval for d_3 is [2.058, 2.577] seconds for the *Translocon-Data* and [0.465, 0.507] seconds for the *Ffh-Data*, respectively. Thus, the translocon enhances the kinetic stability of the SRP-SR complex in the distal conformation by 4-5 fold. Table 7 contrasts the parameter estimates between the *Ffh-Data* and the *Translocon-Data*.

In summary, our statistical analysis shows that the translocon regulates the protein targeting process by (i) restoring the movements of the FtsY-[Ffh-NG] complex to the RNA distal end, (ii) promoting alternative pathways for this movement, in which the FtsY-[Ffh-NG] complex directly transitions from the low-FRET state to the high-FRET state, and (iii) prolonging the time that FtsY-[Ffh-NG] stays at the RNA distal end. It is known that movement of the FtsY-[Ffh-NG] complex away from the RNA capped end is important for vacating the ribosome binding site and initiating ribosome-translocon contacts during the handover of RNC to the translocon. It is also known that GTP-hydrolysis, which disassembles SRP and SR, occurs at the RNA distal end (Shen et al., 2013). Our findings thus reveal that the translocon, via mechanisms (i)-(iii), promotes both of these molecular events and allows them to be synchronized in the pathway. Collectively, these results show that the translocon not only serves as a channel through which the nascent proteins translocate, but also facilitates the productive handover of the RNC onto itself to complete the protein targeting reaction.

5.4 A proposal of detailed mechanism

Our statistical analysis of the single-molecule experimental data in combination with the known biological understanding (Halic et al., 2006; Pool et al., 2002; Peluso et al., 2001; Estrozi et al., 2011; Shen and Shan, 2010; Zhang et al., 2009a; Akopian et al., 2013a; Ataide et al., 2011) suggests the following detailed mechanism of protein targeting, which was conjectured in Shen et al. (2012), corresponding to the four steps of Section 1:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1. SRP recognizes the signal sequence on RNC and binds it. The RNC is delivered to the target membrane where the SR can localize to.
2. When the SRP-SR complex is initially formed, the FtsY-[Ffh-NG] complex binds at the RNA capped end near the ribosome exit site, blocking the site from translocon binding.
3. As the RNC initiates contact with the translocon, the latter actively facilitates the conformation change of SRP-SR complex and drives the FtsY-[Ffh-NG] complex from the capped end to the distal end of RNA.
4. GTP-hydrolysis is initiated at the RNA distal end to disassemble the SRP and SR. Meanwhile, the nascent chain is released from the Ffh M-domain to the translocon on the membrane.

Figure 10 illustrates the detailed mechanism. The movement of the FtsY-[Ffh-NG] complex from the RNA capped end to the distal end is first negatively regulated by RNC, whose pausing effect keeps the SRP-SR complex from disassembly before the translocon is identified, and later positively regulated by the translocon, which actively facilitates the movement of FtsY-[Ffh-NG] to the RNA distal end. This mechanism allows the coordinated exchange of SRP and translocon at the RNC and the effective timing of GTP-hydrolysis, thus minimizing abortive reactions due to premature SRP-SR disassembly or non-productive loss of the RNC.

6 Model Checking

6.1 Check of detailed balance

In biophysics, the principle of microscopic reversibility states that at equilibrium the transition flux between any two states should be equal. In the familiar probability language, the microscopic reversibility translates into the detailed balance condition or the reversibility of the Markov chain: $\pi_i P_{ij} = \pi_j P_{ji}$ for all i and j, where π_i is the equilibrium probability of state i. This can be checked from the posterior samples of the transition matrix $\boldsymbol{P}$.

Figure 11 compares the distribution of $\pi_i P_{ij}$ (first column) with that of $\pi_j P_{ji}$ (second column) from the *Ffh-Data*. The third column shows the distribution of the difference $\pi_i P_{ij} - \pi_j P_{ji}$ compared to zero (the vertical bar), where $i, j \in \{1, 2, 3\}$, $i \neq j$. It is clear that $\pi_i P_{ij} - \pi_j P_{ji} = 0$ holds within the experimental error. The plots on the *FtsY-Data* and the *Translocon-Data* give very similar pictures. We thus confirm that indeed under our hierarchical HMM the principle of microscopic reversibility is satisfied.

6.2 Check of Markovian assumption

In our hierarchical HMM, the Markov assumption of the state transitions (or the conformation changes) plays a fundamental role. If the Markov assumption is correct, then the waiting time at the individual state should be exponentially distributed and that the successive waiting times should be independent of each other. Both can be checked under our Bayesian sampling approach, since we can straightforwardly obtain the waiting time at each state from the posterior samples of the hidden states $\boldsymbol{z}$. Figure 12 shows the posterior

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

distribution of the waiting time at each of the low, middle and high FRET state of the *Ffh-Data* based on the samples of hidden states $\boldsymbol{z}$ in its original scale (left column) and the log-scale (right column). It is seen that on the log-scale the distribution of the waiting time is well fit by a straight line, supporting the exponential distribution. Quantitatively, we performed a chi-squared goodness-of-fit test for the exponential distribution using 30 evenly spaced bins. The resulting p-values for the waiting time at the low, middle and high FRET states are 0.72, 0.20 and 0.35, respectively. Figure 13 shows the autocorrelation of the successive waiting times from the *Ffh-Data* obtained from the samples of the hidden states $\boldsymbol{z}$. It is evident that the successive waiting times are uncorrelated, as the Markov assumption requires. The posterior samples from the *FtsY-Data* and the *Translocon-Data* show quite similar pattern.

7 Summary

The advances in single-molecule experiments enable us to study the detailed mechanism of the co-translational protein targeting process. On the single-molecule level the data are necessarily stochastic. They are often noisy realizations of the underlying stochastic dynamics. To model the stochasticity of each individual experimental trajectory, we use HMM.

The experimental time windows in single-molecule trajectories are often of rather limited length, resulting in relatively short trajectories. As a result, the parameter estimation based on individual trajectories could be quite variable. Furthermore, the determination of the total number of states of the HMM based on individual trajectories is highly unstable. Experimentally, these issues are mitigated by recording hundreds of trajectories repeated under the same experimental condition. In this article, we use the mode of the BIC selection over multiple trajectories for reliable determination of the number of states of the HMM as a preliminary analysis. Then we propose a hierarchical HMM to pool information together from the different trajectories and at the same time to account for the heterogeneity among them. The heterogeneity among the different trajectories arises from the intrinsically stochastic nature of molecular actions, equipment noise, thermal fluctuation and random variations in experimental setups. We find that the proposed hierarchical HMM is highly robust to low signal-to-noise ratios. Finally, assessment of the fitting of each individual trajectory based on parameters estimated from the hierarchical model re-assured us of the model selection at the first stage and the assumption of the hierarchical model at the second stage.

Biologically, we corroborated many conclusions from the previous ad-hoc analysis, giving solid quantitative evidence for the proposed new mechanism of co-translational protein targeting. Instead of being passively involved in the protein targeting process, our analysis shows that the RNC and translocon play active regulatory roles to facilitate the accurate timing of the biological steps. Specifically, the RNC and translocon effectively regulate the movement of the SRP-SR complex between the capped end and the distal end of the RNA, which in turn regulates the assembly and disassembly of the SRP-SR complex and the preference of the RNC for binding the SRP-SR complex versus the translocon. Compared to the previous ad-hoc analysis, our statistical analysis clarifies the pathway for the structural

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

change in the SRP-SR complex, and rigorously showed that the translocon alters the pathway, kinetics, and stability of this structural change, providing stronger evidence that the translocon actively facilitates the loading of RNC onto itself and drives the completion of protein targeting. From a modeling perspective, the hierarchical HMMs that we used for combining information are quite general. They appear effective for dealing with replicated experiments and can be potentially used for analyzing other biological or biochemical experiments. We thus hope that this article would generate further interest in studying these hierarchical models and in applying them for general data analysis.

