A robust approach to estimating rates from time-correlation functions

John D. Chodera,[1,*] Phillip J. Elms,[2,3,†] William C. Swope,[4,‡] Jan-Hendrik Prinz,[5,§] Susan Marqusee,[6,1,3,¶] Carlos Bustamante,[6,1,7,8,9,**] Frank Noé,[5,††] and Vijay S. Pande[10,‡‡]

[1] *California Institute of Quantitative Biosciences (QB3), University of California, Berkeley, California 94720, USA*
[2] *Biophysics Graduate Group, University of California, Berkeley, California 94720, USA*
[3] *Jason L. Choy Laboratory of Single Molecule Biophysics, University of California, Berkeley, CA 94720, USA*
[4] *IBM Almaden Research Center, San Jose, CA 95120*
[5] *DFG Research Center Matheon, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany*
[6] *Department of Molecular & Cell Biology, University of California, Berkeley, CA 94720, USA*
[7] *Department of Physics, University of California, Berkeley, CA 94720, USA*
[8] *Department of Chemistry, University of California, Berkeley, CA 94720, USA*
[9] *Howard Hughes Medical Institute, University of California, Berkeley, CA 94720, USA*
[10] *Department of Chemistry, Stanford University, Stanford, CA 94305*

(Dated: July 21, 2011)

While seemingly straightforward in principle, the reliable estimation of rate constants is seldom easy in practice. Numerous issues, such as the complication of poor reaction coordinates, cause obvious approaches to yield unreliable estimates. When a reliable order parameter is available, the reactive flux theory of Chandler allows the rate constant to be extracted from the *plateau region* of an appropriate reactive flux correlation function. However, when applied to real data from single-molecule experiments or molecular dynamics simulations, the reactive flux correlation function requires the numerical differentiation of a noisy empirical correlation function, which can result in an unacceptably poor estimate of the rate and pathological dependence on the sampling interval. We present a modified version of this theory which does not require numerical derivatives, allowing rate constants to be robustly estimated from the time-correlation function directly. We illustrate the approach using single-molecule passive force spectroscopy measurements of an RNA hairpin.

Section: *Kinetics, Spectroscopy* or *Statistical Mechanics, Thermodynamics, Medium Effects*

The observed dynamics of complex molecular systems such as biomolecules often suggest a simple underlying behavior. Much of chemistry and biophysics revolves around attempting to identify simple models that adequately describe the observed complex dynamics of these systems. In many cases, stochastic conformational dynamics can be modeled to good accuracy using simple first-order phenomenological rate theory, a topic that has been extensively studied theoretically [1, 2]. However, when it is necessary to estimate rates from trajectories generated by computer simulation or observed in single-molecule experiments, numerous pitfalls can frustrate the ability to extract robust, reliable, and accurate estimates of rate constants using seemingly obvious approaches. Here, we consider a simple approach related to reactive flux theory [3–6], but with much greater robustness to various factors that affect real measurements, such as sampling frequency, finite statistics, and measurement noise.

Suppose we have a population of N noninteracting molecules in solution that can occupy one of two conformational states, denoted A and B. Without loss of generality, we assume we are given a trajectory of some order parameter $x(t)$ that allows us to define associated occupation functions $h_A(t)$ and $h_B(t)$ for states A and B, such that

$$h_A(t) = \begin{cases} 1 & \text{if } x(t) \leq x^\ddagger \\ 0 & \text{if } x(t) > x^\ddagger \end{cases} \quad ; \quad h_B(t) = \begin{cases} 1 & \text{if } x(t) > x^\ddagger \\ 0 & \text{if } x(t) \leq x^\ddagger \end{cases}$$

