Maximum likelihood-based analysis of photon arrival trajectories in single-molecule FRET

ARTICLE *in* CHEMICAL PHYSICS · MAY 2012

Impact Factor: 1.65 · DOI: 10.1016/j.chemphys.2012.05.009

Chemical Physics 403 (2012) 52–58

Contents lists available at SciVerse ScienceDirect

Chemical Physics

journal homepage: www.elsevier.com/locate/chemphys

Maximum likelihood-based analysis of photon arrival trajectories in single-molecule FRET

Marta Waligórska, Andrzej Molski *

Adam Mickiewicz University, Faculty of Chemistry, Grunwaldzka 6, 60-780 Poznań, Poland

ARTICLE INFO

Article history:
Received 17 November 2011
In final form 11 May 2012
Available online 22 May 2012

Keywords:
Single molecules
FRET
Two-color photon trajectories
Data analysis

ABSTRACT

When two fluorophores (donor and acceptor) are attached to an immobilized biomolecule, anti-correlated fluctuations of the donor and acceptor fluorescence caused by Förster resonance energy transfer (FRET) report on the conformational kinetics of the molecule. Here we assess the maximum likelihood-based analysis of donor and acceptor photon arrival trajectories as a method for extracting the conformational kinetics. Using computer generated data we quantify the accuracy and precision of parameter estimates and the efficiency of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) in selecting the true kinetic model. We find that the number of observed photons is the key parameter determining parameter estimation and model selection. For long trajectories, one can extract mean dwell times that are comparable to inter-photon times.

1. Introduction

Förster resonance energy transfer (FRET) is the transfer of energy from an excited donor to an acceptor, whose efficiency depends on the distance between the donor and acceptor [1,2]. When two fluorophores (donor and acceptor) are attached to an immobilized molecule, anti-correlated fluctuations of the donor and acceptor fluorescence caused by FRET report on the conformational kinetics of the molecule. Here we are interested in extracting the kinetics from the statistical analysis of the donor and acceptor photon arrival trajectories. We focus on the case when the molecule jumps between discrete states.

Several method have been developed to extract the relevant-kinetic information in single-molecule fluorescence spectroscopy [2–5]. In the common on–off analysis the donor and acceptor photon arrivals are binned and the FRET ratio $E = I^A/(I^D + I^A)$ is calculated, where I^D and I^A are the donor and acceptor fluorescence intensities. A threshold is used to separate a high (on) FRET state and a low (off) FRET state in the FRET ratio trajectory (Fig. 1). Histograms of the on- and off-dwell times are then used to extract the on- and off-state escape rates. The outcome of such an analysis depends on the choice of the bin width and threshold. An increased bin width gives a better signal-to-noise ratio, but lowers the time resolution since rapid transitions cannot be detected. This is a drawback of all methods that involve binning including those based on Hidden Markov models [6–8].

A sequence of photon arrivals marked with the detection channel (donor or acceptor) is called a two-color photon trajectory [9]. This trajectory carries all the information about the internal kinetics of the molecule. Equivalent information is contained in the sequence of interarrival times (Fig. 2). In [10,11] a Bayesian perspective was adopted to the analysis of two-color photon trajectories from single-molecule fluorescence experiments. Recently, Gopich and Szabo [9] developed a formula for the likelihood function of two-color photon trajectories and demonstrated that the maximum likelihood (ML) analysis can be used to extract the kinetic parameters and distinguish models. This approach was applied to experimental data in [12].

The purpose of the present paper is to explore in a systematic way the statistical properties of the ML-based analysis of two-color photon trajectories. We quantify the bias and error of parameter estimates and show that the theory developed by Gopich and Szabo [9] can be useful not only for simple two-state models but also for larger models with different topologies (linear and cyclic). Moreover, we justify the use of the Akaike information criterion (AIC) [13] and the Bayesian information criterion (BIC) [14] for model selection. The preset paper extends our previous work [15] where we demonstrated that the the maximum likelihood-based analysis is effective in extracting kinetic information from one-color photon trajectories.

The outline of the paper is as follows. In the next section, we formulate the problem in terms of a Marked Markov Modulated Poisson Process (MMMPP). In Section 3 we present our procedure for data analysis and give examples of parameter estimation for models with several states. In Section 4 we compare the BIC and the AIC as tools for selecting the true model. In Section 5 we apply the present approach to an exemplary system that was recently studied by Chung et al. [12]. A summary and comments are given in the last section.

* Corresponding author.
E-mail address: amolski@amu.edu.pl (A. Molski).

0301-0104/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemphys.2012.05.009

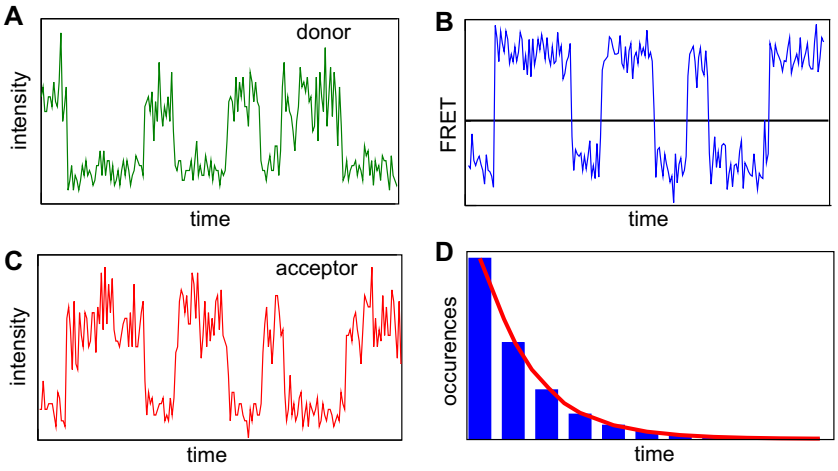


Fig. 1. On–off analysis of a FRET trajectory. Panels A and C show the anti-correlated fluorescence intensity fluctuations in the donor and acceptor channels. The calculated FRET ratio trajectory is shown in panel B together with a threshold separating high (on) and low (off) FRET states. The mean dwell times can be estimated by fitting the on and off time histograms as schematically shown in panel D.