Acknowledgments

S. Shan's research is supported in part by NIH grant GM078024 and the Gordon and Betty Moore Foundation through Grant GBMF2939.

References

Akaike H. A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control. 1974; 19:716–723.

Akopian D, Dalal K, Shen K, Duong F, Shan S. SecYEG activates GTPases to drive the completion of cotranslational protein targeting. The Journal of Cell Biology. 2013a; 200:397–405. [PubMed: 23401005]

Akopian D, Shen K, Zhang X, Shan S. Signal Recognition Particle: an Essential Protein-targeting Machine. Annual Review of Biochemistry. 2013b; 82:693–721.

Ataide SF, Schmitz N, Shen K, Ke A, Shan S, Doudna JA, Ban N. The Crystal Structure of the Signal Recognition Particle in Complex with its Receptor. Science. 2011; 331:881–886. [PubMed: 21330537]

Baum LE, Petrie T. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. The Annals of Mathematical Statistics. 1966; 37:1554–1563.

Baum LE, Petrie T, Soules G, Weiss N. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. The Annals of Mathematical Statistics. 1970; 41:164–171.

Biernacki, C., Celeux, G., Govaert, G. Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood. 1998. Archived article, available at https://hal.inria.fr/inria-00073163/

Blanco M, Walter NG. Analysis of Complex Single-Molecule FRET Time Trajectories. Methods in Enzymology. 2010; 472:153178.

Bronson JE, Fei J, Hofman JM, Ruben LGonzalez J, Wiggins CH. Learning Rates and States from Biophysical Time Series: a Bayesian Approach to Model Selection and Single-molecule FRET Data. Biophysical Journal. 2009; 97:3196–3205. [PubMed: 20006957]

Bulla J, Bulla I, Nenadic O. hsmm - An R Package for Analyzing Hidden semi-Markov Models. Computational statistics and data analysis. 2010; 54:611–619.

Cappe, O., Moulines, E., Ryden, T. Inference in Hidden Markov Models. 2005. Springer Series in Statistics

Celeux G, Durand J-B. Selecting Hidden Markov Model State Number with Cross-Validated Likelihood. Computational Statistics. 2008; 23:541–564.

Chen J, Li P. Hypothesis Test for Normal Mixture Models: the EM Approach. The Annals of Statistics. 2009; 37:25232542.

Dahan M, Deniz AA, Ha T, Chemla DS, Schultz PG, Weiss S. Ratiometric Measurement and Identification of Single Diffusing Molecules. Chemical Physics. 1999; 247:85–106.

Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society B. 1977; 39:1–38.

Eddy SR. Hidden Markov Models. Current Opinion in Structural Biology. 1996; 6:361–365. [PubMed: 8804822]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Estrozi LF, Boehringer D, Shan S, Ban N, Schaffitzel C. Cryo-EM structure of the E. coli translating ribosome in complex with SRP and its receptor. Nature Structural & Molecular Biology. 2011; 18:88–90.

Finesso, L. PhD dissertation. University of Maryland; 1990. Consistent Estimation of the Order for Markov and Hidden Markov Models.

Frühwirth-Schnatter, S. Finite Mixture and Markov Switching Models. Springer-Verlag New York, Inc; 2006.

Geman S, Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984; 6:721–741. [PubMed: 22499653]

Greenfeld M, Pavlichin DS, Mabuchi H, Herschlag D. Single Molecule Analysis Research Tool (SMART): an Integrated Approach for Analyzing Single Molecule Data. PLos One. 2012:7.

Halic M, Becker T, Pool MR, Spahn CMT, Grassucci RA, Frank J, Beckmann R. Structure of the Signal Recognition Particle Interacting with the Elongation-arrested Ribosome. Nature. 2004; 427:808–814. [PubMed: 14985753]

Halic M, Gartmann M, Schlenker O, Mielke T, Pool MR, Sinning I, Beckmann R. Signal Recognition Particle Receptor Exposes the Ribosomal Translocon Binding Site. Science. 2006; 312:745–747. [PubMed: 16675701]

Hawkins DS, Allen DM, Stromberg AJ. Determining the Number of Components in Mixtures of Linear Models. Computational Statistics & Data Analysis. 2001; 38:15–48.

Keenan RJ, Freymann DM, Stroud RM, Walter P. The Signal Recognition Particle. Annual Review of Biochemistry. 2001; 70:755–775.

Keller BG, Kobitski A, Jaschke A, Nienhaus G, Noe F. Complex RNA Folding Kinetics Revealed by Single-Molecule FRET and Hidden Markov Models. Journal of the American Chemical Society. 2014; 136:45344543.

Keribin C. Consistent Estimation of the Order of Mixture Models. The Indian Journal of Statistics, Series A. 2000; 62:49–66.

König SL, Hadzic M, Fiorini E, Börner R, Kowerko D, Blanckenhorn WU, Sigel RKO. BOBA FRET: Bootstrap-based Analysis of Single-molecule FRET Data. PLos One. 2013:8.

Leroux BG. Consistent Estimation of a Mixing Distribution. The Annals of Statistics. 1992; 20:13501360.

Liu, JS. Monte Carlo Strategies in Scientific Computing. Springer-Verlag New York, Inc; 2001.

Liu Y, Park J, Dahmen KA, Chemla YR, Ha T. A Comparative Study of Multivariate and Univariate Hidden Markov Modelings in Time-binned Single-molecule FRET Data Analysis. Journal of Physical Chemistry. 2010; 114:5386–5403. [PubMed: 20361785]

Lodish, H., Berk, A., Zipursky, SL., Matsudaira, P., Baltimore, D., Darnell, J. Molecular Cell Biology. 4. New York: W. H. Freeman; 2000.

MacKAY RJ. Estimating the Order of a Hidden Markov Model. Canadian Journal of Statistics. 2002; 30:573–589.

McKinney SA, Joo C, Ha T. Analysis of Single-molecule FRET Trajectories Using Hidden Markov Modeling. Biophysical Journal. 2006; 91:1941–1951. [PubMed: 16766620]

McLachlan, G., Peel, D. Finite Mixture Models. 2005. Wiley Series in Probability and Statistics

Moerner WE. A Dozen Years of Single-molecule Spectroscopy in Physics, Chemistry, and Biophysics. The Journal of Physical Chemistry B. 2002; 106:910–927.

Nie S, Zare RN. Optical Detection of Single Molecules. Annual Review of Biophysics and Biomolecular Structure. 1997; 26:567–596.

Nyathi Y, Wilkinson BM, Pool MR. Co-translational Targeting and Translocation of Proteins to the Endoplasmic Reticulum. Biochimica et Biophysica Acta (BBA) -Molecular Cell Research. 2013; 1833:2392–2402. [PubMed: 23481039]

Peluso P, Shan S, Nock S, Herschlag D, Walter P. Role of SRP RNA in the GTPase Cycles of Ffh and FtsY. Biochemistry. 2001; 40:15224–15233. [PubMed: 11735405]

Pool MR, Stumm J, Fulga TA, Sinning I, Dobberstein B. Distinct Modes of Signal Recognition Particle Interaction with the Ribosome. Science. 2002; 297:1345–1348. [PubMed: 12193787]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Qian H, Kou SC. Statistics and Related Topics in Single-Molecule Biophysics. Annual Review of Statistics and Its Application. 2014; 1:465–492.

Rabiner LR. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE. 1989; 77:257–286.

