Rate Constants for Slow Conformational Transitions and Their Sampling Errors Using Single-Molecule Fluorescence Spectroscopy

Marián Boguñá, Alexander M. Berezhkovskii,[†] and George H. Weiss*

Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892

Received: October 31, 2000

Single-molecule fluorescence spectroscopy can be used to estimate the rate constants for slow transitions between two states characterized by different lifetimes of a fluorescent probe. The fluorescence decay for this system is a biexponential, the coefficients of each exponential being the fraction of time spent by the system in each of the states. This paper explains how to estimate the rate constants for the transitions and derives the sampling error for the estimates on the assumption that the data are gathered on *M* molecules, each monitored for a time *T*. Results of the analysis indicate that the dominant factor in determining the precision of the rate constant is *M* and that the role of *T* is less significant.

1. Introduction

A variety of techniques, among them single-molecule spectroscopy (SMS),[1−5] can be used to monitor the dynamic behavior of individual molecules in condensed phases. The information available from single-molecule experiments is more detailed than that provided by bulk measurements. One implementation of these ideas is based on measurements of fluorescence of the probes attached to single molecules. It has been applied to the study of conformational changes of DNA,[6−8] as well as to tRNA molecules.[9] In these experiments the molecule interconverts between two states, 1 and 2, with different fluorescence lifetimes of the probe, τ_1 and τ_2, respectively. The interconversion is described by a first-order kinetic scheme:

$$1 \underset{k_2}{\overset{k_1}{\rightleftarrows}} 2 \tag{1.1}$$

where k_1 and k_2 are rate constants whose values are sought. When the fluorescence decay rate is much greater than that of the interconversion, the decay of the fluorescence intensity measured in a bulk experiment is biexponential:

$$\frac{I_{\text{bulk}}(t)}{I_{\text{bulk}}(0)} = P_{\text{eq}}(1)e^{-t/\tau_1} + P_{\text{eq}}(2)e^{-t/\tau_2} \tag{1.2}$$

where $P_{\text{eq}}(i)$ is the probability that a molecule is in state i (= 1, 2) at equilibrium. If we let $k = k_1 + k_2$ these probabilities can be expressed as

$$P_{\text{eq}}(1) = \frac{k_2}{k}, \quad P_{\text{eq}}(2) = 1 - P_{\text{eq}}(1) = \frac{k_1}{k} \tag{1.3}$$

Equation 1.2 shows that the bulk experiment allows one to estimate only the ratio of the rate constants, rather than the rate constants themselves. The single-molecule experiment provides additional information with which one can estimate the individual rate constants. In this experiment a randomly chosen molecule is periodically excited by a train of laser pulses for a time *T*. The fluorescence decay is again described by biexponential with the same decay times as in the bulk experiment, but with random amplitudes:

$$\frac{I_{\text{SM}}(t)}{I_{\text{SM}}(0)} = xe^{-t/\tau_1} + (1-x)e^{-t/\tau_2} \tag{1.4}$$

The amplitude *x* is the fraction of time (out of the total time *T*) that the molecule spent in state 1.[6,10,11] Repetition of the experiment allows one to estimate the probability density for *x*, $p(x|T)$, conditioned on the total monitoring time of a single molecule. The bulk result in eq 1.2 is recovered from the single-molecule experiment in the limit $T \to \infty$, since, because of ergodicity, $\lim_{T\to\infty} p(x|T) = \delta[x - P_{\text{eq}}(1)]$. When *T* is finite, $p(x|T)$ is no longer a delta function and contains the information that can be used to find the individual rate constants.

A general theory of this kind of SM fluorescence experiment has been developed in refs 10−12. The theory developed in these references assumes that the number of molecules, *M*, studied in the experiment, is infinite. Errors in the estimates of k_1 and k_2 arise because *M* is necessarily finite. In this paper we explain how the rate constants can be estimated from the set of random amplitudes found experimentally and calculate the error in these estimates due to the finiteness of *M*, i.e., the sampling error. Our analysis deals only with the sampling error and neglects any other source of errors.