If there is a separation of timescales between the short relaxation time within the conformational states and the long time the system must wait, on average, in one conformational state before undergoing a transition to another state, the asymptotic relaxation behavior of an initial population of $N_A(0)$ molecules in conformation A and $N_B(0)$ molecules in conformation B can be described by a simple linear rate law:

$$\frac{d}{dt} N_A(t) = -k_{A\to B}\, N_A(t) + k_{B\to A}\, N_B(t) \tag{1}$$

where $k_{A\to B}$ and $k_{B\to A}$ are microscopic rate constants. In terms of time-dependent expectations over trajectories initiated from some initial nonequilibrium state, Eq. 1 is equivalent to

$$\frac{d}{dt} \langle h_A(t) \rangle_{ne} = -k_{A\to B}\, \langle h_A(t) \rangle_{ne} + k_{B\to A}\, \langle h_B(t) \rangle_{ne} \tag{2}$$

where $\langle h_A(t) \rangle_{ne}$ denotes the nonequilibrium probability of finding a given molecule in conformation A at time t given that the fraction of molecules that were initially in conformation A was $\langle h_A(0) \rangle_{ne} = N_A(0)/N$. We hereafter write $h_A(t)$ as shorthand for $h_A(x(t))$.

Were Eq. 2 to govern dynamics at all times, the expected fraction of molecules in conformation A as a function of time

* jchodera@berkeley.edu
† elms@berkeley.edu
‡ swope@us.ibm.com
§ jan.prinz@fu-berlin.de
¶ marqusee@berkeley.edu
** carlos@alice.berkeley.edu
†† frank.noe@fu-berlin.de
‡‡ Corresponding author; pande@stanford.edu

would be given by an exponential decay function

$$\langle h_A(t)\rangle_{ne} = \langle h_A\rangle + [\langle h_A(0)\rangle_{ne} - \langle h_A\rangle]\, e^{-kt} \tag{3}$$

where the quantity $k \equiv k_{A\to B} + k_{B\to A}$ denotes the *phenomenological rate constant* because it is the effective rate that dominates the observed exponential asymptotic relaxation decay behavior. $\langle h_A\rangle$ denotes the standard equilibrium expectation, giving the equilibrium fraction of molecules in conformation A. Note that we do not expect Eq. 3 to hold for short times $t < \tau_{\rm mol}$, where $\tau_{\rm mol}$ is the timescale associated with relaxation processes that damp out *recrossings* that occur due to imperfect definition of the separatrix between the reactant and product states [3, 4].

If the system were purely two-state, a number of naïve approaches to estimation of the rate constant from observed trajectory data would yield useful rate estimates. For example, given an observed trajectory $x(t)$, we could simply compute the number of times n_c the dividing surface $x^\ddagger$ was crossed in either direction in total trajectory time T, estimating the rate as,

$$k_{\rm crossings} \approx \frac{n_c}{T}. \tag{4}$$

Alternatively, we could partition the trajectory into segments in which the system remains in one state, and estimate the *mean lifetime* of these segements, from which the rate is estimated as,

$$k_{\rm lifetime} \approx \tau^{-1}. \tag{5}$$

Both approaches will yield rate estimates that converge to the true rate k as $T \to \infty$ when x provides a perfect reaction coordinate for a perfectly two-state system, in that $x^\ddagger$ correctly divides the two conformations states that interconvert with first-order kinetics.

However, when considering trajectories obtained from computer simulations or single-molecule experiments, these naïve approaches can lead to substantially erroneous estimates. The observed coordinate x might function as a good order parameter, in that it allows the conformational states to be well-resolved at extreme values of x, but a poor reaction coordinate, in that both conformational states are populated in some region near the optimal dividing surface $x^\ddagger$ [7**?**]. The rate estimates from Eqs. 4 and 5 will therefore overestimate the number of crossings or underestimate the state lifetimes, instead converging to the transition state theory rate estimate $k_{\rm TST}$ that gives the instantaneous flux across the dividing surface and overestimating the true rate k. Additionally, if the trajectories are not continuous $x(t)$ but instead consist of discrete observations made with a sampling resolution Δt, additional issues develop. As the sampling interval Δt increases, some crossing of the dividing surface $x^\ddagger$ will be missed, and the perceived lifetimes of states will be increased, having the *opposite* effect of a poor reaction coordinate in *diminishing* the rate estimates of Eqs. 4 and 5. As a result, it can be difficult to predict whether the overall result is an underestimate or overestimate of the true rate k. An example illustrating these effects for a model system where the true rate is known is given in the *Supplementary Material*.