Rapoport TA. Protein Transport across the Endoplasmic Reticulum Membrane: Facts, Models, Mysteries. The FASEB Journal. 1991; 5:2792–2798. [PubMed: 1916103]

Rapoport TA. Protein Translocation across the Eukaryotic Endoplasmic Reticulum and Bacterial Plasma Membranes. Nature. 2007; 450:663–669. [PubMed: 18046402]

Roy R, Hohng S, Ha T. A Practical Guide to Single-molecule FRET. Nature Methods. 2008; 5:507–516. [PubMed: 18511918]

Ryden T. Estimating the Order of Hidden Markov Models. Statistics. 1995; 26:345354.

Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics. 1978; 6:461–464.

Shen K, Arslan S, Akopian D, Ha T, Shan S. Activated GTPase Movement on an RNA Sca3old Drives Co-translational Protein Targeting. Nature. 2012; 492:271–275. [PubMed: 23235881]

Shen K, Shan S. Transient Tether between the SRP RNA and SRP Receptor Ensures Efficient Cargo Delivery during Cotranslational Protein Targeting. Proceedings of the National Academy of Sciences. 2010; 107:7698–7703.

Shen K, Wang Y, Hwang Fu Y-H, Zhang Q, Feigon J, Shan S. Molecular Mechanism of GTPase Activation at the Signal Recognition Particle (SRP) RNA Distal End. The Journal of Biological Chemistry. 2013; 288:36385–36397. [PubMed: 24151069]

Tamarat P, Maali A, Lounis B, Orrit M. Ten Years of Single-Molecule Spectroscopy. The Journal of Physical Chemistry A. 2000; 104:1–16.

Tanner MA, Wong WH. The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association. 1987; 82:528–540.

van de Meent JW, Bronson JE, Wiggins CH, Gonzalez RL Jr. Empirical Bayes Methods Enable Advanced Population-level Analyses of Single-molecule FRET Experiments. Biophysical Journal. 2014; 106:1327–1337. [PubMed: 24655508]

Voigts-Hoffmann F, Schmitz N, Shen K, Shan S, Ataide SF, Ban N. The Structural Basis of FtsY Recruitment and GTPase Activation by SRP RNA. Molecular Cell. 2013; 52:643–654. [PubMed: 24211265]

Watkins LP, Yang H. Detection of Intensity Change Points in Time-Resolved Single-Molecule Measurements. The Journal of Physical Chemistry B. 2005; 109:617628.

Weiss S. Measuring Conformational Dynamics of Biomolecules by Single Molecule Fluorescence Spectroscopy. Nature Structural Biology. 2000; 7:724–729. [PubMed: 10966638]

Windham MP, Cutler A. Information Ratios for Validating Mixture Analysis. Journal of the American Statistical Association. 1992; 87:1188–1192.

Xie XS, Lu HP. Single-Molecule Enzymology. The Journal of Biological Chemistry. 1999; 274:15967–15970. [PubMed: 10347141]

Xie XS, Trautman JK. Optical Studies of Single Molecules at Room Temperature. Annual Review of Physical Chemistry. 1998; 49:441–480.

Zhang X, Jantama K, Moore JC, Jarboe LR, Shanmugam KT, Ingrama LO. Metabolic evolution of energy-conserving pathways for succinate production in Es-cherichia coli. Proceedings of the National Academy of Sciences. 2009a; 106:20180–20185.

Zhang X, Kung S, Shan S. Demonstration of a Multi-step Mechanism for Assembly of the SRP-SR Receptor Complex: Implications for the Catalytic Role of SRP RNA. Journal of Molecular Biology. 2008; 381(3):581–593. [PubMed: 18617187]

Zhang X, Schaffitzel C, Ban N, Shan S. Multiple Conformational Switches in a GTPase Complex Control Co-translational Protein Targeting. Proceedings of the National Academy of Sciences. 2009b; 106:1754–1759.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

A Baum-Welch/EM algorithm for HMM

For a given value of K, the total number of states, we can use the EM algorithm (Dempster et al., 1977), a.k.a. the Baum-Welch algorithm for HMM (Baum and Petrie, 1966; Baum et al., 1970), to infer $\boldsymbol{\theta}$. For the ease of presentation, we assume here that the initial distribution of the first hidden state z_1 is flat. The full likelihood function is

$$L(\boldsymbol{\theta})=\prod_{n=2}^{N} p(z_n|z_{n-1},\boldsymbol{P})\prod_{n=1}^{N} p(y_n|z_n,\boldsymbol{\mu},\boldsymbol{\sigma^2})=\prod_{j,k=1}^{K} P_{jk}^{T_{jk}}\cdot\prod_{n=1}^{N}\mathcal{N}(y_n;\mu_{z_n},\sigma_{z_n}^2),$$

where T_{jk} denotes the total number of transitions in $\boldsymbol{z}$ from state j to state k, and $\mathcal{N}(y;\mu,\sigma^2)$ denotes the normal density with mean μ and variance σ^2 evaluated at y. For the EM algorithm, in the E-step, the expectation step, we have

$$E\log L(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})=\sum_{j,k=1}^{K}\sum_{n=2}^{N} v_{n,j,k}\log P_{jk}+\sum_{k=1}^{K}\sum_{n=1}^{N} u_{n,k}\log\mathcal{N}(y_n;\mu_k,\sigma_k^2),$$

where $u_{n,k}=p(z_n=k|\boldsymbol{y},\boldsymbol{\theta}^{old})$ and $v_{n,j,k}=p(z_{n-1}=j,z_n=k|\boldsymbol{y},\boldsymbol{\theta}^{old})$ can be expressed in terms of $\alpha(z_n):=p(\boldsymbol{y}_{1:n},z_n|\boldsymbol{\theta}^{old})$ and $\beta(z_n):=p(\boldsymbol{y}_{(n+1):N}|z_n,\boldsymbol{\theta}^{old})$:

$$\begin{aligned} v_{n,z_n} &= \alpha(z_n)\beta(z_n)/p(\boldsymbol{y}_{1:N}|\boldsymbol{\theta}^{old}),\\ v_{n,z_{n-1},z_n} &= \alpha(z_{n-1})\beta(z_n)p(y_n|z_n,\boldsymbol{\theta}^{old})p(z_n|z_{n-1},\boldsymbol{\theta}^{old})/p(\boldsymbol{y}_{1:N}|\boldsymbol{\theta}^{old}).\end{aligned}$$

$\alpha(z_n)$ and $\beta(z_n)$ can be efficiently calculated by the forward-backward algorithm (Rabiner, 1989), a recursive formula that allows fast computation: evaluating the α's forwardly from 1 to N and the β's backwardly from N to 1:

$$\alpha(z_n)=p(y_n|z_n,\boldsymbol{\theta}^{old})\sum_{z_{n-1}=1}^{K}\alpha(z_{n-1})p(z_n|z_{n-1},\boldsymbol{\theta}^{old}), \tag{3}$$

$$\beta(z_n)=\sum_{z_{n+1}=1}^{K}\beta(z_{n+1})p(y_{n+1}|z_{n+1},\boldsymbol{\theta}^{old})p(z_{n+1}|z_n,\boldsymbol{\theta}^{old}),\quad \beta(z_N)\equiv 1. \tag{4}$$

In addition, the forward-backward algorithm gives the marginal likelihood evaluated at the maximum likelihood estimate $p(\boldsymbol{y}|\hat{\boldsymbol{\theta}})=\Sigma_{z_N}\alpha(z_N)=\Sigma_{z_N}p(\boldsymbol{y}_{1:N},z_N|\hat{\boldsymbol{\theta}})$.