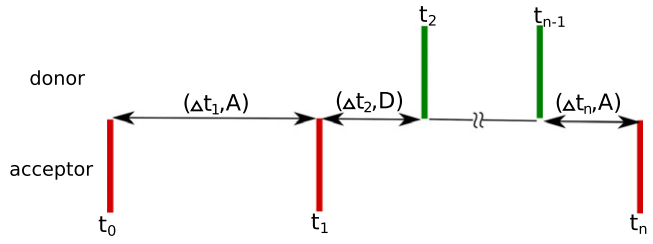


Fig. 2. Graphical representation of a two-color photon trajectory with arrival times $t_0, t_1, \ldots, t_{n-1}, t_n$. The green bars indicate arrivals at the donor channel, and the red bars indicate the arrivals at the acceptor channel. The corresponding marks are c_0 =A, c_1 =A, c_2 =D,. . ., c_{n-1} =D, c_n =A.

2. Marked Markov Modulated Poisson Processes

Below we model two-color photon arrivals as a Marked Markov Modulated Poisson Process (MMMPP) and recall a formula for calculating the likelihood of an arrival trajectory [9].

We assume that the molecule has m internal states, and that its kinetics are determined by a rate matrix $\mathbf{Q}$, whose elements are the rate constants k_{ij} for transitions between states $i \to j$. If the molecule is in state i at time t, it can jump in an infinitesimal time interval $(t, t+dt)$ to a different state $j \neq i$ with conditional probability $k_{ij}dt$, or it can stay in state $i$with probability $1 - \sum_{j\neq i} k_{ij}dt$. The off-diagonal elements $k_{ij},\ i \neq j$, of matrix $\mathbf{Q}$ are positive and determine the jump rates between different states. The diagonal elements k_{ii} are negative, $k_{ii} = -\sum_{j\neq i} k_{ij}$, and determine the escape rates from states i.

The Poisson rates of arrivals in the donor (D) and acceptor (A) channels, i.e. the average number of photons detected per unit time, are determined by arrival rate matrices $\mathbf{\Lambda}^{\mathrm{D}}$ and $\mathbf{\Lambda}^{\mathrm{A}}$. The diagonal elements of $\mathbf{\Lambda}^{\mathrm{D}}$ and $\mathbf{\Lambda}^{\mathrm{A}}$ are the Poisson arrival rates I_i^{D}, I_i^{A} for states $i = 1, \ldots, m$, and the off-diagonal elements are all zeros. The sum of $\mathbf{\Lambda}^{\mathrm{D}}$ and $\mathbf{\Lambda}^{\mathrm{A}}$ is denoted $\mathbf{\Lambda}$. When the molecule stays in state i in $(t, t+dt)$, it generates an arrival with conditional probability $\left(I_i^{\mathrm{D}} + I_i^{\mathrm{A}}\right)dt$. Each arrival $t_i$is characterized by the corresponding detection channel $c_i = \mathrm{D}$ or A. Arrival times $t_0, t_1, \ldots, t_n$, marked with the corresponding detection channels $c_0, c_1, \ldots, c_n$ form a MMMPP that is used here to model photon arrival time trajectories of single immobilized molecules. We focus on the statistics of the interarrival times $\Delta t_k = t_k - t_{k-1}$ and the detection channels c_k.

Let $s(t)$ denote the state of the underlying rate processes at time t, and $s_k,\ k \geqslant 0$, denote the state right after a k-th arrival $s_k = s(t_k)$. Thus $s_k = i$, if the arrival happened when the processes stayed in state i. Let $\Delta t_k,\ k \geqslant 0$ denote the interval time between the $(k-1)$st and the kth arrivals, $\Delta t_k = t_k - t_{k-1},\ \Delta t_0 = 0$. Let c_k denote the detection channel (D or A) at the k-th arrival. The key observation for developing a formula for the likelihood of the observed arrival trajectory $(t_0, c_0), (t_1, c_1), \ldots, (t_n, c_n)$ is that the bi-variate process $\{(s_k, \Delta t_k);\ k \geqslant 0\}$ is a Markov renewal process. Note that the states s_k at arrivals t_k are not observed. The likelihood function of an arrival trajectory $(t_0, c_0), (t_1, c_1), \ldots, (t_n, c_n)$ was obtained by Gopich and Szabo [9] and can be written as

$$L = \mathbf{p}^{\mathrm{ini}} \mathbf{\Lambda}^{c_0} \prod_{k=1}^{n} \{\exp\left[(\mathbf{Q} - \mathbf{\Lambda})\Delta t_k\right] \mathbf{\Lambda}^{c_k}\} \mathbf{1} \tag{1}$$

where $\mathbf{p}^{\mathrm{ini}}$ is a row vector of the initial state probabilities, $\mathbf{\Lambda} = \mathbf{\Lambda}^{\mathrm{D}} + \mathbf{\Lambda}^{\mathrm{A}}$, and $\mathbf{1}$ is a column vector of ones $\mathbf{1} = [1, 1, \ldots, 1]^T$. Analogous expressions were obtained by Kou et al. for single-molecule photon delay trajectories [16], and Ryden for one-color Markov Modulated Poisson Processes [17,18].