To understand how these pathologies can affect real measurements, we examined the behavior of the p5ab RNA hairpin in an optical trap under passive conditions. This hairpin has been the subject of previous single-molecule force spectroscopy studies [8–10], and exhibits apparent two-state kinetics as the hairpin folds and unfolds under an external biasing force. The force trace $x(t)$ is shown in Fig. 1, and reports on the instantaneous force on the bead along the bead-bead axis; for a harmonic trap, this force is linearly proportional to the displacement of the bead from the center of the trap, and hence the bead-to-bead extension. As the hairpin folds, the bead-to-bead distance contracts, increasing the applied force as the polystyrene bead conjugated to the end of the polymer moves away from the center of the optical trap. At the stationary trap position used for data collection, the hairpin makes many transitions between the two states resolvable from the measured force in the 60-second trajectory, populating each state nearly equally (Fig. 3, top). Data was collected at 50 kHz using a dual-beam counter-propagating optical trap [11, 12], a high sampling rate far above the corner frequency for bead response under these conditions, as previously published [10]. To examine the dependence on sampling interval Δt, the data was also subsampled to 1 kHz, a frequency found to be below the corner frequency of the bead, such that the bead velocity has decorrelated between sequential observations due to hydrodynamic interactions [10].

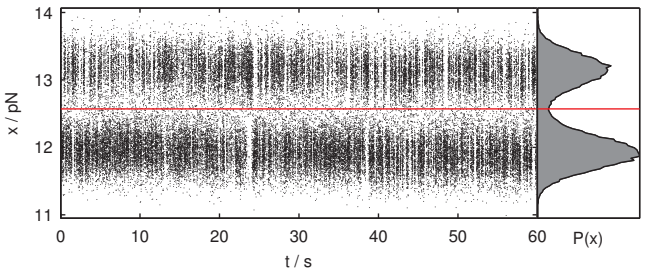

FIG. 1. **Force trace of p5ab RNA hairpin in a stationary optical trap.** A 60-second observation is shown, where the force history $x(t)$ recorded at 50 kHz and subsampled to 1 kHz is plotted. A histogram of the observed force values is shown as $P(x)$ to the right. The red line indicates the optimal dividing surface for rate calculations, $x^\ddagger \approx 12.57$ pN.

The rate constant was estimated using the naïve crossing rate scheme (Eq. 4) as a function of the dividing surface choice $x^\ddagger$, and plotted in Fig. 2 (middle upper and middle lower, dashed lines). Two issues are quickly discerned: First, near the optimal choice of dividing surface ($x^\ddagger \sim 12.57$ pN), the estimated rate $k_{\rm crossing}$ differs greatly depending on whether the 1 kHz data (black dashes, 45.4 s^{-1}) or 50 kHz data (red dashes, 552 s^{-1}) were used to compute the rate estimate. Second, as the dividing surface is perturbed slightly, the rate estimate for either sampling rate changes rapidly. Both properties are highly undesirable, as practical estimators of the rate should yield results insensitive to the sampling rate and exact placement of dividing surface.

As a solution to some of the issue of an imperfect dividing surface, Chandler (and subsequent workers) demonstrated how, despite the lack of a good reaction coordinate, the phenomenological rate could be computed using time-correlation functions through the *reactive flux correlation*