In the M-step of the EM algorithm, which maximizes $E\log L(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})$ over $\boldsymbol{\theta}$, we obtain $\boldsymbol{\theta}^{new}$ according to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$P_{jk}=\frac{\sum_{n=2}^{N} v_{n,j,k}}{\sum_{k=1}^{K}\sum_{n=2}^{N} v_{n,j,k}},\ \mu_k=\frac{\sum_{n=1}^{N} y_n u_{n,k}}{\sum_{n=1}^{N} u_{n,k}},\ \sigma_k^2=\frac{\sum_{n=1}^{N} u_{n,k}(y_n-\mu_k)^2}{\sum_{n=1}^{N} u_{n,k}}.$$

B Gibbs Sampling for HMM

In addition to the EM algorithm, which quickly obtains the MLE of the parameters, we can also use Bayesian MCMC sampling (Liu, 2001) to assess the entire (posterior) distribution of the parameters. Our MCMC sampling can be viewed as a special case of data augmentation (Tanner and Wong, 1987): augment the parameter space $\boldsymbol{\theta}$ with the hidden states $\boldsymbol{z}$, and iteratively sample one given the other (i.e., sample $\boldsymbol{\theta}$ given $\boldsymbol{z}$ and sample $\boldsymbol{z}$ given $\boldsymbol{\theta}$).

Specifically, in our MCMC sampling, we adopt flat priors for $\boldsymbol{P}$ and μ_k, $k = 1, \ldots, K$, and independent inverse-χ^2 priors with parameters ν, s^2 for σ_k^2 (the prior on $\boldsymbol{\mu}$ is flat over the region $0 < \mu_1 < \cdots < \mu_K < 1$). The posterior distribution is

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}) &= p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}) p_0(\boldsymbol{P}) p_0(\boldsymbol{\mu}) p_0(\boldsymbol{\sigma^2}) \\ &\propto \prod_{j=1}^{K}\prod_{k=1}^{K} P_{jk}^{T_{jk}} \prod_{n=1}^{N} \mathcal{N}(y_n;\mu_{z_n},\sigma_{z_n}^2) \prod_{k=1}^{K} p_0(\sigma_k^2;\nu,s^2). \end{aligned}$$

It follows that in our (group Gibbs) sampler, the conditional distribution of the jth row of the transition matrix $P_{j\cdot} = (P_{j1}, P_{j2}, \ldots, P_{jK})$ is a Dirichlet distribution, the conditional distribution of $\boldsymbol{\mu}$ is a multivariate normal distribution, the conditional distribution of $\boldsymbol{\sigma^2}$ is a multivariate inverse-χ^2 distribution and that the hidden states $\boldsymbol{z}$ can be sampled sequentially from 1 to N through the following recursion:

$$\begin{aligned} p(z_n=k|z_{n-1}=j,\boldsymbol{\theta},\boldsymbol{y}) &\propto P_{jk}\,\mathcal{N}(y_n;\mu_k,\sigma_k)\,p(\boldsymbol{y}_{n+1:N}|z_n=k) \\ &= P_{jk}\,\mathcal{N}(y_n;\mu_k,\sigma)\,\beta(k), \quad n=1,2,\ldots,N, \end{aligned}$$

where $\beta(k)$ is the the backward probability defined in equation (4).

C MCMC sampling of the hierarchical HMM

The posterior distribution is proportional to

$$p(\boldsymbol{\mu}_0,\boldsymbol{\eta}_0^2,\boldsymbol{s}^2)\prod_l p(\boldsymbol{y}^{(l)},\boldsymbol{z}^{(l)}|I^{(l)},\boldsymbol{\mu}^{(l)},\boldsymbol{\sigma}^{(l)},\boldsymbol{P})\times\prod_l p(\boldsymbol{\mu}^{(l)}|\boldsymbol{\mu}_0,\boldsymbol{\eta}_0^2,I^{(l)})p((\boldsymbol{\sigma}^{(l)})^2|\boldsymbol{\nu},\boldsymbol{s}^2,I^{(l)})p(I^{(l)}).$$

We use the Gibbs sampler to update a group of parameters at a time, conditioning on the others, and iterate until convergence. The sampling details are given below, where $I(\omega)$ and I_ω denote the indicator function.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1. Initialization. Fit each trajectory independently using the EM algorithm in Appendix A and set the initial values of $\{\boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}\}$ at the corresponding MLEs. The initial values of $\{I^{(l)}\}$ are set to be $\{1, \ldots, K\}$.

2. Update global parameters $\boldsymbol{\mu}_0, \boldsymbol{\eta}_0^2, \boldsymbol{s}^2$. For $1 \le k \le K$,

$$\begin{aligned}
&\text{Sample } \mu_{0,k} \text{ from } \mathcal{N}\Big(\textstyle\sum_{l=1,k\in I^{(l)}}^{T} \mu_k^{(l)} / (\sum_{l=1}^{T} I_{k\in I^{(l)}}), \eta_{0,k}^2 / (\sum_{l=1}^{T} I_{k\in I^{(l)}})\Big),\\
&\text{Sample } \eta_{0,k}^2 \text{ from Inv-}\chi^2\Big(\textstyle\sum_{l=1}^{T} I_{k\in I^{(l)}} - 2, \sum_{l=1,k\in I^{(l)}}^{T} (\mu_k - \mu_{0,k})^2 / (\sum_{l=1}^{T} I_{k\in I^{(l)}} - 2)\Big),\\
&\text{Sample } s_k^2 \text{ from } \Big\{\nu_k \textstyle\sum_{l=1}^{T} I_{k\in I^{(l)}} / (\sigma_k^{(l)})^2\Big\}^{-1} \chi_{df}^2,\ df = \nu_k \textstyle\sum_{l=1}^{T} I_{k\in I^{(l)}} + 2.
\end{aligned}$$

3. Update transition probabilities $\boldsymbol{P}$ according to

$$p(\boldsymbol{P}) \propto \prod_{i,j} P_{ij}^{\sum_l N_{i,j}^{(l)}} \Big/ \prod_{I^{(l)} \neq \{1,2,\ldots,K\}} \prod_{i \in I^{(l)}} \Big(\sum_{k \in I^{(l)}} P_{ik}\Big)^{\sum_{k\in I^{(l)}} N_{i,k}^{(l)}}.$$

4. Update parameters for individual trajectories.

   - Update $\{\boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}\}$. For $k \in I^{(l)}$, $l = 1, \ldots, T$,

$$\begin{aligned}
\mu_k^{(l)} &\sim \mathcal{N}\left(\frac{\mu_{0k}/\eta_{0k}^2 + \sum_{z_n^{(l)}=k} y_n^{(l)} / (\sigma_k^{(l)})^2}{1/\eta_{0k}^2 + \sum_{z_n^{(l)}=k} 1/(\sigma_k^{(l)})^2}, \frac{1}{1/\eta_{0k}^2 + \sum_{z_n^{(l)}=k} 1/(\sigma_k^{(l)})^2}\right);\\
(\sigma_k^{(l)})^2 &\sim INV\text{-}\chi^2\left(\nu_k + \sum_{n=1}^{N_l} I(z_n^{(l)}=k), \frac{\nu_k s_k^2 + \sum_n (y_n^{(l)} - \mu_k^{(l)})^2 I(z_n^{(l)}=k)}{\nu_k + \sum_n I(z_n^{(l)}=k)}\right).
\end{aligned}$$

   - Update $\{z^{(l)}\}$. This is essentially the same as introduced in Appendix B except that when $I^{(l)} \neq \{1, 2, \ldots, K\}$, the transition matrix is a re-normalized submatrix of $\boldsymbol{P}$ according to which states are present in trajectory l.