We simulated photon arrivals trajectories following the definition of a MMMPP, i.e. assuming a Markov state jump process with escape rate $k_i = \sum_{j\neq i} k_{ij}$ in state i and Poisson photon arrivals with intensities $I_i = I_i^{\mathrm{D}} + I_i^{\mathrm{A}}$. Arrivals are detected in channel D with probability $P_i^{\mathrm{D}} = I_i^{\mathrm{D}} / \left(I_i^{\mathrm{D}} + I_i^{\mathrm{A}}\right)$ or in channel A with probability $P_i^{\mathrm{A}} = 1 - P_i^{\mathrm{D}}$. Note that $P_i^{\mathrm{A}} = I_i^{\mathrm{A}} / \left(I_i^{\mathrm{D}} + I_i^{\mathrm{A}}\right)$ is often referred to as the FRET ratio E_i. State trajectories were simulated by drawing dwell times from exponential distributions with rates k_i. At each state jump the destination state was drawn from a discrete distribution determined by individual transition rates $k_{ij},\ j \neq i$. Photon arrivals in state $i$were simulated by drawing photon waiting times from an exponential distribution with rate I_i. Each arrival was marked D or A according to the probabilities P_i^{D} and P_i^{A}. When the waiting time for a next photon included a state jump, the arrival was discarded, and a new photon waiting time starting at the state jump was drawn with the rate corresponding to the new state.

3. Parameter estimation

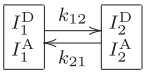
Estimates of the rate constants $\widehat{k}_{ij}$ for transition between states i and j, and intensities in the detection channels, $\widehat{I}_i^{\mathrm{D}}, \widehat{I}_i^{\mathrm{A}}$, were obtained by maximizing the likelihood function $L(1)$. The maximiza-

tion was carried out using an algorithm developed and implemented in C++ by Klemm et al. [19].

The rate constants and intensities have units of inverse time. Here we use relative units which makes the presentation slightly more general. We repeated each experiment 10^3 times, recorded the fit results, and then explored the statistical properties of the estimates. The parameters estimates are scaled by the true value $\kappa_{ij} = \widehat{k}_{ij}/k_{ij}$, $\iota_i^{\mathrm{D}} = \widehat{I}_i^{\mathrm{D}}/I_i^{\mathrm{D}}$, $\iota_i^{\mathrm{A}} = \widehat{I}_i^{\mathrm{A}}/I_i^{\mathrm{A}}$. In the presentation we focus on the rate constants for state transitions because they are more interesting to the experimenter.

3.1. Two-state model

We start off with the simplest two-state model



model 2 (2)

with $k_{12} = k_{21} = 1$ and explore the effect of the fluorescence intensity and the trajectory length on parameter recovery. We assume that the total fluorescence intensity in both FRET states are equal $I_1^{\mathrm{D}} + I_1^{\mathrm{A}} = I_2^{\mathrm{D}} + I_2^{\mathrm{A}} = I$. The subscripts represent the FRET state and the superscripts represent the detection channel (photon color). Also we assume that $I_1^{\mathrm{A}}/I_1^{\mathrm{D}} = I_2^{\mathrm{D}}/I_2^{\mathrm{A}} = 0.25$, which corresponds to the difference in the FRET ratio $\Delta E = E_2 - E_1 = 0.6$. Table 1 presents the mean and standard deviation of the scaled parameter estimates for the trajectory length n ranging from $n = 10^3$ to 5×10^4 photons and the total intensity I from 1 to 100. The kinetic parameters can be recovered with acceptable precision (say the standard deviation $\leqslant 20\%$) even for short trajectories made up of 5×10^3 photons when the fluorescence intensity is appropriate, I = 5, 10, 50. When the intensity is too low, $I = 1$, the inter-photon times are comparable to the mean dwell times and the mean dwell times cannot be extracted. When the intensity is too high, I = 100, only a small number of state transitions is observed. It follows from Table 1 that the optimal fluorescence intensity is about 10, i.e. the optimal mean number of arrivals per dwell time is about 10.

The key question when assessing the ML method is whether it gives a better time resolution than methods involving binning. Table 2 shows the mean and standard deviation of 10^3 fits using the on–off analysis and should be compared with Table 1. For I = 1, 5 and 10 the fluorescence intensity was too low for the FRET trajectory to be quantified using on–off analysis. For I = 50 and 100, we used the optimal trajectory bin sizes h = 0.22 and 0.15, respectively. For I = 50 the kinetic parameters were recovered for all trajectory lengths. For I = 100 and n= 5×10^3 photons, too few state jumps were observed.

Comparing the on–off analysis with the ML analysis, we conclude that the latter works better. We note that the on–off analysis cannot resolve dwell times when they are comparable or shorter than the bin width. On the other hand the ML can recover short *mean* dwell times that are comparable with the inter-photon times (see below). Thus the time resolution of the on–off analysis is limited by the bin width, whereas that of the ML method is limited by the mean inter-photon time determined by fluorescence intensity.