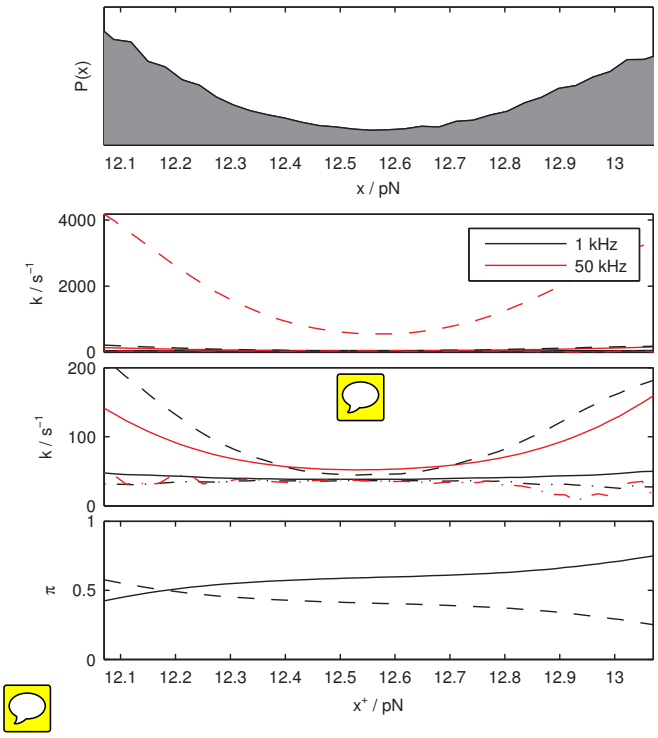

FIG. 2. **Dependence of rate estimates on dividing surface.** *Top:* Histogram of observed forces near transition region between conformational states states. *Upper middle:* Rate estimate from crossing rate k_{crossing} (dashed line), reactive flux rate estimated $k_{\mathrm{RF}}(\tau)$ near plateau time of $\tau = 3$ ms (dash-dotted line), and implied rate $k_{\mathrm{im}}(\tau)$ evaluated at $\tau = 60$ ms (solid line), estimated from 1 kHz data (black) or 50 kHz data (red) as a function of dividing surface $x^{\ddagger}$ choice. *Lower middle:* Same, but zoomed view near true rate constant. *Bottom:* Estimates of equilibrium probabilities π_A (dashed line) and π_B (solid line) estimated from 1 kHz data (black) and 50 kHz data (red) as a function of dividing surface placement $x^{\ddagger}$.

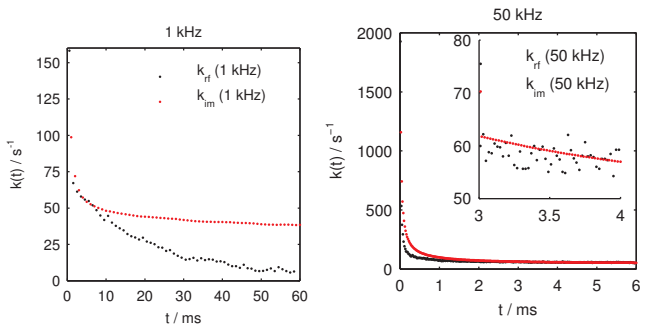

FIG. 3. **Reactive flux correlation function and implied rates from p5ab hairpin single-molecule force trajectory.** The implied rate $k_{\mathrm{im}}(t)$ (red) and reactive flux rate correlation function $k_{\mathrm{RF}}(t)$ (black) are computed for the optimal dividing surface $x^{\ddagger} \approx 12.57$ pN for 1 kHz (*left*) and 50 kHz (*right*). A close-up view compares the rate estimates in the plateau region between 3–4 ms for 50 kHz data (*right inset*).

function $k_{\mathrm{RF}}(t)$ [3–6],

$$k_{\mathrm{RF}}(t) = -\frac{d}{dt}\frac{\langle \delta h_A(0)\, \delta h_A(t)\rangle}{\langle \delta h_A^2\rangle}, \tag{6}$$

where $\delta h_A(t) \equiv h_A(t) - \langle h_A\rangle$ is the instantaneous deviation from the equilibrium population for some trajectory $x(t)$. The reactive flux function $k_{\mathrm{RF}}(t)$ measures the flux across the boundary between A and B that is *reactive*, in the sense that the system has crossed a dividing surface placed between A and B at time zero and is located on the *product* side of the boundary at time t. The reactive flux is bound[illegible]om above by the transition state theory rate estimate k_{TST}, the instantaneous flux across the boundary, because recrossings back to the reactant state will diminish the reactive flux; $k_{\mathrm{RF}}(t)$ becomes identical to k_{TST} as $t \to 0^{+}$ [3]. At t larger than some τ_{mol}—the timescale of relaxation processes *within* the conformational states—thermalization processes will cause the molecule to be captured either in its reactant or product states and remain there for a long time. As a result, the asymptotic rate constant (whose existence requires the presupposed separation of timescales) is only obtained at $\tau_{\mathrm{mol}} < t \ll \tau_{\mathrm{rxn}}$, where $k_{\mathrm{RF}}(t)$ reaches a *plateau value*, decaying to zero at $t \gg \tau_{\mathrm{rxn}}$ with a time constant of τ_{rxn} [3, 4]. Subsequent work extends these concepts to the case of multiple conformational states [5, 6].