A main result of this paper is a general formula relating the sampling errors to *M* and the measurement time *T*. This formula shows that when the rate constants are not too different, say $(1/3) < k_1/k_2 < 3$, the dependence on *T* will not be significant so long as $kT \geq 10$. This implies that in this circumstance long records would contain redundant information. Nevertheless this additional information can be utilized by cutting the long records into shorter ones, thereby increasing the effective number of molecules without performing further experiments. The same strategy of partitioning the records can also be used when the measurement times for the molecules differ.

2. Rate Constants

Since there are two rate constants, one needs to have two equations. When *T* is finite one can use the first two moments

* Author to whom correspondence should be addressed.

[†] Permanent address: Karpov Institute of Physical Chemistry, 10 Vorontsovo Pole St., 103064 Moscow K-64, Russia.

10.1021/jp004023b CCC: $20.00 © 2001 American Chemical Society
Published on Web 04/26/2001

of x. These are functions of the rate constants and can be found using the results in ref 12. Let $\langle x\rangle$ and $\sigma^2(T)$ be the mean and variance of x. Under the assumption that the system is initially in a state of equilibrium the exact mean and variance of x are[12,13]

$$\langle x\rangle = P_{\mathrm{eq}}(1) = \frac{k_2}{k}, \quad \sigma^2(T) = \frac{2P_{\mathrm{eq}}(1)P_{\mathrm{eq}}(2)}{kT}\left[1 - \frac{1}{kT}(1 - e^{-kT})\right] \tag{2.1}$$

If $\langle x\rangle$ and $\sigma^2(T)$ are known exactly, the rate constants can be expressed in terms of $\langle x\rangle$ and k as

$$k_1 = (1 - \langle x\rangle)k, \quad k_2 = \langle x\rangle k \tag{2.2}$$

where k can be found as the solution to the transcendental equation

$$\frac{\sigma^2(T)}{2\langle x\rangle(1 - \langle x\rangle)}(kT)^2 - kT - e^{-kT} + 1 = 0 \tag{2.3}$$

The last two equations allow one to calculate the two rate constants, which can be expressed as

$$k_i = k_i(\langle x\rangle,\sigma^2(T)),\ i = 1, 2 \tag{2.4}$$

When kT is large enough, say $kT \geq 10$, the last two terms in eq 2.3 can be neglected. In this approximation the rate constants can be related to the mean and variance of x by

$$k_1 \approx \frac{2\langle x\rangle(1 - \langle x\rangle)^2}{T\sigma^2(T)}, \quad k_2 \approx \frac{2\langle x\rangle^2(1 - \langle x\rangle)}{T\sigma^2(T)} \tag{2.5}$$

Equations 2.4 and 2.5 formally relate the rate constants to the mean and variance of x. In practice, neither $\langle x\rangle$ nor $\sigma^2(T)$ is known exactly, but must be estimated from the experimental data. The experimental output consists of a set of amplitudes $\{x_1, x_2, \ldots, x_M\}$, where x_i is the amplitude found from measurements on molecule i. The sampling error in estimating the rate constants is the error incurred by replacing $\langle x\rangle$ and $\sigma^2(T)$, by their estimates, $\bar{x}$ and $\bar{\sigma}^2$. Our strategy will be to first calculate the sampling errors in the estimates of $\langle x\rangle$ and $\sigma^2(T)$ and then to find the sampling error in the estimates of the k_i using the relations in eq 2.4.