   - Update $\{I^{(l)}\}$. $I^{(l)}$ is equal to $A \subset \{1, 2, \ldots, K\}$ with probability proportional to

$$p(\boldsymbol{y}^{(l)}, \boldsymbol{z}^{(l)} | \boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}, \boldsymbol{P}, I^{(l)}=A) p(\boldsymbol{\mu}^{(l)} | \boldsymbol{\mu}_0, \boldsymbol{\eta}_0^2, I^{(l)}=A) p((\boldsymbol{\sigma}^{(l)})^2 | \boldsymbol{\nu}, \boldsymbol{s}^2, I^{(l)}=A)$$

     where A stands for $\{1, 2, 3\}$, $\{1, 2\}$, $\{1, 3\}$ or $\{2, 3\}$ when $K = 3$, and $\{1, 2, 3, 4\}$, $\{1, 2, 3\}$, $\{1, 2, 4\}$, $\{1, 3, 4\}$, $\{2, 3, 4\}$, $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, $\{2, 4\}$, or $\{3, 4\}$ when $K = 4$.

5. Iterate Steps 2 to 4 until convergence.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

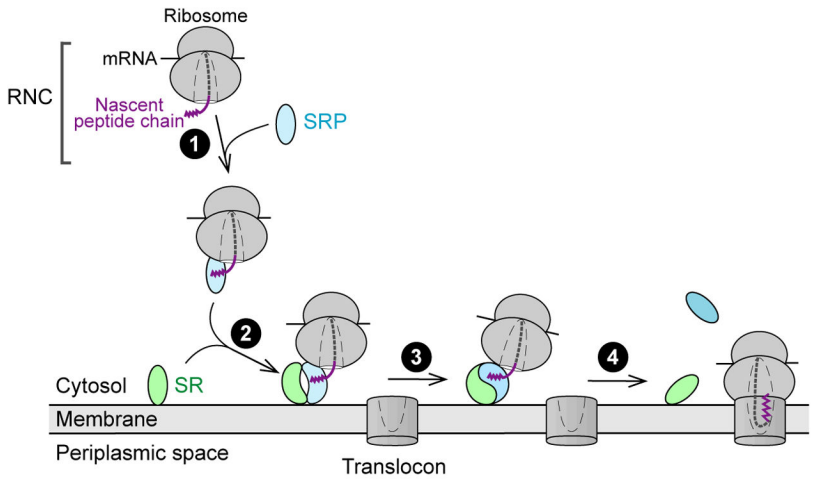

Figure 1.
The four steps of protein targeting.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

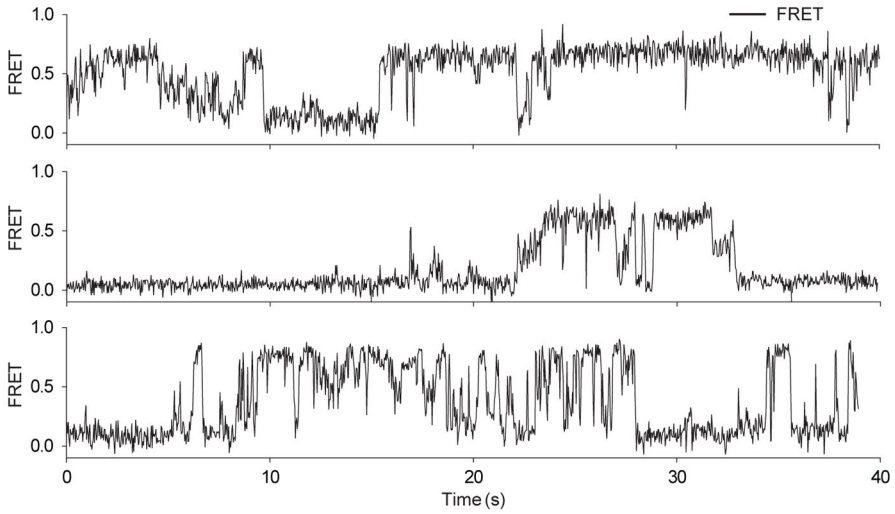

Figure 2.
Three sample FRET trajectories.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

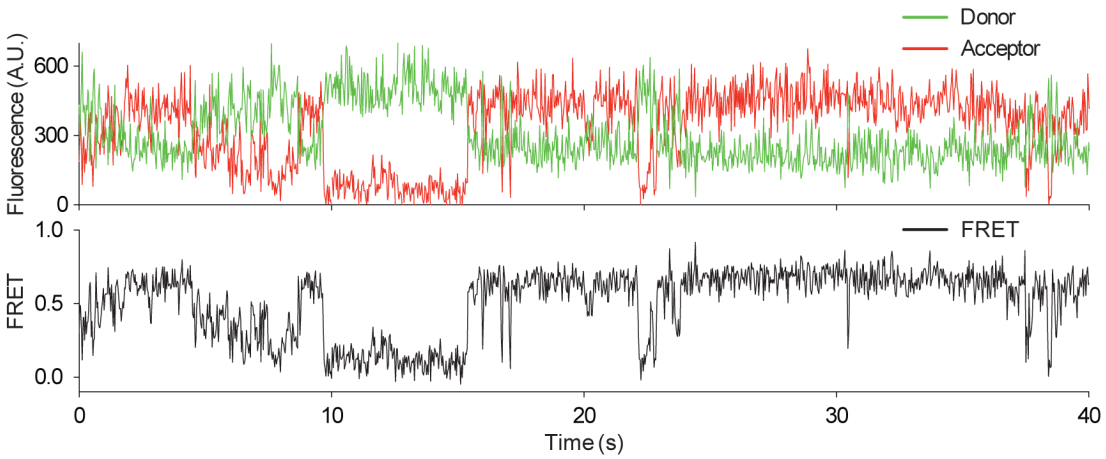

Figure 3.

Sample trajectory of FRET observations. The upper panel is the fluorescence of the donor and the acceptor, respectively; the lower panel shows the FRET values.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

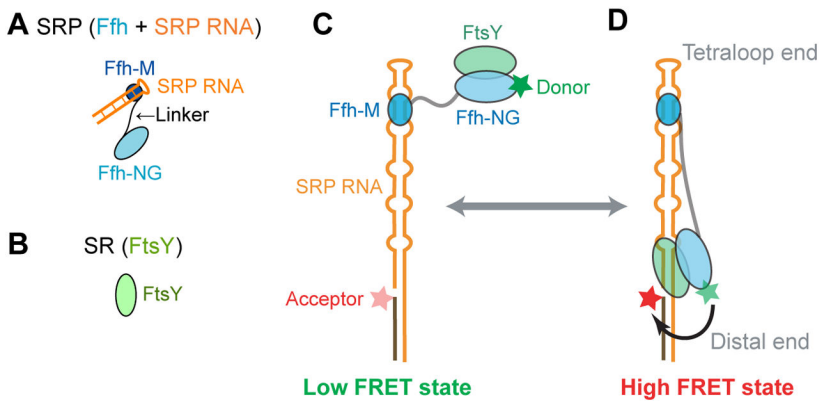

Figure 4.

