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

Yang Chen\(^ab\), Kuang Shen\(^b\), Shu-Ou Shan\(^c\), and S. C. Kou\(^a\)

\(^a\)Department of Statistics, Harvard University, Cambridge, MA, USA; \(^b\)Life Sciences Research Foundation, Whitehead Institute for Biomedical Research, Cambridge, MA, USA; \(^c\)Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA

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.

1. Introduction
In cells, proteins often need to be transported to appropriate destinations inside or outside of a cell 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 (Rapoport 1991; Lodish et al. 2000). Cointernalization protein targeting is such a process in which proteins still being synthesized on the ribosome (called ribosome nascent-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-transportal protein targeting process consists of four basic steps (Zhang et al. 2009b; Nyathi, Wilkinson, and Pool 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. 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, Kung, and Shan 2008; Shen and Shan 2010; Ataide et al. 2011; Voigt-Hoffmann et al. 2013; Nyathi, Wilkinson, and Pool 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); 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. FRET (Förster resonance energy transfer) (Roy, Hohng, and Ha 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.

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 heterogenous. 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 dataset: Does the collection of FRET trajectories represent chemically homogenous molecules or molecular complexes? If not, is the heterogeneity biologically relevant?

2. How many states are there in these FRET trajectories? Previous analysis used an arbitrary number of states for HMM (Shen et al. 2012). However, there is no statistical analysis to legitimize 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

![Figure 1](https://example.com/figure1.png)

**Figure 1.** The four steps of protein targeting.

![Figure 2](https://example.com/figure2.png)

**Figure 2.** Three sample FRET trajectories.
equilibrium information from individual trajectories. Rigorous statistical analysis using the information from all trajectories is required to extract this information and understand 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 \((y_1, y_2, \ldots)\) 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, Joo, and Ha 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, Laird, and Rubin 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 photobleaching). 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 question, 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, Bulla, and Nenadic 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 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 appendices contain 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, Hofng, and Ha 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 \((y_1, y_2, \ldots)\), obtained at every 30 millisecond (ms) in our case. \(y_i \in [0, 1]\) is calculated as \(y_i = \text{acceptor fluorescence}/(\text{donor fluorescence} + \text{acceptor fluorescence})\). 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, that is, 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 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 (Keenan et al. 2001; Halic et al. 2004; Akopian et al. 2013b; Zhang, Kung, and Shan 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 \textit{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 Figure 4(c) and 4(d) 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 \textit{Ffh-Data}, \textit{FtsY-Data}, \textit{RNC-Data}, and \textit{Translocon-Data} obtained from these experiments, and Table 2 summarizes the lengths of the trajectories in each dataset. 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.

Table 1. Datasets and number of recorded trajectories in each set.

<table>
<thead>
<tr>
<th>Data abbreviation</th>
<th>FRET donor</th>
<th>FRET acceptor</th>
<th>Complexes in experiments</th>
<th>No. trajectories</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{Ffh-Data}</td>
<td>Ffh-NG</td>
<td>RNA distal end</td>
<td>SRP-SR</td>
<td>142</td>
</tr>
<tr>
<td>\textit{FtsY-Data}</td>
<td>FtsY</td>
<td>RNA distal end</td>
<td>SRP-SR</td>
<td>208</td>
</tr>
<tr>
<td>\textit{RNC-Data}</td>
<td>Ffh-NG</td>
<td>RNA distal end</td>
<td>SRP-SR, RNC</td>
<td>97</td>
</tr>
<tr>
<td>\textit{Translocon-Data}</td>
<td>Ffh-NG</td>
<td>RNA distal end</td>
<td>SRP-SR, RNC, Translocon</td>
<td>138</td>
</tr>
</tbody>
</table>

Table 2. Summary of the lengths (number of data points) of the recorded trajectories in each dataset.