We also use the two-state model to illustrate the effect of the difference ΔE of the FRET ratio in the two states. Fig. 3 shows how the estimates change for different ΔE and different trajectory lengths from 5×10^3 to 5×10^4 photons, when $I_1^{\mathrm{D}} + I_1^{\mathrm{A}} = I_2^{\mathrm{D}} + I_2^{\mathrm{A}} = 50$. The estimates strongly dependent on the trajectory length. For long trajectories (5×10^4 photons) one can get acceptable (standard deviation $\leqslant 20\%$) estimates even when $\Delta E \approx 0.2$. Note that the ML analysis does not introduce substantial bias in estimated parameters, however the distribution of the estimated parameters can be skewed for short trajectories (cf. Fig. 9a).

Interestingly, the ML analysis can recover the rate constants when the dwell times are comparable with the inter-photon time. Fig. 4 shows the parameter estimates when the transition rate k_{21} changes from slow ($k_{21} \approx 1$), to fast ($k_{21} = 50$). The faster the escape rate the worse the parameter estimates. Fig. 5 shows the parameter estimates when both kinetic rates $k_{12} = k_{21} = k$ change from slow ($k = 1$) to fast ($k = 100$). Note that the rate constants can be recovered even when the dwell times are smaller than the mean inter-photon time.

3.2. 3- and 4-state models

We use the following models to show that the maximum likelihood analysis is effective for linear (L) and cyclic (C) models with more than two states:

Table 1
The mean (and standard deviation) of 10^3 scaled parameter estimates for a two-state model with unit transition rates and different donor, $I_1^{\mathrm{D}}, I_2^{\mathrm{D}}$, and acceptor, $I_1^{\mathrm{A}}, I_2^{\mathrm{A}}$, fluorescence intensities and trajectory lengths n.

Channel	I_1	I_2	I		$n = 10^3$	$n = 5 \times 10^3$	$n = 10^4$	$n = 5 \times 10^4$
D	0.8	0.2	1	κ_{12}	1.15 (0.68)	1.00 (0.24)	1.01 (0.15)	1.001 (0.065)
A	0.2	0.8		κ_{21}	1.09 (0.65)	1.02 (0.24)	1.00 (0.16)	1.003 (0.063)
D	4	1	5	κ_{12}	1.03 (0.29)	1.00 (0.12)	1.007 (0.087)	1.000 (0.040)
A	1	4		κ_{21}	1.04 (0.29)	1.01 (0.12)	1.006 (0.088)	0.999 (0.038)
D	8	2	10	κ_{12}	1.03 (0.29)	1.01 (0.12)	1.006 (0.082)	0.999 (0.037)
A	2	8		κ_{21}	1.02 (0.29)	1.00 (0.12)	1.004 (0.083)	1.000 (0.038)
D	40	10	50	κ_{12}	1.08 (0.41)	1.01 (0.17)	1.01 (0.12)	1.003 (0.055)
A	10	40		κ_{21}	1.06 (0.43)	1.01 (0.17)	1.00 (0.12)	1.004 (0.051)
D	80	20	100	κ_{12}	1.16 (0.60)	1.03 (0.23)	1.01 (0.16)	1.000 (0.070)
A	20	80		κ_{21}	1.16 (0.63)	1.02 (0.23)	1.01 (0.16)	0.999 (0.069)

Table 2
On–off analysis. The mean (and standard deviation) of 10^3 scaled estimates of the transition rates for a two-state model with unit transition rates and different trajectory lengths n. The bin width $h = 0.22$ for $I = 50$, and $h = 0.15$ for $I = 100$.

Channel	I_1	I_2	I		$n = 10^3$	$n = 5 \times 10^3$	$n = 10^4$	$n = 5 \times 10^4$
D	40	10	50	κ_{12}	1.02 (0.58)	1.07 (0.29)	1.05 (0.16)	1.013 (0.060)
A	10	40		κ_{21}	1.00 (0.56)	1.04 (0.28)	1.04 (0.16)	1.011 (0.063)
D	80	20	100	κ_{12}	-	0.93 (0.42)	1.02 (0.29)	1.000 (0.086)
A	20	80		κ_{21}	-	0.94 (0.41)	1.03 (0.27)	1.030 (0.084)

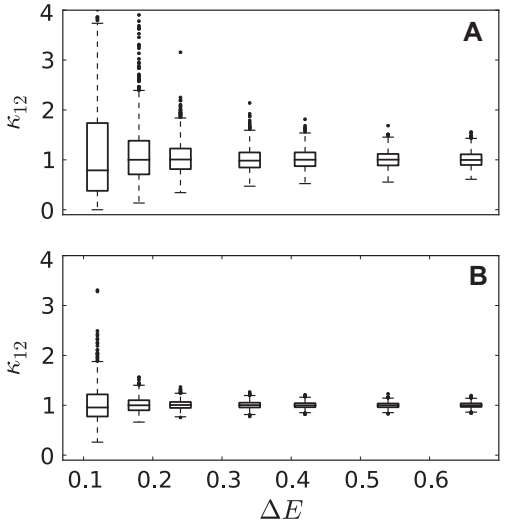


Fig. 3. Effect of the FRET ratio difference ΔE. Box plots of 10^3 scaled estimates of k_{12} for a two-state model with $k_{12} = k_{21} = 1$ and $I_i^{\mathrm{D}} + I_i^{\mathrm{A}} = 50$. Panel A: the trajectory length $n = 5 \times 10^3$ photons, panel B: $n = 5 \times 10^4$ photons.