The reactive flux correlation function $k_{\mathrm{RF}}(t)$ can, in principle, be used to estimate the phenomenological rate constant k and microscopic rate constants $k_{A\to B}$ and $k_{B\to A}$ from an order parameter timeseries obtained from a computer simulation or single-molecule experiment, but this still presents a number of practical difficulties. For observations recorded at fixed intervals in time Δt, the time derivative of the correlation function (Eq. 6) must be estimated by finite-difference methods, but the presence of statistical error in the estimated correlation function often produces unacceptably large noise in the resulting estimate of $k_{\mathrm{RF}}(t)$. Alternatively, the correlation function $\langle \delta h_A(0)\delta h_A(t)\rangle$ could be smoothed by fitting a polynomial to produce a continuous estimate of the derivative, but this introduces a bias that is difficult to quantify. Additionally, if the reaction timescale τ_{rxn} is not very long compared to the observation interval Δt, then the plateau region where $k_{\mathrm{RF}}(t)$ is identical to the rate may be small and difficult to detect before $k_{\mathrm{RF}}(t)$ decays to zero. Lastly, while alternative expressions to Eq. 6 exist where the *velocity* normal to the separatrix at the time of barrier crossing is utilized instead of a time derivative of the empirical correlation function [3, 4], it is difficult to compute this velocity for complex dividing surfaces in computer simulations, and difficult to measure experimentally in single-molecule experiments.

We computed the reactive flux $k_{\mathrm{RF}}(t)$ from this force trajectory for both 1 kHz and 50 kHz sampling frequencies, using one-sided finite-difference to estimate the derivative in Eq. 6 (Figure 3, black points). When estimated from 50 kHz data (Fig. 3, right), the rate smoothly stabilizes to $\sim 36\ \mathrm{s}^{-1}$ after a transient time of $\tau_{\mathrm{mol}} \approx 3$ ms, but the numerical derivative introduces a great deal of noise into the estimate (Fig. 3, right inset). However, when k_{RF} was estimated from the 1 kHz data (Fig. 3, left), the plateau region near 3–4 ms is relatively narrow and difficult to detect, and the $k_{\mathrm{RF}}(t)$ falls (decaying as ke^{-kt}) as t reaches times comparable to τ_{rxn}. While the reactive flux *does* yield rates that are relatively insensitive to the placement of the dividing surface (Fig. 2, middle upper and middle lower panels, dash-dotted lines), the difficulty of locating the plateau region may make this scheme impracti-

cal for certain sampling frequencies.

We propose an alternative approach, similar in spirit to reactive flux correlation, that avoids the need to compute the time derivative of the correlation function in Eq. 6. Instead, we estimate the rate $k_{\rm im}(t)$ *implied* by the state-to-state transition probabilities observed for a given observation interval t—referring to this quantity as the *implied* rate constant. As with the reactive flux correlation function, for times t where $\tau_{\rm mol} < t \ll \tau_{\rm rxn}$, the phenomenolgical rate constant (if it exists, by virtue of a separation of timescales) is recovered by $k_{\rm im}(t)$, but our modified estimator provides a much larger plateau for times $t > \tau_{\rm mol}$ where a usable rate estimate can be extracted.

As before, if a separation of timescales exists, relaxation behavior for times $t > \tau_{\rm mol}$ is defined in terms of first order rate equations (Eq. 2), here recast in matrix form,

$$\frac{d}{dt}\mathbf{p}(t) = \mathbf{K}\,\mathbf{p}(t) \tag{7}$$

where $\mathbf{p} = [p_A(t)\; p_B(t)]^{\rm T}$ is the vector of equilibrium probabilities, $p_A(t) = \langle h_A(t)\rangle_{ne}$ and $p_B(t) = \langle h_B(t)\rangle_{ne}$ denote the nonequilibrium occupation probabilities of states A and B at time t, and $\mathbf{K}$ is the matrix of rate constants

$$\mathbf{K} = \begin{bmatrix} -k_{A\to B} & k_{B\to A} \\ k_{A\to B} & -k_{B\to A} \end{bmatrix}. \tag{8}$$

The eigenvalues of $\mathbf{K}$ are $\lambda_1 = 0$, reflecting conservation of probability mass, and $\lambda_2 = -(k_{A\to B} + k_{B\to A}) = -k$, which governs the recovery toward equilibrium populations at the phenomenological relaxation rate k.