<table>
<thead>
<tr>
<th></th>
<th>5% Quantile</th>
<th>Median</th>
<th>Mean</th>
<th>95% Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{Ffh-Data}</td>
<td>518</td>
<td>1484</td>
<td>1681</td>
<td>3390</td>
</tr>
<tr>
<td>\textit{FtsY-Data}</td>
<td>357</td>
<td>1027</td>
<td>1248</td>
<td>2993</td>
</tr>
<tr>
<td>\textit{RNC-Data}</td>
<td>317</td>
<td>746</td>
<td>873</td>
<td>1864</td>
</tr>
<tr>
<td>\textit{Translocon-Data}</td>
<td>338</td>
<td>918</td>
<td>1071</td>
<td>2357</td>
</tr>
</tbody>
</table>
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 hr. 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 (532 nm) and red (638 nm) lasers were aligned and speaking, the FRET value is between 0 and 1, the Gaussian distribution is a good approximation (Dahan et al. 1999; McKinney, Joo, and Ha 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 three-dimensional structure of a protein or a protein complex. For example, the low- and high-FRET states in Figure 4(c) and 4(d) correspond to two distinct conformations of the SRP-SR complex. Let $P = (P_k)$ be the $K \times K$ transition matrix of $z_i$; it represents the conformational kinetics of a complex. For each FRET trajectory, the parameters are $\Theta = (\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$, where $\mu_k$ and $\sigma_k^2$ are the mean and variance of the FRET value at state $k$; $k = 1, \ldots, K$. Let $\pi = (\pi_1, \ldots, \pi_K)$ be the probabilities that the first hidden state $z_1$ is in state 1, $\ldots, K$. The joint likelihood of observations $y_{1:N}$ and the hidden states $z_{1:N}$ is

$$p(y_{1:N}, z_{1:N} | \Theta) = \pi_{z_1} \prod_{n=2}^{N} p(z_n | z_{n-1}, P) \prod_{n=1}^{N} p(y_n | z_n, \mu, \sigma^2).$$

Please note that for notational ease, we use $y_{m:n}$ to denote the vector $(y_m, y_{m+1}, \ldots, y_n)$ for $m < n$ throughout this article. The marginal likelihood $L(\Theta | y_{1:N}) = \int p(y_{1:N} | z_{1:N} | \Theta)dz_{1:N}$ is given by integrating out $z_{1:N}$ 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, Laird, and Rubin 1977), to calculate the maximum likelihood estimator (MLE) $\hat{\Theta}$. The Baum–Welch/EM algorithm, in addition, can yield the marginal likelihood evaluated at the MLE, $L(\hat{\Theta} | y_{1:N})$. 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 Figure 4(c) and 4(d) 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 nonidentifiable 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

\[ y_i \mid (z_i = k) \sim N(\mu_k, \sigma_k^2), \quad (1) \]

where $z = (z_1, z_2, \ldots, z_N)$ 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, Joo, and Ha 2006; Liu et al. 2010).
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, Allen, and Stromberg 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 (Leroux 1992; Biernacki, Celeux, and Govaert 1998; McLachlan and Peel 2005; Frühwirth-Schnatter 2006). Though the consistency of BIC for Gaussian HMMs has not been completely established (Finesso 1990; Ryden 1995; Cappe, Moulines, and Ryden 2005), it has been shown through simulations that BIC empirically tends to select the correct model when the sample size is large but could give highly variable results when the sample size is small or moderate (Ryden 1995; Keribin 2000; MacKAY 2002; Watkins and Yang 2005; Frühwirth-Schnatter 2006; Celeux and Durand 2008). In the context of FRET trajectories, the variability of BIC for HMMs has also been observed (Blanco and Walter 2010; van de Meent et al. 2014; 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, Joo, and Ha 2006; Celeux and Durand 2008; Bulla, Bulla, and Nenadic 2010; Greenfeld et al. 2012; Keller et al. 2014). We adopt this recommendation.