Molecular details of SRP and SR in *E.coli*. (A) SRP in *E.coli* is composed of RNA, Ffh-M and Ffh-NG. Ffh-M binds the RNA and the signal sequence (not shown); Ffh-NG binds the ribosome (not shown) and SR. (B) SR in *E.coli* is the FtsY protein. (C) FtsY-[Ffh-NG] complex is near the capped end of the RNA with a low FRET value. (D) FtsY-[Ffh-NG] complex is near the distal end of the RNA with a high FRET value. The red and green stars denote the FRET acceptor and donor, respectively.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

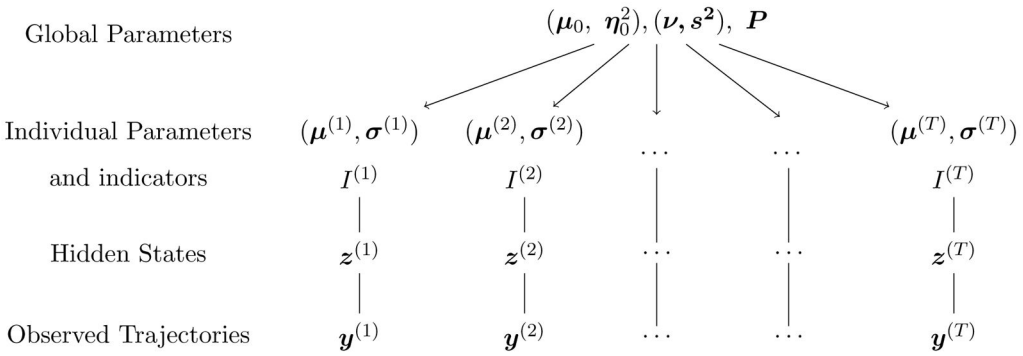

Figure 5.
Diagram of the hierarchical HMM.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

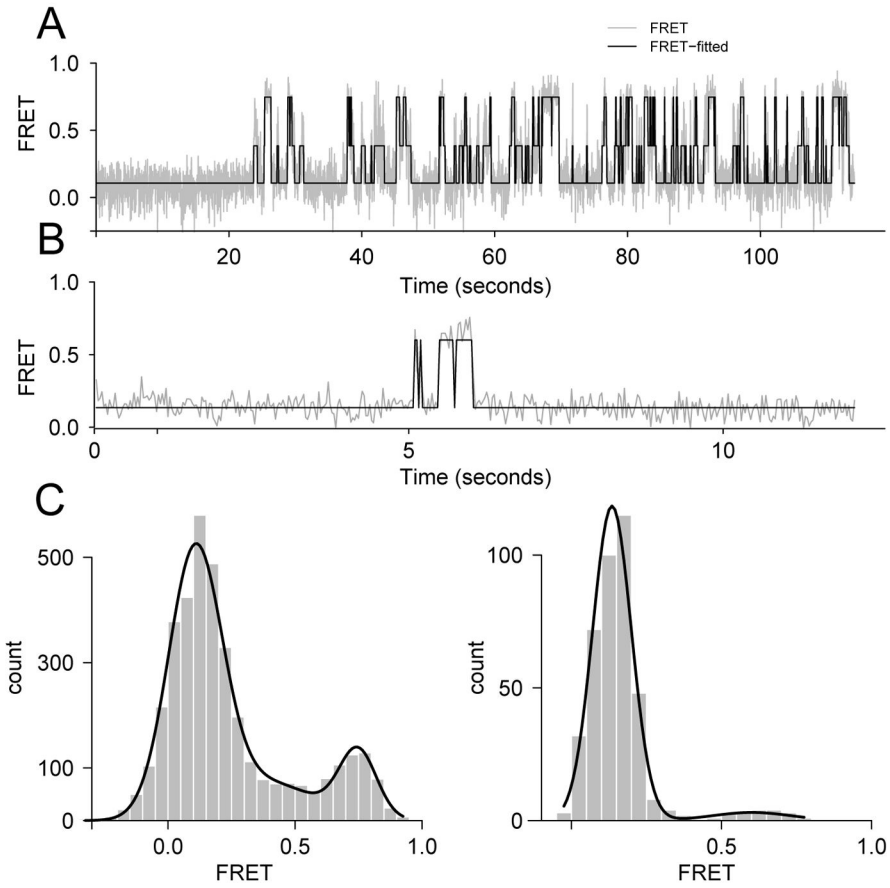

Figure 6.

Two sample FRET trajectories, one long trajectory from the *Ffh-Data* and one short trajectory from the *Translocon-Data*. The trace plots show the fitted hidden states. The lower panel shows the histograms of the experimental FRET values together with the fitted Gaussian mixtures.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

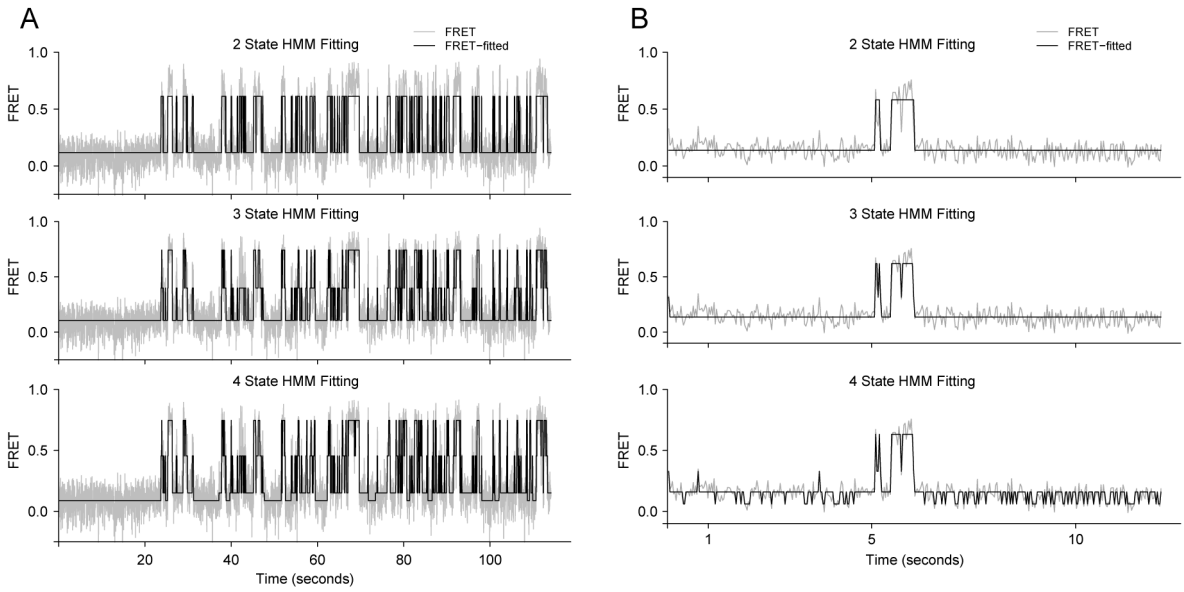

Figure 7.

Fitting of individual FRET trajectories. The left column (A) shows the fitting of the 2-state, 3-state and 4-state HMMs to the long trajectory of Figure 6(A) alone. The right column (B) shows the fitting of 2-state, 3-state and 4-state HMMs to the short trajectory of Figure 6(B) by itself.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

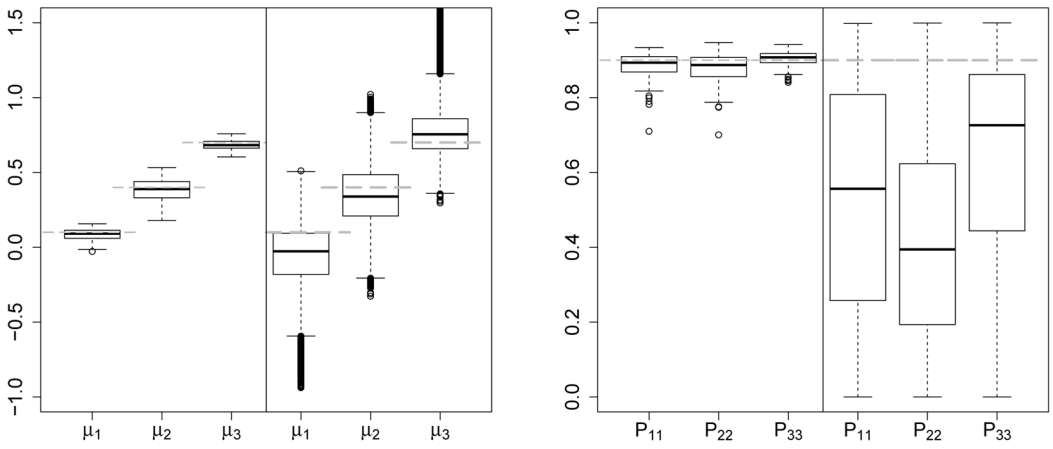

Figure 8.