The solution to Eq. 7 (corresponding to Eq. 3) is given by

$$\mathbf{p}(t) = e^{\mathbf{K}t}\,\mathbf{p}(0) = \mathbf{T}(t)\,\mathbf{p}(0) \tag{9}$$

where $e^{\mathbf{A}} \equiv \sum_{n=0}^{\infty}\mathbf{A}^n/n!$ is the formal matrix exponential and $\mathbf{T}(t)$ can be identified as the column-stochastic *transition probability matrix* whose elements $T_{ji}(t)$ give the conditional probability of observing the system in conformation j at time t given that it was initially in conformation i at time 0.

The elements of $\mathbf{T}(t)$ for a given observation interval t are conveniently given in terms of the time-correlation function

$$T_{ji}(t) \equiv \frac{\langle h_i(0)\,h_j(t)\rangle}{\langle h_i\rangle} \equiv \frac{C_{ij}(t)}{\pi_i} \tag{10}$$

For $t > \tau_{\rm mol}$, we have $\mathbf{T}(t) \approx e^{\mathbf{K}t}$ for a constant matrix $\mathbf{K}$, but this will not hold for $t < \tau_{\rm mol}$. Instead, we can establish a one-to-one correspondence between $\mathbf{T}(t)$ and the rate matrix $\mathbf{K}_{im}(t)$ it *implies* for any t,

$$\mathbf{T}(t) = e^{\mathbf{K}_{im}(t)\,t} \;\Leftrightarrow\; \mathbf{K}_{im}(t) = t^{-1}\log\mathbf{T}(t), \tag{11}$$

where the logarithm denotes the matrix logarithm. For $t > \tau_{mol}$, all $\mathbf{K}_{im}(t) \approx \mathbf{K}$, and the rates are identical to those from reactive flux theory.

Because of their relationship through the exponential (Eq. 11), $\mathbf{T}(t)$ and $\mathbf{K}_{im}(t)$ share the same eigenvectors $\mathbf{u}_k$, and their respective eigenvalues $\mu_k(t)$ and $\lambda_k(t)$ are simply related [13],

$$\mu_k(t) = e^{\lambda_k(t)\,t}. \tag{12}$$

An estimate of the phenomenological rate constant $k_{\rm im}(t)$ for observation time t can be obtained from the second eigenvalue of $\mathbf{K}_{im}(t)$, which for $t > \tau_{\rm mol}$ assumes the value of $-(k_{A\to B} + k_{B\to A}) = -k$,

$$k_{\rm im}(t) = -\lambda_2(t) = -t^{-1}\ln\mu_2(t) \tag{13}$$

where $\mu_2(t)$ can be computed to be,

$$\mu_2(t) = \frac{C_{AA}(t) - \pi_A^2}{\pi_A - \pi_A^2} = \frac{\langle \delta h_A(0)\delta h_A(t)\rangle}{\langle \delta h_A^2\rangle}, \tag{14}$$

which is simply the normalized fluctuation autocorrelation function for the indicator function h_A for state A (or, equivalently, for state B), assuming the value of unity at $t = 0$ and decays to zero at large t.

Combining these expressions gives the expression for the implied rate estimate $k_{\rm im}(t)$,

$$k_{\rm im}(t) = -t^{-1}\ln\frac{\langle \delta h_A(0)\delta h_A(t)\rangle}{\langle \delta h_A^2\rangle} \tag{15}$$

which is the main novel result of this paper.

In the limit $t \to 0^+$, $k_{\rm im}(t)$ reduces to the transition state theory estimate $k_{\rm TST}$. To see this, we expand $C_{AA}(t)$ in terms of its behavior near $t = 0$,

$$\begin{aligned} C_{AA}(t) &= C_{AA}(0) + t\,\dot{C}_{AA}(0) + \mathcal{O}(t^2) \\ &= \pi_A + t\,\dot{C}_{AA}(0) + \mathcal{O}(t^2) \end{aligned} \tag{16}$$

and so,

$$\mu_2(t) = 1 + t\frac{\dot{C}_{AA}(0)}{\pi_A\pi_B} + \mathcal{O}(t^2) \tag{17}$$