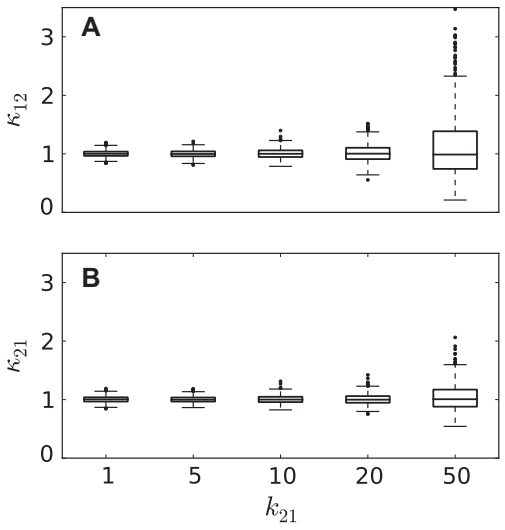
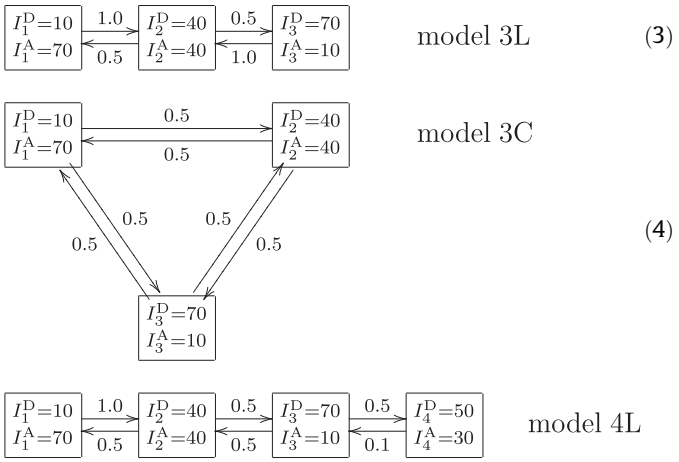


Fig. 4. Effect of the transition rate. Box plots of 10^3 scaled estimates of k_{12} and k_{21} for a two-state model with $k_{12} = 1$, $I_1^{\mathrm{D}} = 10$, $I_1^{\mathrm{A}} = 40$, $I_2^{\mathrm{D}} = 40$, $I_2^{\mathrm{A}} = 10$ and different k_{21}. The trajectory length $n = 5 \times 10^4$ photons.



model 3L (3)

model 3C (4)

model 4L (5)

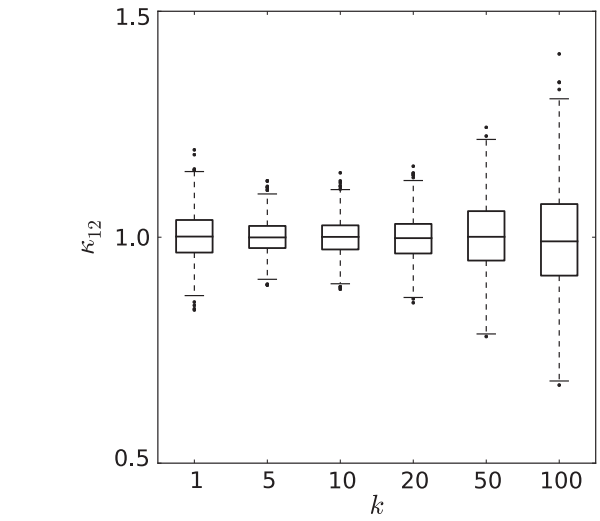


Fig. 5. Effect of the transition rate. Box plots of 10^3 scaled estimates of k_{12} and k_{21} for a two-state model with $I_1^{\mathrm{D}} = 10$, $I_1^{\mathrm{A}} = 40$, $I_2^{\mathrm{D}} = 40$, $I_2^{\mathrm{A}} = 10$ and different $k = k_{12} = k_{21}$. The trajectory length $n = 5 \times 10^4$ photons.

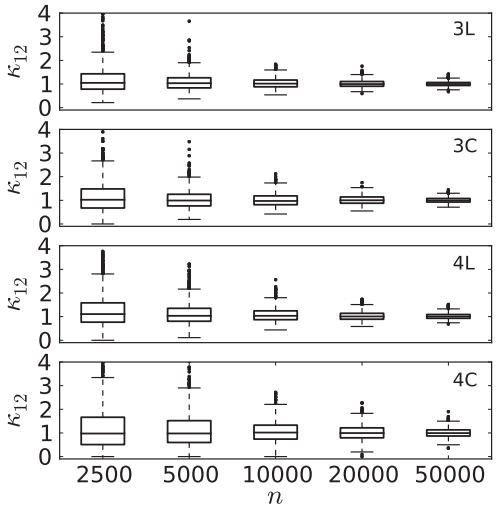
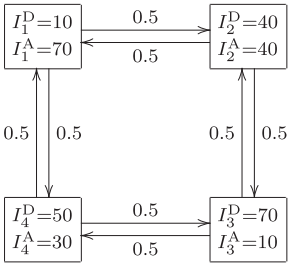


Fig. 6. Effect of the trajectory length. Box plots of 10^3 scaled estimates of k_{12} for models 3L (3), 3C (4), 4L (5) and 4C (6) and different trajectory lengths n.



model 4C (6)

Fig. 6 presents the scaled estimates of k_{12} for 3- and 4-state models where the transition rates were selected so that the mean dwell time in each state is 1. Clearly, this example demonstrates that the ML analysis is not limited to the simplest two-state model but also can be applied to more involved models. As expected, the larger number of states the longer trajectories are needed to get comparable model estimates. Note however, the present ML analysis estimates not only the transition rates but also the fluorescence intensities.