Comparison of fitting of the hierarchical HMM versus the fitting of individual trajectories. The left panel compares the estimation of the global means $\boldsymbol{\mu}_0$. The right panel compares the estimation of the transition probabilities P_{11}, P_{22}, P_{33}. Both panels use the boxplots. In each panel, the left half shows the posterior distribution under the hierarchical HMM; the right half shows the aggregated posterior distribution based on fitting the 3-state HMM to individual trajectories. The grey horizontal lines correspond to the true values of the parameters.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

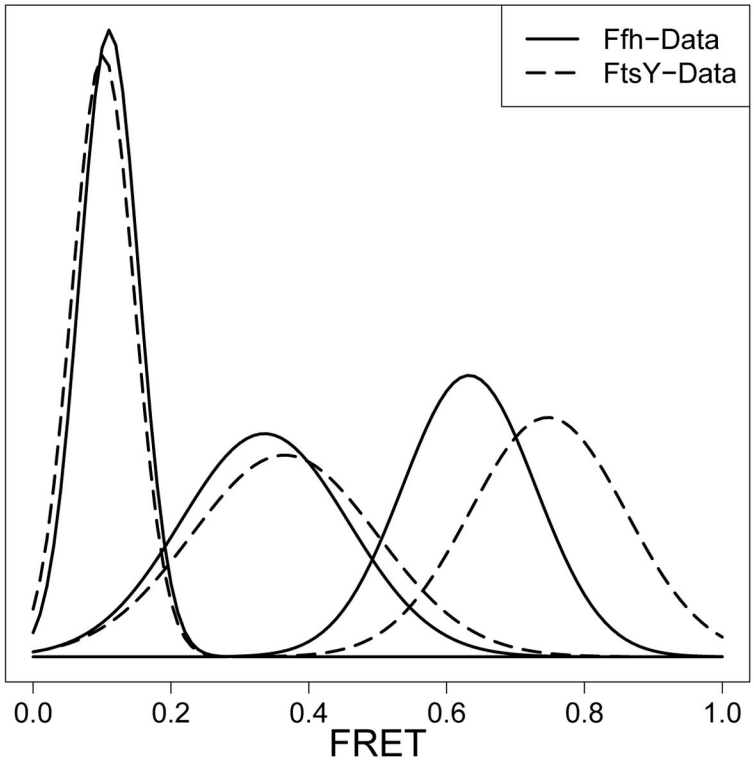

Figure 9.
The posterior distributions of the mean parameters for the *Ffh-Data* and *FtsY-Data*.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

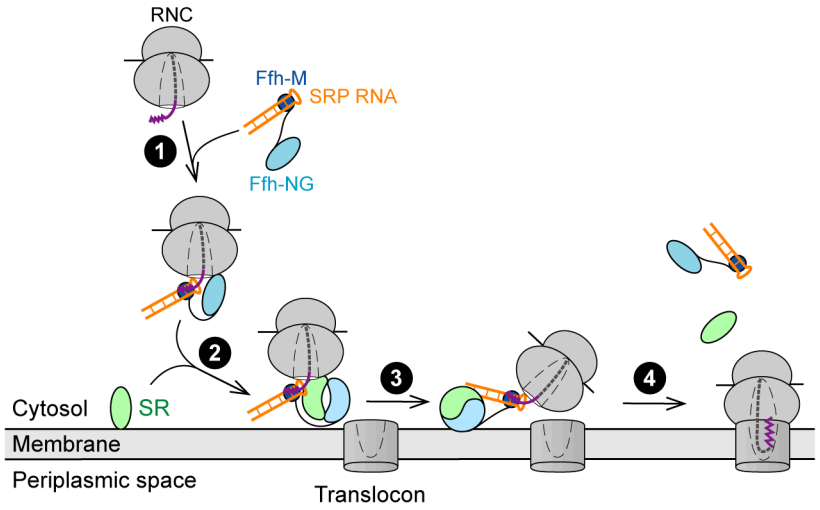

Figure 10.

The refined mechanism. Steps 1 & 2: SRP binds RNC at the RNA capped end and carries it to the membrane by forming a complex with SR located at the membrane. Step 3: The FtsY-[Ffh-NG] complex goes to the distal end so that RNC can be loaded at the translocon. Step 4: SRP-SR disassembles through GTP-hydrolysis and the nascent chain goes through the translocon on the target membrane.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

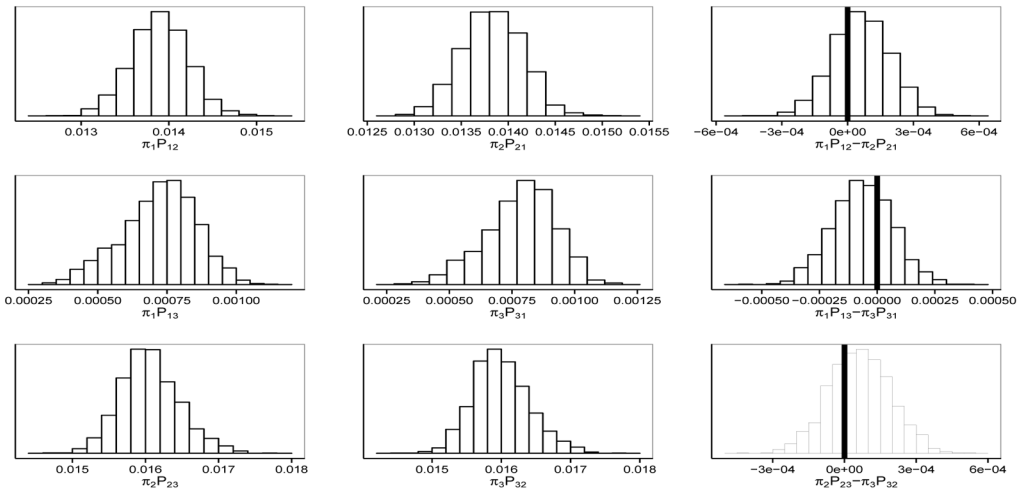

Figure 11.

Check of detailed balance for the *Ffh-Data*. The first column is the posterior distribution of $\pi_i P_{ij}$, and the second column is that of $\pi_j P_{ji}$, where $i, j \in \{1, 2, 3\}$, $i \neq j$. The third column shows the distribution of their difference $\pi_i P_{ij} - \pi_j P_{ji}$; the thick vertical bar is at zero.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

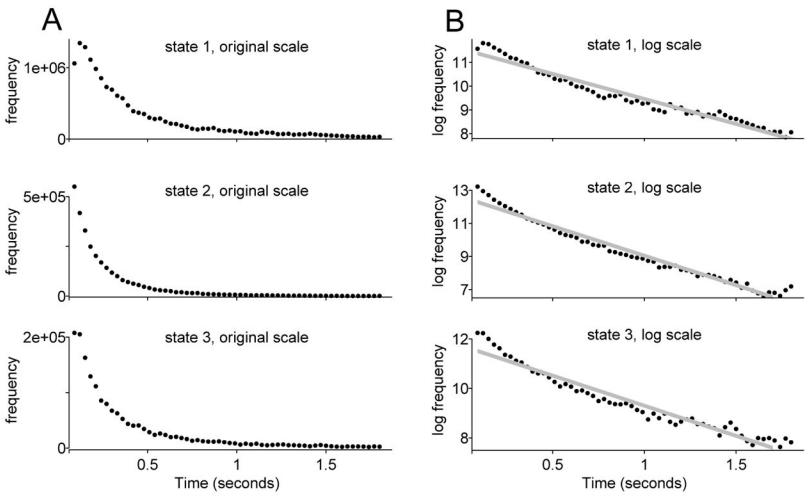

Figure 12.