Near $t = 0$, $\mu_2(t) \approx 1$, allowing us to expand the argument to the logarithm appearing in $k_{\rm im}(t)$ to first order in t about unity:

$$\lim_{t\to 0^+} k_{\rm im}(t) = \lim_{t\to 0^+} -t^{-1}\ln\mu_2(t) = -\frac{\dot{C}_{AA}(0)}{\pi_A\pi_B} = k_{\rm TST} \tag{18}$$

Similarly, the true phenomenological rate k is given by the long-time limit of $k_{\rm im}(t)$:

$$k = \lim_{t\to\infty} k_{\rm im}(t) = \lim_{t\to\infty} -t^{-1}\ln\frac{\langle \delta h_A(0)\delta h_A(t)\rangle}{\langle \delta h_A^2\rangle} \tag{19}$$

However, when estimating the phenomenological rate through this expression, evaluation of the correlation function should be for some $t \ll \tau_{rxn} = k^{-1}$, as the statistical error in the estimate of $k_{\rm im}(t)$ grows with t (see Appendix).

When there is a separation of timescales such that $\tau_{\rm mol} \ll \tau_{\rm rxn}$, such that a phenomenological rate exists, we can see that $k_{\rm im}(t)$ and $k_{\rm RF}(t)$ are expected to provide similar estimates in the regime $\tau_{\rm mol} < t \ll \tau_{\rm rxn}$. We note Eq. 15 can be rearranged to yield a correlation function

$$\frac{\langle \delta h_A(0)\delta h_A(t)\rangle}{\langle \delta h_A^2\rangle} = e^{-k_{\rm im}(t)\,t} \tag{20}$$

By the definition of reactive flux correlation function (Eq. 6), we can write $k_{\rm RF}(t)$ in terms of $k_{\rm im}(t)$ as,

$$\begin{aligned} k_{\rm RF}(t) &= -\frac{d}{dt}\frac{\langle \delta h_A(0)\delta h_A(t)\rangle}{\langle \delta h_A^2\rangle} = -\frac{d}{dt}e^{-k_{\rm im}(t)\,t} \\ &= e^{-k_{\rm im}(t)\,t}\left[k_{\rm im}(t) + t\frac{d}{dt}k_{\rm im}(t)\right] \end{aligned} \tag{21}$$

When $t \gg \tau_{\rm mol}$, then $k_{\rm im}(t) \approx k$, and we have $k_{\rm RF}(t) \approx k e^{-k\,t}$.

To illustrate the estimation of the phenomenological rate k using the implied timescale $k_{\rm im}(t)$, we computed it for the p5ab hairpin force trajectory described above. At the 50 kHz sampling rate (Fig. 3, right), the rate estimates are almost

identical to those from $k_{\mathrm{RF}}(t)$ for a broad range of times where $t > \tau_{\mathrm{mol}}$, though there is much less noise in the $k_{\mathrm{im}}(t)$ rate estimate than in $k_{\mathrm{RF}}(t)$ (Fig. 3, right inset). At the 1 kHz sampling rate (Fig. 3, left), however, the rate estimate from $k_{\mathrm{im}}(t)$ remains stable over several times τ_{rxn}, even though the $k_{\mathrm{RF}}(t)$ has already decayed from the plateau region. The implied rate estimate, $k_{\mathrm{im}}(t)$, therefore appears to provide a more robust estimate of the phenomenological rate under a variety of conditions.

This robustness also carries over to an insensitivity to the placement of dividing surface $x^{\ddagger}$, the problem reactive flux theory was originally envisioned to solve. Using an observation time of $\tau = 60$ ms, the implied rate estimate $k_{\mathrm{im}}(\tau)$ varies only a few percent over a wide range near the boundaries between states (Fig. 2, lower middle), which is striking compared to the large range over which the estimate from Eq. 4 varies in the same region (Fig. 2, upper middle).