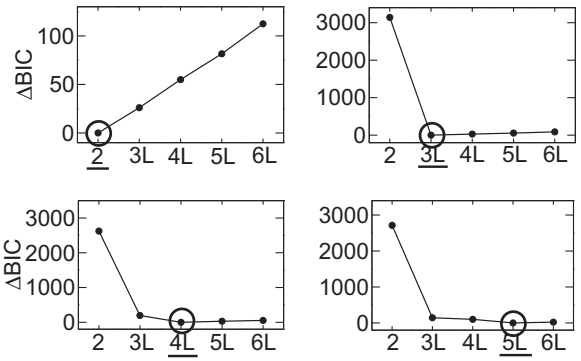


Fig. 7. Examples of model selection. Photon trajectories ($n = 5 \times 10^4$) simulated for models 2, 3L, 4L, 5L were fitted with models 2, 3L, 4L, 5L, 6L. The fitted models are located along the x axis with increasing number of states m = 2–6. The y axis gives ΔBIC = BIC − BIC_{min} for each model. The simulated models are marked with a circle, and the models selected using the BIC criterion are underlined.

Table 3
Efficiency of model selection expressed as the number of experiments (out of 10^2) where the AIC and BIC selected the true model. Models 3L, 3C, 4L and 4C were simulated for different trajectory lengths n. Models 2, 3L, 3C, 4L, 4C and 5L were fitted.

Criterion simulated model	AIC				BIC			
	3L	3C	4L	4C	3L	3C	4L	4C
$n = 5 \times 10^3$	83	75	81	42	100	87	1	0
$n = 10^4$	86	90	96	90	100	100	44	3
$n = 5 \times 10^4$	85	87	98	100	100	100	100	97

4. Model selection

In our model selection study we simulated 10^2 photon arrival trajectory for a given model, fitted different models to the same data and used the Akakie information criterion (AIC) [13] and and the Bayesian information criterion (BIC) [14] to select the most probable model.

The simulated photon trajectories were fitted using models with different number of states and different topologies. The AIC allows for the model likelihood L and the number of model parameters p:

$$\text{AIC} = -2\ln L + 2p \tag{7}$$

where L is the maximized likelihood for the estimated model, and p is the number of free parameters to be estimated. For given data the best model is the one with the lowest AIC in (7). The BIC allows not only for the model likelihood L and the number of model parameters $p$but also for the trajectory length n:

$$\text{BIC} = -2\ln L + p\ln n \tag{8}$$

Fig. 7 shows examples of model selection for models with increasing number of states $m = 2, \ldots, 5$. The difference ΔBIC between the BIC calculated for each fit and the lowest BIC value is plotted against the fitted model. For a selected model $\Delta\text{BIC} = 0$. Correct selection of the true model is possible even in the case of a 5-state model.

Table 3 compiles the results of model selection for different trajectory lengths n. Photon trajectories for models 3L, 3C, 4L and 4C were simulated and models 2, 3L, 3C, 4L, 4C and 5L were fitted to the trajectories. In general the BIC works better for long trajectories. However, even for short trajectories where the total number of donor and acceptor photons is $n = 5 \times 10^3$ a simple (3L or 3C) model was correctly selected in 87 out of 10^2 cases.

5. Illustrative example

So far we have explored general trends in the performance of the maximum likelihood method as applied to photon arrival trajectories. To further asses this approach, in this section we focus on an exemplary system that was studied recently [12]. Chung et al. used 2-color photon trajectories to measure the folding kinetics of α_3 D protein at different concentrations of GdmCl [12]. They proposed a two-state model

$$U \underset{k_U}{\overset{k_F}{\rightleftarrows}} F \tag{9}$$

where U and F denote the unfolded and folded states of the protein, and k_U and k_F are the *feeding* rates of states U and F, respectively.

To illustrate the statistical properties of the maximum likelihood approach, we show in Figs. 8,9a,9b,10 how the accuracy and precision of scaled estimates of k_U and k_F, $\kappa_U = \widehat{k}_U/k_U$ and $\kappa_F = \widehat{k}_F/k_F$, change with trajectory length n for two-state model (9) with $k_F = 0.45\ \text{ms}^{-1}$ and $k_U = 0.55\ \text{ms}^{-1}$. The simulation parameters correspond to those for 2.25 M of GdmCl (cf. Tables 1 and 2 in [12]), except that in [12] only long trajectories were analyzed, $n \approx 3 \times 10^5$ photons. The total (donor + acceptor) fluores-

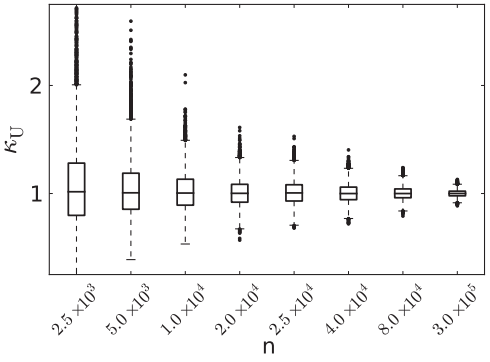


Fig. 8. Effect of the trajectory length n on scaled estimates of k_U. Box plots of 10^4 estimates for two-state model (9) with $k_F = 0.45\ \text{ms}^{-1}$, $k_U = 0.55\ \text{ms}^{-1}$, equal total fluorescence intensities in the unfolded and folded states, $I_U = I_F = 50\ \text{ms}^{-1}$, and the FRET ratios $E_U = 0.61$ and $E_F = 0.93$, respectively.