3. Sampling Error in Estimates of $\langle x\rangle$ and $\sigma^2(T)$

We assume that all of the x_i are identically distributed independent random variables described by the probability density $p(x|T)$. The standard estimates of the mean and variance are[14]

$$\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i, \quad \bar{\sigma}^2 = \frac{1}{M - 1}\sum_{i=1}^{M}(x_i - \bar{x})^2 \tag{3.1}$$

Because M is finite, $\bar{x}$ and $\bar{\sigma}^2$ are random variables. It is easy to check that

$$\langle\bar{x}\rangle = \langle x\rangle \text{ and } \langle\bar{\sigma}^2\rangle = \sigma^2(T) \tag{3.2}$$

This property of the estimates is usually referred to as unbiasedness.[14]

In our further analysis the estimates $\bar{x}$ and $\bar{\sigma}^2$ will be decomposed into a sum of deterministic and random parts by writing

$$\bar{x} = \langle x\rangle + \delta x,\ \bar{\sigma}^2 = \sigma^2(T) + \delta\sigma^2 \tag{3.3}$$

where, for example, δx is the random component of $\bar{x}$. These definitions ensure that $\langle\delta x\rangle = \langle\delta\sigma^2\rangle = 0$. One can check that the second-order moments of δx and $\delta\sigma^2$ are[15]

$$\mathrm{var}(\bar{x}) = \langle(\delta x)^2\rangle = \frac{1}{M}\langle(x - \langle x\rangle)^2\rangle = \frac{\sigma^2(T)}{M}$$

$$\mathrm{var}(\bar{\sigma}^2) = \langle(\delta\sigma^2)^2\rangle = \frac{1}{M}\{[\langle(x - \langle x\rangle)^4\rangle - [\sigma^2(T)]^2]\} + \frac{2}{M(M - 1)}[\sigma^2(T)]^2$$

$$\mathrm{cov}(\delta x\delta\sigma^2) = \langle\delta x\delta\sigma^2\rangle = \frac{1}{M}\langle(x - \langle x\rangle)^3\rangle \tag{3.4}$$

The important feature of these relations is that all of these averages are $O(M^{-1})$. This result will be used later to show that the estimates of the rate constants are unbiased to $O(M^{-1})$.

All of the moments on the right-hand side of eq 3.4 can be calculated exactly using results derived in ref 13. The exact moments are too complicated to be included here, although they will be used later in their exact form to generate numerical results. In the long-time limit the normalized moments take the form

$$\frac{\mathrm{var}(\bar{x})}{\langle x\rangle^2} \approx \frac{2P_{\mathrm{eq}}(2)}{MP_{\mathrm{eq}}(1)kT}$$

$$\frac{\mathrm{cov}(\delta x\delta\sigma^2)}{\langle x\rangle\sigma^2(T)} \approx \frac{3[P_{\mathrm{eq}}(2) - P_{\mathrm{eq}}(1)]}{MP_{\mathrm{eq}}(1)kT}$$

$$\frac{\mathrm{var}\ (\bar{\sigma}^2)}{[\sigma^2(T)]^2} \approx \frac{2}{M}\left[1 + \frac{3\{1 - 5P_{\mathrm{eq}}(1)P_{\mathrm{eq}}(2)\}}{P_{\mathrm{eq}}(1)P_{\mathrm{eq}}(2)kT}\right] \tag{3.5}$$

where a term that is $O(M^{-2})$ is omitted in the last line. A brief derivation of the moments is given in the Appendix.

4. Sampling Errors in Rate Constant Estimates

Equation 2.4 is an exact relation between the k_i and $\langle x\rangle$ and $\sigma^2(T)$. When $\langle x\rangle$ and $\sigma^2(T)$ in this relation are replaced by their estimates $\bar{x}$ and $\bar{\sigma}^2$, we arrive at an estimate of the rate constants:

$$\bar{k}_i = k_i(\bar{x}, \bar{\sigma}^2) \tag{4.1}$$

These estimates are random variables whose values are close to the exact values in eq 2.4 when M is large enough. To estimate the error in $\bar{k}_i$ we write

$$\bar{k}_i = k_i(\bar{x}, \bar{\sigma}^2) = k_i(\langle x\rangle,\sigma^2(T)) + \delta k_i \tag{4.2}$$