In our case of a $K$-state HMM, the BIC statistic, denoted by $\text{BIC}_K$, is

$$\text{BIC}_K = -2 \log L(\hat{\theta}|y_{1:N}) + \log N \times (K^2 + 2K - 1),$$

where $\hat{\theta}$ is the MLE of $\theta$ 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 $\text{BIC}_K$ 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, Laird, and Rubin 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 $\mu$ are uniformly generated from $[0, 1]$, the initial values of each row of the transition matrix $P$, and the distribution $\pi$ 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 $\sigma$ 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 $\text{Ffh}$- and $\text{FtsY}$-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 $\text{Ffh}$-, $\text{FtsY}$-, and $\text{Translocon}$-Data and $K = 1$ for $\text{RNC}$-Data. Using the estimation mode to select $K$ reflects “majority rule,” that is, 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.

### Table 3. The number of trajectories with hidden states $K$ allocated by minimizing $\text{BIC}_K$

<table>
<thead>
<tr>
<th>Data</th>
<th>$K = 1$</th>
<th>$K = 2$</th>
<th>$K = 3$</th>
<th>$K = 4$</th>
<th>$K = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Ffh}$-$\text{FtsY}$-Data</td>
<td>1</td>
<td>21</td>
<td>159</td>
<td>136</td>
<td>33</td>
</tr>
<tr>
<td>$\text{Translocon}$-Data</td>
<td>2</td>
<td>13</td>
<td>75</td>
<td>44</td>
<td>4</td>
</tr>
<tr>
<td>$\text{RNC}$-Data</td>
<td>92</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

NOTE: Boldface highlights the number of states allocated by the majority of the trajectories.

**4. Modeling FRET Trajectories with Hierarchical Hidden Markov Model**

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

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 $\{y^{(l)}, z^{(l)}\}$ are the observations and hidden states for trajectory $l$. We assume that the same transition matrix $P$ is shared by all trajectories; for trajectory $l$, the means $(\mu_1^{(l)}, \ldots, \mu_K^{(l)})$ come from a higher level distribution $\mu_i^{(l)} \sim N(\mu_0, \eta_0^2)$ with (vector) hyper-parameters $\mu_0$ and $\eta_0^2$, and the variances $(\sigma_i^2)^{(l)}, \ldots, (\sigma_K^2)^{(l)}$ come from scaled inverse-$\chi^2$ distributions with (vector) hyper-parameters $(v, s^2)$, where $v$ denotes the degrees of freedom and $s^2$ are the scale parameters. The intuition behind this hierarchical HMM is that (i) the transition matrix $P$ 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 $\mu^{(l)}$ and $(\sigma^2)^{(l)}$ 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: \( I^{(l)} \) indicates which states are present in trajectory \( l \). For example, if the maximum number of states is \( K = 3 \), \( I^{(l)} \) can take four values \( I^{(l)} = \{1, 2, 3\}, I^{(l)} = \{1, 2\}, I^{(l)} = \{1, 3\}, \) or \( I^{(l)} = \{2, 3\} \), 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 \( N^{(l)} \) be the number of transitions from state \( i \) to \( j \) in trajectory \( l \); \( N^{(l)} = 0 \) if either state \( i \) or \( j \) does not appear in trajectory \( l \). The likelihood function for trajectory \( l \) is

\[
p(y^{(l)}, z^{(l)} | \mu^{(l)}, \sigma^{(l)}, P, I^{(l)}) = \prod_{i,j=1}^{K} \left( \frac{R_{ij}}{P_{ik}} \right)^{N_{ij}^{(l)}} \prod_{n=1}^{N_{l}} N(y_{n}^{(l)}, \mu_{zi_{n}}, \sigma_{zi_{n}}^{2}),
\]

where \( \left( \frac{R_{ij}}{P_{ik}} \right)_{i,j \in I^{(l)}} \) is the renormalized 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

\[
\prod_{l} p(y^{(l)}, z^{(l)} | \mu^{(l)}, \sigma^{(l)}, P, I^{(l)}) p(\mu^{(l)} | \mu_{0}, \sigma^{(l)} p((\sigma^{(l)})^{2}) | v, s^{2}).
\]