Posterior distribution of waiting time at the three states of the *Ffh-Data* on the original scale, the left column (**A**); and the log scale, the right column (**B**).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

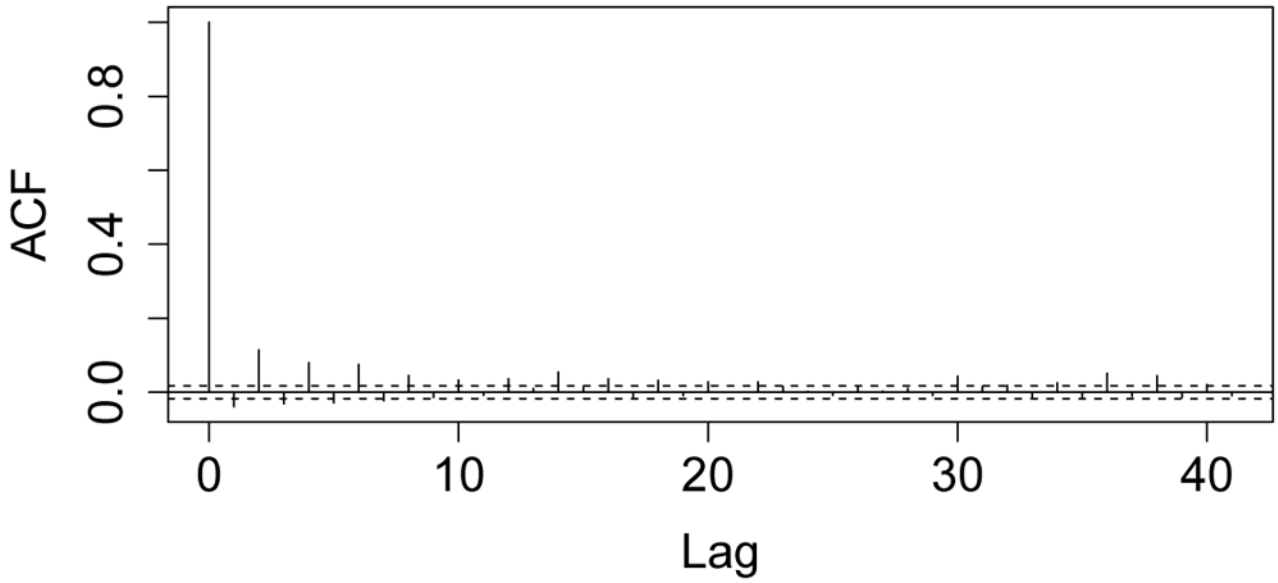

Figure 13.
Autocorrelation of the successive waiting times from the *Ffh-Data*.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Data sets and number of recorded trajectories in each set.

Data Abbreviation	FRET Donor	FRET Acceptor	Complexes in experiments	No. Trajectories
Ffh-Data	Ffh-NG	RNA distal end	SRP-SR	142
FtsY-Data	FtsY	RNA distal end	SRP-SR	208
RNC-Data	Ffh-NG	RNA distal end	SRP-SR, RNC	97
Translocon-Data	Ffh-NG	RNA distal end	SRP-SR, RNC, Translocon	138

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Summary of the lengths (number of data points) of the recorded trajectories in each data set.

	5% Quantile	Median	Mean	95% Quantile
Ffh-Data	518	1484	1681	3390
FtsY-Data	357	1027	1248	2993
RNC-Data	317	746	873	1864
Translocon-Data	338	918	1071	2357

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

The number of trajectories with hidden states K allocated by minimizing BIC_K.

	No. of trajectories allocated				
Data	***K* = 1**	***K* = 2**	***K* = 3**	***K* = 4**	***K* = 5**
Ffh, FtsY-Data	1	21	**159**	136	33
Translocon-Data	2	13	**75**	44	4
RNC-Data	**92**	3	1	1	0

J Am Stat Assoc. Author manuscript; available in PMC 2017 October 18.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Number of trajectories from the *Ffh/FtsY-Data* and *Translocon-Data* assigned to 2, 3, 4 hidden states based on the posterior mode of $|I^{(l)}|$. The hierarchical HMM was fitted twice with three states maximum and four states maximum, respectively. As in Section 3.2, we put the *Ffh-Data* and *FtsY-Data* together in the table.

	No. of trajectories allocated				
Hierarchical HMM	**three states maximum**		**four states maximum**		
No. States	**2**	**3**	**2**	**3**	**4**
Ffh, FtsY-Data	56	**294**	26	**201**	123
Translocon-Data	39	**99**	50	**60**	28

J Am Stat Assoc. Author manuscript; available in PMC 2017 October 18.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5

95% posterior intervals of the global means μ_{0i} and global standard deviations η_{0i}; $i \in \{1, 2, 3\}$ for *Ffh-Data*, *FtsY-Data*, *RNC-Data* and *Translocon-Data*.

Parameters	Ffh-Data	FtsY-Data	RNC-Data	Translocon-Data
$\mu_{0,1}$	[0.105, 0.116]	[0.096, 0.107]	[0.091, 0.099]	[0.097, 0.104]
$\mu_{0,2}$	[0.319, 0.353]	[0.348, 0.382]	NA	[0.380, 0.441]
$\mu_{0,3}$	[0.619, 0.646]	[0.733, 0.761]	NA	[0.619, 0.635]
$\eta_{0,1}$	[0.039, 0.048]	[0.041, 0.049]	[0.017, 0.022]	[0.019,0.023]
$\eta_{0,2}$	[0.110, 0.135]	[0.122, 0.148]	NA	[0.131, 0.169]
$\eta_{0,3}$	[0.087, 0.107]	[0.101, 0.126]	NA	[0.044, 0.058]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 6

Posterior estimates of the transition probabilites (mean ± 2× standard deviations) of *Ffh-Data*, *FtsY-Data*, *Translocon-Data* based on the hierarchical model fitting.

Data	Ffh-Data	FtsY-Data	Translocon-Data
P_{11}	0.9703 ± 0.0014	0.9798 ± 0.0013	0.9976 ± 0.0005
P_{22}	0.8732 ± 0.0054	0.8776 ± 0.0058	0.9713 ± 0.0076
P_{33}	0.9384 ± 0.0027	0.9217 ± 0.0039	0.9870 ± 0.0015
P_{12}	0.0283 ± 0.0014	0.0186 ± 0.0015	0.0011 ± 0.0004
P_{13}	0.0015 ± 0.0005	0.0015 ± 0.0005	0.0013 ± 0.0004
P_{21}	0.0587 ± 0.0034	0.0579 ± 0.0044	0.0044 ± 0.0015
P_{23}	0.0681 ± 0.0036	0.0646 ± 0.0037	0.0244 ± 0.0072
P_{31}	0.0029 ± 0.0010	0.0057 ± 0.0017	0.0022 ± 0.0006
P_{32}	0.0587 ± 0.0031	0.0726 ± 0.0045	0.0108 ± 0.0015

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 7

Compare *Ffh-Data* and *Translocon-Data*: 95% posterior intervals of mean values of the states ($\mu_{0,1}$, $\mu_{0,2}$, $\mu_{0,3}$), dwell time at the high-FRET state (d_3) and the probability that a transitions from low- to high-FRET state goes through the middle state (p_{middle}).

Parameters	Ffh-Data	Translocon-Data
$\mu_{0,1}$	[0.105, 0.116]	[0.097, 0.104]
$\mu_{0,2}$	[0.319, 0.353]	[0.380, 0.441]
$\mu_{0,3}$	[0.619, 0.646]	[0.619, 0.635]
d_3	[0.465, 0.507]	[2.058, 2.577]
p_{middle}	91.2%	40.7%