To find an approximation to δk_i we substitute the expressions for $\bar{x}$ and $\bar{\sigma}^2$ in eq 3.3 into $k_i(\bar{x}, \bar{\sigma}^2)$ and expand the result to second order in δx and $\delta\sigma^2$:

$$\delta k_i \approx \frac{\partial k_i}{\partial\langle x\rangle}\delta x + \frac{\partial k_i}{\partial\sigma^2(T)}\delta\sigma^2 + \frac{1}{2}\frac{\partial^2 k_i}{\partial\langle x\rangle^2}(\delta x)^2 + \frac{\partial^2 k_i}{\partial\langle x\rangle\partial\sigma^2(T)}\delta x\delta\sigma^2 + \frac{1}{2}\frac{\partial^2 k_i}{\partial(\sigma^2(T))^2}(\delta\sigma^2)^2 \tag{4.3}$$

Since $\langle\delta x\rangle = \langle\delta\sigma^2\rangle = 0$ it follows that the estimate $\bar{k}_i$ is unbiased up to terms that are $O(M^{-1})$.

On squaring both sides of eq 4.3 we find

$$\mathrm{var}(\bar{k}_i) = \langle(\delta k_i)^2\rangle = \left(\frac{\partial k_i}{\partial\langle x\rangle}\right)^2 \mathrm{var}(\bar{x}) + 2\left(\frac{\partial k_i}{\partial\langle x\rangle}\right)\left(\frac{\partial k_i}{\partial\sigma^2(T)}\right)\mathrm{cov}(\bar{x}, \bar{\sigma}^2) + \left(\frac{\partial k_i}{\partial\sigma^2(T)}\right)^2 \mathrm{var}(\bar{\sigma}^2) + O\left(\frac{1}{M^2}\right) \quad (4.4)$$

The derivatives in this equation can be evaluated exactly by appealing to eq 2.1. Figure 1 shows plots of the relative error $[\mathrm{var}(\bar{k}_i)]^{1/2}/k_i$ as a function of kT to lowest order in M^{-1} for $M = 50$.

In the large-kT limit ($kT \geq 10$) the expressions can be simplified by appealing to eq 2.3. In this limiting case we find, by using eq 3.5, that to lowest order in $1/M$ the relative error in the estimate of the rate constant is

$$\frac{[\mathrm{var}(\bar{k}_i)]^{1/2}}{k_i} \approx \left(\frac{2}{M}\right)^{1/2}\left[1 + \frac{1 - 4P_{eq}(1)P_{eq}(2)}{2P_{eq}(1)P_{eq}(2)kT}\right] = \left(\frac{2}{M}\right)^{1/2}\left[1 + \frac{(K-1)^2}{2KkT}\right] \quad (4.5)$$

independent of i. In this formula $K = k_1/k_2$. The expression in eq 4.5, which is one of the main results of the paper, shows how the relative error depends on the number of molecules, M, and monitoring time, T. We see from this equation that the dominant factor in determining the relative variance of the estimate of k_i is the number of molecules rather than the monitoring time. Any deviation from the condition $k_1 = k_2$ will increase the error at a fixed value of kT. One further result is immediately available. The joint probability density of the random variables $\bar{x}$ and $\bar{\sigma}^2(T)$ is essentially a two-dimensional Gaussian.[14] Since eq 4.3 shows that δk_i is a linear combination of two Gaussian random variables to lowest order in M^{-1}, it follows that $\bar{k}_i$ is also a Gaussian with mean k_i and a variance calculated from eq 4.5. This is confirmed by a plot of simulated data given in Figure 2.

Looking at Figure 1 we see that when the two rate constants are approximately equal the curves will contain a small dip as a function of kT. Otherwise, one sees that the greater the difference between k_1 and k_2, the slower will be the approach to the large kT value $(2/M)^{1/2}$. This effect of asymmetry is in agreement with results shown in Figure 1 of ref 12.