### 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 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 two-state, three-state, and four-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 \( \mu = (\mu_1, \mu_2, \mu_3) \) 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 \( \sigma = (\sigma_1, \sigma_2, \sigma_3) \) 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 \( \mu_0 = (0.1, 0.4, 0.7) \). The right panel compares the estimation of the transition probabilities \( P_{12}, P_{23}, P_{31} \). 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 three-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 = \( \min_k \{ \frac{1}{\sigma_k}\} \) (Hawkins, Allen, and Stromberg 2001; Greenfeld et al. 2012). For the 100 trajectories of Figure 8, the median SNR is 0.3. In contrast, we find from our 16 simulated datasets 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 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 heterogenous 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 heterogenous trajectories are well explained by (ii).

![Figure 5](image-url) Diagram of the hierarchical HMM.

![Figure 6](image-url) 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.
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 $\mu_{0i}$ and $\eta_{0i}$ for the datasets. 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 Figure 4(c)). The state with a high FRET value,

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Ffh-Data</th>
<th>FtsY-Data</th>
<th>RNC-Data</th>
<th>Translocon-Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{0.1}$</td>
<td>[0.105, 0.116]</td>
<td>[0.096, 0.107]</td>
<td>[0.091, 0.099]</td>
<td>[0.097, 0.104]</td>
</tr>
<tr>
<td>$\mu_{0.2}$</td>
<td>[0.319, 0.353]</td>
<td>[0.348, 0.382]</td>
<td>NA</td>
<td>[0.380, 0.441]</td>
</tr>
<tr>
<td>$\mu_{0.3}$</td>
<td>[0.619, 0.646]</td>
<td>[0.733, 0.761]</td>
<td>NA</td>
<td>[0.619, 0.635]</td>
</tr>
<tr>
<td>$\eta_{0.1}$</td>
<td>[0.039, 0.048]</td>
<td>[0.041, 0.049]</td>
<td>[0.017, 0.022]</td>
<td>[0.019, 0.023]</td>
</tr>
<tr>
<td>$\eta_{0.2}$</td>
<td>[0.110, 0.135]</td>
<td>[0.122, 0.148]</td>
<td>NA</td>
<td>[0.131, 0.169]</td>
</tr>
<tr>
<td>$\eta_{0.3}$</td>
<td>[0.087, 0.107]</td>
<td>[0.101, 0.126]</td>
<td>NA</td>
<td>[0.044, 0.058]</td>
</tr>
</tbody>
</table>

Figure 7. Fitting of individual FRET trajectories. The left column (a) shows the fitting of the two-state, three-state, and four-state HMMs to the long trajectory of Figure 6(a) alone. The right column (b) shows the fitting of two-state, three-state, and four-state HMMs to the short trajectory of Figure 6(b) by itself.

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 $\mu_{0i}$. The right panel compares the estimation of the transition probabilities $P_{ij}$, $P_{11}$, $P_{22}$, $P_{33}$. 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 three-state HMM to individual trajectories. The gray horizontal lines correspond to the true values of the parameters.
Figure 9. The posterior distributions of the mean parameters for the Ffh-Data and FtsY-Data.