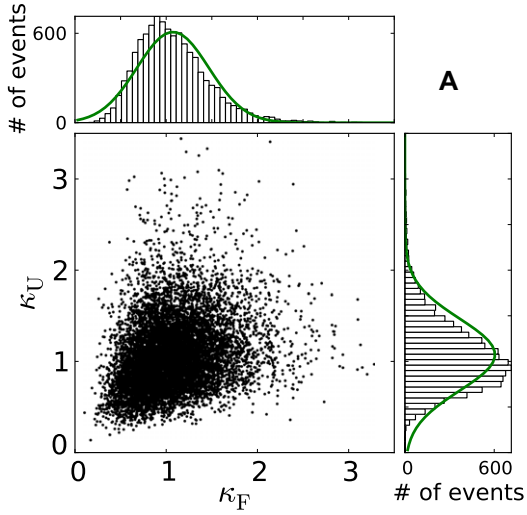


Fig. 9a. Correlations between scaled estimates of k_U and k_F. The trajectory length is $n = 2.5 \times 10^3$ photons. Solid lines: Gaussian probability densities corresponding to the mean values and standard deviations from 10^4 estimates. Simulation parameters as in Fig. 8.

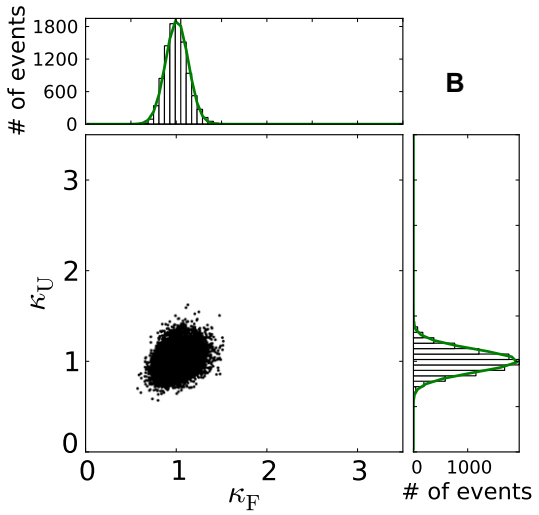


Fig. 9b. Same as in Fig. 9a, except for the trajectory length $n = 2 \times 10^4$ photons.

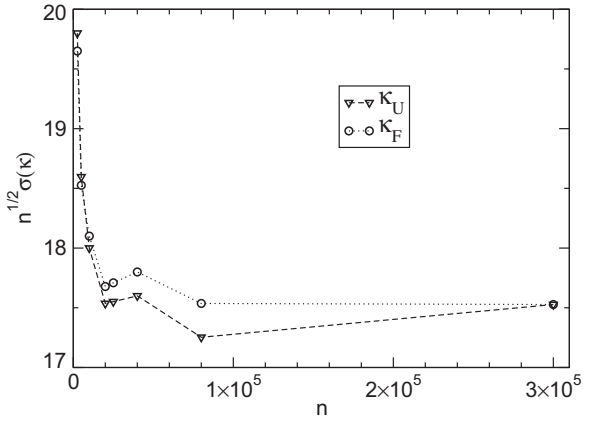


Fig. 10. Effect of the trajectory length n on the standard deviation of scaled estimates of k_U(triangles) and k_F(circles): $\sqrt{n}\sigma(\kappa_U)$ vs n for two-state model (9). Simulation parameters as in Fig. 8.

cence intensities in both states are equal, $I_U = I_F = 50\ \mathrm{ms}^{-1}$. The FRET ratios in the folded and unfolded states are $E_U = 0.61$ and $E_F = 0.93$, respectively.

Fig. 8 presents a boxplot of 10^4 scaled estimates of k_U for different trajectory lengths n. For short trajectories, $n \leqslant 10^4$, the bias is significant. Fig. 8 suggests that the asymptotic regime, where the estimates are well represented by a Gaussian, may begin at about $n = 2 \times 10^4$. Figs. 9a and 9b show the distributions of scaled estimates of k_F and k_U and the correlations between the estimates for trajectories with (a) $n = 2.5 \times 10^3$ and (b) $n = 2 \times 10^4$ photons. The solid lines are Gaussian distributions calculated for the mean value and standard deviation of 10^4 estimates. Note the clear deviations of the distributions of estimates from the predicted Gaussian distributions for short trajectories in 9. Thus longer trajectories (as in 10) are needed to approach the asymptotic regime. Fig. 9b presents a plot of $\sqrt{n}\sigma(\kappa_U)$ vs n, where $\sigma(\kappa_U)$ is the standard deviation of the scaled estimates of k_U. This plot shows that $\sqrt{n}\sigma(\kappa_U)$ approaches a limiting value for large n. Thus, as a rule of thumb, when the trajectory length is increased by a factor of four, the error of the rate constant is halved.