In general, increasing the monitoring time will always improve the precision of the estimate of the normalized variance. However, there is no benefit to be gained by increasing T beyond $kT = 10$ if the value of k_1/k_2 is in the interval (1/3, 3) (see Figure 1). This would seem to suggest that data from measurements made at such long times provide no usable information. However, this is not the case since the data can be used to increase what we will term "virtual molecules". This is done by decomposing the total data set into subsets, each of which can be regarded as being the result of measurements on new virtual molecules. When the rate constants are not too dissimilar the monitoring time for these virtual molecules should be of the order of $kT \approx 10$. On one hand, this is long enough to guarantee the independence of data collected from the virtual molecules as the relaxation function for the kinetic scheme in eq 1.1 is $\exp(-\,kt)$.[13] On the other hand, this value of kT is short enough to make the cutting procedure efficient in increasing the number of virtual molecules. Making use of the data in this way can potentially increase the precision in the estimates.

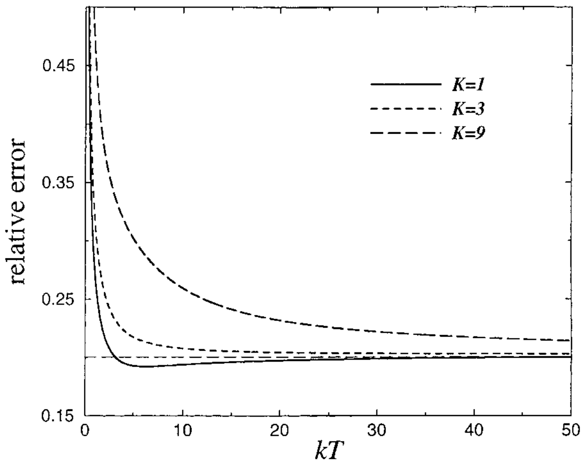

Figure 1. Curves of the normalized standard deviation, $[\mathrm{var}(\bar{k}_i)]^{1/2}/k_i$, as a function of kT for $M = 50$ molecules. The curves shown are for values of $K = k_1/k_2 = 1$, 3, and 9. These illustrate the point that the quickest approach to the limiting value of 0.2 occurs for $K = 1$ and is quite slow when the rates are very asymmetric.

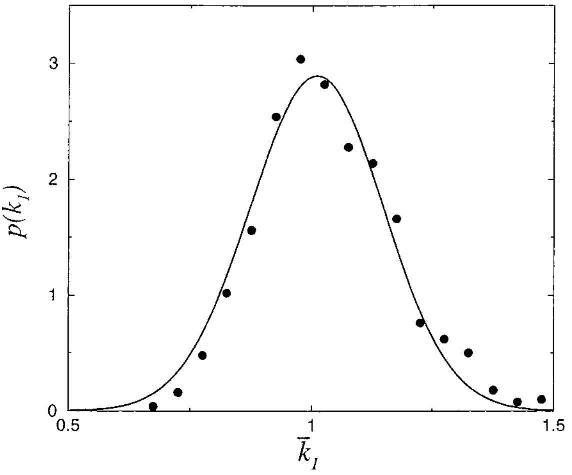

Figure 2. Results for the approximate value of the probability density for the estimate $\bar{k}_i$ generated by the simulation of 1000 experiments. The parameters used to generate the points were $M = 100$ molecules, $k_1 = 1$, $k_2 = 0.3$, and $kT = 10$. The solid line is a Gaussian calculated with the same mean and variance as in the simulated data.

The general idea of decomposing the data set into smaller data sets (after the biexponential form of fluorescence decay with lifetimes identical to those found in the bulk experiment has been established) can also be applied when the monitoring times differ for the different molecules. Reference 16 discusses how to calculate rate constants in this case. Implementing the decomposition strategy not only increases the number of effective molecules, but may also allow one to convert an initial data set with different monitoring times into one with identical monitoring times.

Appendix. Moments of x.