\( \mu_{0.3} \approx 0.6 \sim 0.8 \), corresponds to the conformation where the FtsY-[Ffh-NG] complex is near the distal end of the RNA (see Figure 4(d)). It is noteworthy that in addition to these two major conformations, our analysis identifies a “middle” state with the FtsY-[Ffh-NG] complex is near the distal end of the RNA (see Figure 4(d)). The observation that both \( d_1 \) and \( d_2 \) are significantly larger than \( d_3 \) 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” steps. 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 datasets. 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 \([0.966, 1.057]\) sec for \( d_1 \), the low-FRET state; \([0.228, 0.249]\) sec for \( d_2 \), the middle state; and \([0.465, 0.507]\) sec for \( d_3 \), the high-FRET state. The observation that both \( d_1 \) and \( d_2 \) are significantly larger than \( d_3 \) 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 \( P_{3} \) is significantly smaller than \( P_{2} \) 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 \( P_{i,j}^{(2)} \) 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 \( \sum_{k=1}^{\infty} P_{1,k}^{(2)} P_{k,3} \) (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 \( P_{1,3} + \sum_{k=1}^{\infty} P_{1,k}^{(1)} P_{k,1} P_{1,3} \). Table 6 satisfies the following recursive formulas, owing to the first-step analysis:

\[
\begin{align*}
P_{1,2}^{(k+1)} &= P_{1,1} P_{1,2}^{(k)} + P_{1,2} P_{2,2}^{(k)} + P_{1,2} P_{2,3}^{(k)} \\
P_{2,1}^{(k+1)} &= P_{2,1} P_{1,1}^{(k)} + P_{2,1} P_{2,1}^{(k)} + P_{2,1} P_{2,2}^{(k)} \\
P_{2,2}^{(k+1)} &= P_{2,1} P_{1,2}^{(k)} + P_{2,2} P_{2,2}^{(k)} + P_{2,2} P_{2,3}^{(k)} \\
\end{align*}
\]

Summing over \( k \) on both sides of the equations yields:

\[
\begin{align*}
\sum_{k=1}^{\infty} P_{1,k}^{(2)} P_{2,3} &= \frac{P_{1,2} P_{2,3}}{(1 - P_{1,1})(1 - P_{1,2}) - P_{1,2} P_{2,1}} \\
P_{1,3} + \sum_{k=1}^{\infty} P_{1,k}^{(1)} P_{1,3} &= \frac{(1 - P_{1,2}) P_{1,3}}{(1 - P_{1,1})(1 - P_{1,2}) - P_{1,2} P_{2,1}}. \\
\end{align*}
\]

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.

<table>
<thead>
<tr>
<th>Data</th>
<th>Ffh-Data</th>
<th>FtsY-Data</th>
<th>Translocon-Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_{1,1} )</td>
<td>0.9703 ± 0.0014</td>
<td>0.9798 ± 0.0013</td>
<td>0.9976 ± 0.0005</td>
</tr>
<tr>
<td>( P_{1,2} )</td>
<td>0.8732 ± 0.0054</td>
<td>0.8776 ± 0.0058</td>
<td>0.9713 ± 0.0076</td>
</tr>
<tr>
<td>( P_{1,3} )</td>
<td>0.9384 ± 0.0027</td>
<td>0.9271 ± 0.0039</td>
<td>0.8707 ± 0.0015</td>
</tr>
<tr>
<td>( P_{2,1} )</td>
<td>0.0283 ± 0.0014</td>
<td>0.0186 ± 0.0015</td>
<td>0.0011 ± 0.0004</td>
</tr>
<tr>
<td>( P_{2,2} )</td>
<td>0.0015 ± 0.0005</td>
<td>0.0015 ± 0.0005</td>
<td>0.0015 ± 0.0004</td>
</tr>
<tr>
<td>( P_{2,3} )</td>
<td>0.0587 ± 0.0034</td>
<td>0.0579 ± 0.0044</td>
<td>0.0444 ± 0.0015</td>
</tr>
<tr>
<td>( P_{3,1} )</td>
<td>0.0681 ± 0.0036</td>
<td>0.0646 ± 0.0037</td>
<td>0.0244 ± 0.0007</td>
</tr>
<tr>
<td>( P_{3,2} )</td>
<td>0.0029 ± 0.0010</td>
<td>0.0007 ± 0.0007</td>
<td>0.0022 ± 0.0006</td>
</tr>
<tr>
<td>( P_{3,3} )</td>
<td>0.0587 ± 0.0031</td>
<td>0.0726 ± 0.0045</td>
<td>0.0108 ± 0.0015</td>
</tr>
</tbody>
</table>
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 large-scale 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 $P_2 > P_1$); 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 $y = \text{const} + \text{Gaussian noise}$, 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 site (see Figure 4(c)). 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 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 quaternary 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 toward 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.