In the present work photon arrival trajectories are parameterized by the rate constants k_U and k_F, and the donor (D) and acceptor (A) fluorescence intensities in the unfolded (U) state, I_U^D and I_U^A, and in the folded (F) state, I_F^D and I_F^A. Chung et al. used a different parametrization in terms of the relaxation rate $k = k_U + k_F$, the population of the folded state $p_F = k_F/k$, and the FRET ratios in the unfolded state $E_U = I_U^A/\left(I_U^D + I_U^A\right)$, and in the folded state $E_F = I_F^A/\left(I_F^D + I_F^A\right)$. Thus from estimates of k_U, k_F, I_U^D, I_U^A I_F^D and I_F^A one can calculate estimates of k, p_F, E_U and E_F. Using this approach for 10^4 trajectories with $n = 3 \times 10^5$ photons, we found the following standard errors for the estimated parameters: $\sigma(\widehat{k}) = 0.023$, $\sigma(\widehat{p}_F) = 0.0092$, $\sigma(\widehat{E}_U) = 0.0014$ and $\sigma(\widehat{E}_F) = 0.00082$, which are consistent with the estimates for 2.25 M of GdmCl in Table 2 of [12]. The estimated biases are negligible compared to magnitude of the standard errors. We also simulated photon trajectories corresponding to the other concentrations of GdmCl analyzed in [12] and obtained analogous results.

Note the differences in the procedures applied here and in [12]. First, we simulate entire trajectories, whereas in [12] a long trajectory is a combination of short ones. Second, we fit both the rates and the intensities for the donor and acceptor channel in each state, whereas in [12] only the rates and FRET ratios are estimated. Third, we estimate the standard deviation by repeating the simulation/estimation process whereas in [12] the errors where estimated from a fit to a single trajectory. Thus our study justifies this approach for long trajectories as used in [12].

Our procedure can also be used when the total fluorescence changes with the state. We simulated photon arrivals for the case when the total (donor and acceptor) fluorescence intensity in one state is different by 20% from the total intensity in the other, and found that the quality of parameter estimates is similar to the case of equal total intensities.

6. Summary and conclusions

In this paper we used the maximum likelihood-based approach to explore in a systematic way parameter estimation and model selection from two-color photon arrival trajectories. We simulated photon arrival trajectories for immobilized single molecules as realizations of Marked Markov Modulated Poisson Processes. From the simulated two-color photon arrival trajectories the fluorescence intensities and kinetic rates were estimated by maximizing the likelihood function (1). In order to select the true model among competing models we used the Akaike and Bayes information criteria (7) and (8).

The present work has several implications for single-molecule data analysis. First, relatively short trajectories can be used for parameter estimation. For instance, for two-state kinetics, a trajectory made up of 5×10^3 photons allows for recovery of the rates with standard deviation lower than 15%. However, for short trajectories the bias and standard deviation of recovered parameter may be difficult to asses from a single photon trajectory, and one may need to resort to Monte Carlo simulations. Second, the method has a better time resolution than the standard on–off analysis. This is because the maximum likelihood analysis of photon arrival trajectories does not involve binning and therefore no information is lost. For sufficiently long trajectories, one can extract the mean dwell times that are comparable to the inter-photon times. Third, the maximum likelihood-based approach can be effectively used not only for two-state models, but also for more complex models, e.g. for 3- and 4-state models with different topologies (linear and cyclic). Fourth, the Akaike and Bayes information criteria are efficient in selecting the true model. Fifth, our results support the usefulness of the maximum likelihood-based approach put forward in [9,12]. Note that the likelihood function (1) can be used

for immobilized molecules when the total (donor and acceptor) fluorescence is different in different states.

Acknowledgment

This work has been supported by a KBN Grant NN204 166736.

References

[1] R. Roy, S. Hohng, T. Ha, Nat. Methods 6 (2008) 507.
[2] M. Blanco, N. Walter, Methods Enzymol. 472 (2010) 153.
[3] E. Barkai, Y. Jung, R. Silbey, Ann. Rev. Phys. Chem. 55 (2004) 457.
[4] M. Lippitz, F. Kulzer, M. Orrit, ChemPhysChem 6 (2005) 770.
[5] E. Barkai, F. Brown, M. Orrit, H. Yang (Eds.), Theory and Evaluations of Single-molecule Signals, World Scientific Publishing Co. Inc., River Edge, NJ, 2008.
[6] S. McKinney, C. Joo, T. Ha, Biophys. J. 91 (2006) 1941.
[7] T-H. Lee, J. Phys. Chem. B 113 (2009) 11535.
[8] S. Jung, R.M. Dickson, J. Phys. Chem. B 113 (2009) 13886.
[9] I.V. Gopich, A. Szabo, J. Phys. Chem. B 113 (2009) 10965.
[10] M. Andrec, R.M. Levy, D.S. Talaga, J. Phys. Chem. A 107 (2003) 7454.
[11] J.E. Bronson, J. Fei, J.M. Hofman, R.L. Gonzalez Jr, C.H. Wiggins, Biophys. J. 97 (2009) 3196.
[12] H.S. Chung, I.V. Gopich, K. McHale, T. Cellmer, J.M. Louis, W.A. Eaton, J. Phys. Chem. A (2010). DOI: 10.1021/jp1009669.
[13] H. Akaike, IEEE Trans. Autom. Control 19 (1974) 716.
[14] G. Schwartz, Ann. Stat. 6 (1978) 461.
[15] M. Hajdziona, A. Molski, J. Chem. Phys. 134 (2011) 054112.
[16] S.C. Kou, X.S. Xie, J.S. Liu, J.R. Stat. Soc. Ser. C 54 (2005) 469.
[17] T. Ryden, Comput. Stat. Data Anal. 21 (1996) 431.
[18] T. Burzykowski, J. Szubiakowski, T. Ryden, Chem. Phys. 288 (2003) 291.
[19] A. Klemm, C. Lindemann, M. Lohmann, Perform. Eval. 54 (2003) 149.