The moments of x can be calculated using the results derived in some detail in ref 13. The formula for $\langle x^n\rangle$ as a function of T can be expressed in terms of a function $g(t)$ defined by

$$g(t) = P_{\mathrm{eq}}(1) + P_{\mathrm{eq}}(2)e^{-kt} \quad (\mathrm{A1})$$

as

$$\langle x^n\rangle = \frac{n!P_{\mathrm{eq}}(1)}{T^n}\int_0^T \mathrm{d}t_1 \int_0^{t_1} \mathrm{d}t_2 \ldots \int_0^{t_{n-1}} \mathrm{d}t_n g(t_1 - t_2) \times g(t_2 - t_3) \ldots g(t_{n-1} - t_n) \quad (\mathrm{A2})$$

A more succinct representation can be found in terms of Laplace transforms as shown in ref 13. Exact expressions for the first four moments can be expressed in terms of $K = k_1/k_2$ as

$$\langle x \rangle = P_{\mathrm{eq}}(1)$$

$$\langle x^2 \rangle = P_{\mathrm{eq}}^2(1)\left[1 + \frac{2K}{kT} - \frac{2K}{(kT)^2}(1 - e^{-kT})\right]$$

$$\langle x^3 \rangle = P_{\mathrm{eq}}^3(1)\left[1 + \frac{6K}{kT} + \frac{6K(K-2)}{(kT)^2} + \frac{12K(1-K)}{(kT)^3} - \frac{6K}{(kT)^2}\left\{K + \frac{2(1-K)}{kT}\right\}e^{-kT}\right]$$

$$\langle x^4 \rangle = P_{\mathrm{eq}}^4(1)\left[1 + \frac{12K}{kT} + \frac{36K(K-1)}{(kT)^2} + \frac{24K(K^2 - 6K + 3)}{(kT)^3} - \frac{12K(6K^2 - 18K + 1)}{(kT)^4} + \frac{12K}{(kT)^2}\left\{K^2 + \frac{2K(2K-3)}{kT} + \frac{6K^2 - 18K + 1}{(kT)^2}\right\}e^{-kT}\right] \quad \text{(A3)}$$

References and Notes

(1) Moerner, W. E., Ed. *Acc. Chem. Res.* **1966**, *29*, 561 (a special issue entirely devoted to SMS).

(2) Nie, S.; Zare, R. N. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 567.

(3) Xie, X. S.; Trautman, J. K. *Annu. Rev. Phys. Chem.* **1998**, *49*, 441.

(4) *Science* **1999**, *283*, No. 5408 (contains a collection of articles on SMS).

(5) *Chem. Phys.* **1999**, *247*, 1 (a special issue entirely devoted to SMS).

(6) Edman, L.; Mets, Ü. Rigler, R. *Exp. Technol. Phys.* **1995**, *41*, 157.

(7) Edman, L.; Mets, Ü. Rigler, R. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 6710.

(8) Wennmalm, S.; Edman, L.; Rigler, R. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10641.

(9) Jia, Y.; Sytnick, A.; Li, L.; Vladimirov, S.; Cooperman, B. S.; Hochstrasser, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 7932.

(10) Geva, E.; Skinner, J. L. *Chem. Phys. Lett.* **1998**, *258*, 225.

(11) Edman, L.; Wennmalm, S.; Tamsen, F.; Rigler, R. *Chem. Phys. Lett.* **1998**, *292*, 15.

(12) Berezhkovskii, A. M.; Szabo, A.; Weiss, G. H. J. *Chem. Phys.* **1999**, *110*, 9145.

(13) Boguñá, M.; Berezhkovskii, A. M.; Weiss, G. H. *Physica A* **2000**, *282*, 475.

(14) Rohatgi, V. K. *Statistical Inference*; John Wiley: New York, 1984.

(15) Stuart, A.; Ord, J. K. *Kendall's Advanced Theory of Statistics, Vol. 1,* 5th ed.; Oxford University Press: New York, 1987; Chapter 10.

(16) Berezhkovskii, A. M.; Boguñá, M.; Weiss, G. H. *Chem. Phys. Lett.* **2001**, *336*, 321, erratum (to appear).