<table>
<thead>
<tr>
<th>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 ($P_{middle}$).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
</tr>
<tr>
<td>$\mu_{0.1}$</td>
</tr>
<tr>
<td>$\mu_{0.2}$</td>
</tr>
<tr>
<td>$\mu_{0.3}$</td>
</tr>
<tr>
<td>$d_3$</td>
</tr>
<tr>
<td>$P_{middle}$</td>
</tr>
</tbody>
</table>

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 $d_3 = 1/(1 - P_3)$, between the Translocon-Data and the Ffh-Data. The 95% posterior interval for $d_3$ is [2.058, 2.577] sec for the Translocon-Data and [0.465, 0.507] sec for the Ffh-Data, respectively. Thus, the translocon enhances the kinetic stability of the SRP-SR complex in the distal conformation by four- to five-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 (Peluso et al. 2001; Pool et al. 2002; Halic et al. 2006; Zhang et al. 2009a; Shen and Shan 2010; Ataide et al. 2011; Estrozzi et al. 2011; Akopian et al. 2013a) suggests the following detailed mechanism of protein targeting, which was conjectured in Shen et al. (2012), corresponding to the four steps of Section 1:

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.
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 \( \pi_i \) is the equilibrium probability of state \( i \). This can be checked from the posterior samples of the transition matrix \( P \).

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 \( z \). Figure 12 shows the posterior 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 \( z \) 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.

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.
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 $z$. 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 reassured 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 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.
Appendix 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, Laird, and Rubin 1977), a.k.a. the Baum–Welch algorithm for HMM (Baum and Petrie 1966; Baum et al. 1970), to infer $\theta$. 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

$$L(\theta) = \sum_{n=2}^{N} p(z_n | z_{n-1}, P) \prod_{n=1}^{N} p(y_n | z_n, \mu, \sigma^2).$$

In addition, the forward–backward algorithm gives the marginal distribution of the $j$th row of the transition matrix $P_j = (P_{j1}, P_{j2}, \ldots, P_{jK})$ is a Dirichlet distribution, the conditional distribution of $\mu$ is a multivariate normal distribution, the conditional distribution of $\sigma^2$ is a multivariate inverse-$\chi^2$ distribution, and that the hidden states $z$ can be sampled sequentially from 1 to $N$ through the following recursion:

$$p(z_n = k | z_{n-1} = j, \theta, y) \propto P_n N(y_n; \mu_k, \sigma^2_k) / P_k N(y_k; \mu, \sigma^2).$$

where $\beta(k)$ is the backward probability defined in Equation (A.2).

Appendix 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 $\theta$ with the hidden states $z$, and iteratively sample one given the other (i.e., sample $\theta$ given $z$ and sample $z$ given $\theta$).

Specifically, in our MCMC sampling, we adopt flat priors for $P$ and $\mu_k, k = 1, \ldots, K$, and independent inverse-$\chi^2$ priors with parameters $v, s^2$ for $\sigma^2_k$ (the prior on $\mu$ is flat over the region $0 < \mu_1 < \cdots < \mu_K < 1$). The posterior distribution is

$$p(\theta, z | y) = p(y, z|\theta) p_0(P) p_0(\mu) p_0(\sigma^2) \propto \prod_{j=1}^{K} \prod_{k=1}^{K} P_{jk} \prod_{n=1}^{N} N(y_n; \mu_k, \sigma^2_k) \prod_{k=1}^{K} p_0(\sigma^2_k; v, s^2).$$

Appendix C: MCMC Sampling of the Hierarchical HMM

The posterior distribution is proportional to

$$p(\mu_0, \eta_0^2, s^2) \prod_{l} p(y^{(l)}, z^{(l)} | I^{(l)}, \mu^{(l)}, \sigma^{(l)}, P) \times \prod_{l} p(\mu^{(l)} | \mu_0, \eta_0^2, I^{(l)}) p(\sigma^{(l)} | v, s^2, I^{(l)}) p(I^{(l)}).$$

where $I^{(l)}$ is the indicator probability defined in Equation (A.2).

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 $I^{(l)}$ and $I_0$ denote the indicator function.

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

2. Update global parameters $\mu_0, \eta_0^2, s^2$. For $1 \leq k \leq K$,

   Sample $\mu_{0,k}$ from $N(\sum_{l=1}^{T} \mu^{(l)}_k / \sum_{l=1}^{T} I_k^{(l)})$.

   Sample $\eta_{0,k}^2$ from Inv-$\chi^2(\sum_{l=1}^{T} I_k^{(l)} - 2, \sum_{l=1}^{T} I_k^{(l)})$.

   Sample $s^2_k$ from $\{v_k \sum_{l=1}^{T} I_k^{(l)} / (\sigma^{(l)}_k)^2\}^{-1}\chi^2_{df}$. $df = v_k \sum_{l=1}^{T} I_k^{(l)} + 2$. 

In addition, the forward–backward algorithm gives the marginal likelihood evaluated at the maximum likelihood estimate $p(y|\theta) = \sum_{z_1} \alpha(z_1) = \sum_{z_1} p(y_{1:N}, z_1 | \theta)$. In the M-step of the EM algorithm, which maximizes $E \log L(\theta|\beta^{old})$ over $\theta$, we obtain $\theta^{new}$ according to

$$P_{jk} = \frac{\sum_{n=2}^{N} u_{n,k,j}}{\sum_{k=1}^{K} \sum_{n=2}^{N} u_{n,k,j}}, \quad \mu_k = \frac{\sum_{n=1}^{N} y_n u_{n,k}}{\sum_{n=1}^{N} u_{n,k}^2},$$

$$\sigma^2_k = \frac{\sum_{n=1}^{N} u_{n,k} (y_n - \mu_k)^2}{\sum_{n=1}^{N} u_{n,k}}.$$
3. Update transition probabilities \( P \) according to

\[
p(P) \propto \prod_{i,j} \sum_k P_{ij}^{(0)} \prod_{l=1}^{K} \prod_{n=1}^{N} \left( \sum_{k} P_{k} \right) \sum_{o(e)} N_{ij}^{(e)} \].

4. Update parameters for individual trajectories.

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

\[
\mu^{(l)}_k \sim N\left( \frac{\mu_0 n_0^2 + \sum_{n=1}^{N} I(z_{n}^{(l)} = k)}{\sigma^{(l)}_k^2 + \sum_{n=1}^{N} I(z_{n}^{(l)} = k)}, \frac{1}{\sigma^{(l)}_k^2 + \sum_{n=1}^{N} I(z_{n}^{(l)} = k)} \right)
\]

- Update \( \{ \sigma^{(l)} \} \). 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 renormalized submatrix of \( P \) 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

\[
p(y^{(l)}, z^{(l)} ; \mu^{(l)}, \sigma^{(l)}, P, I^{(l)} = A) \propto p(\mu^{(l)}; \eta_0^2, I^{(l)} = A)
\]

5. Iterate Steps 2 to 4 until convergence.

Funding

S. Shan’s research is supported in part by NIH grant GM078024 and the Gordon and Betty Moore Foundation through Grant GBMF2939. S. C. Kou’s research is supported in part by grants from NSF and ARO.

References


Akopian, D., Dalal, K., Shen, K., Duong, F., and Shan, S. (2013a), “SecYEG Activates GTPases to Drive the Completion of Cotranslational Protein Targeting,” The Journal of Cell Biology, 200, 397–405. [961]