To obtain individual microscopic rates $k_{A\to B}$ and $k_{B\to A}$, we recall that the phenomenological rate k represents the sum of the forward and backward rates,

$$k = k_{A\to B} + k_{B\to A} \tag{22}$$

as well as the fact that the flux across the dividing surface must be balanced at equilibrium,

$$\pi_A k_{A\to B} = \pi_B k_{B\to A} \tag{23}$$

which allows us to deduce that the individual rates are simply

$$k_{A\to B} = \pi_B\, k \;\; ; \;\; k_{B\to A} = \pi_A\, k \tag{24}$$

A difficulty immediately appears—the estimate of π_A and π_B can be quite sensitive to the choice of dividing surface between states A and B (Fig. 2, bottom). As a result, the individual rates $k_{A\to B}$ and $k_{B\to A}$ will also be sensitive to the choice of dividing surface, even though the phenomenological rate k may not be; this problem appears unavoidabl

ACKNOWLEDGMENTS

The authors would like to thank Ken Dill (University of California, San Francisco), Phillip L. Geissler (University of California, Berkeley), and Jed W. Pitera (IBM Almaden Research Center) for stimulating discussions on this topic. PJE would like to thank Steve Smith (University of California, Berkeley) for help with the instrumentation and Jin Der Wen (National Taiwan University) and Ignacio Tinoco (University of California, Berkeley) for providing the p5ab RNA hairpin. This work was supported in part by NIH grants GM 32543 (C.B.), GM 50945 (S.M.) and a grant from the NSF (S.M.). JDC gratefully acknowledges support from the HHMI and IBM predoctoral fellowship programs, NIH grant GM34993 through Ken A. Dill (UCSF), and NSF grant for Cyberinfrastructure (NSF CHE-0535616) through Vijay S. Pande (Stanford), and a QB3-Berkeley Distinguished Postdoctoral Fellowship at various points throughout this work. FN acknowledges support from DFG Research Center Matheon.

[1] P. Hänggi, P. Talkner, and M. Borkovec, Rev. Mod. Phys., **62**, 251 (1990).
[2] H.-X. Zhou, Quarterly Rev. Biophys., **43**, 219 (2010).
[3] D. Chandler, J. Chem. Phys., **68**, 2959 (1978).
[4] J. A. M. Jr., D. Chandler, and B. J. Berne, J. Chem. Phys., **70**, 4056 (1979).
[5] J. E. Adams and J. D. Doll, Surface Science, **111**, 492 (1981).
[6] A. F. Voter and J. D. Doll, J. Chem. Phys., **82**, 80 (1985).
[7] J. D. Chodera and V. S. Pande, Submitted (2011).
[8] J. Liphardt, B. Onoa, S. B. Smith, I. Tinoco Jr., and C. Bustamante, Science, **292**, 733 (2001).
[9] J.-D. Wen, M. Manosas, P. T. X. Li, S. B. Smith, C. Bustamante, F. Ritort, and I. Tinoco, Jr., Biophys. J., **92**, 2996 (2007).
[10] P. J. Elms, J. D. Chodera, C. J. Bustamante, and S. Marqusee, In preparation (2010).
[11] S. B. Smith, Y. Cui, and C. Bustamante, Meth. Enzym., **361**, 134 (2003).
[12] C. Bustamante and S. B. Smith, "Light-force sensor and method for measuring axial optical-trap forces from changes in light momentum along an optical axis," (2006), united States Patent 7133132.
[13] N.-V. Buchete and G. Hummer, J. Phys. Chem. B, **112**, 6057 (2008).
[14] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok, and K. A. Dill, J. Chem. Theor. Comput., **3**, 26 (2007).

Appendix A: Statistical error of implied rate estimate

[JDC: This might be moved to *Supplementary Material*.]

To estimate the statistical error in the implied rate $k_{\mathrm{im}}(t)$, we first define two trajectory functionals:

$$\begin{aligned} \mathcal{F}[X] &\equiv h_A(x_0) \\ \mathcal{G}[X] &\equiv h_A(x_0)\, h_B(x_\tau) \end{aligned} \tag{A1}$$

We define a timeseries F_n and G_n obtained from evaluating these functionals on sequential (potentially overlapping) segments X_n of a much longer trajectory, or multiple independent short trajectories, depending on what kind of data has been collected,

$$\begin{aligned} F_n &= \mathcal{F}[X_n] \\ G_n &= \mathcal{G}[X_n] \end{aligned} \tag{A2}$$

Estimates of their expectations are given by the sample means:

$$\begin{aligned} \pi_A \approx \hat{F} &= \frac{1}{N}\sum_{n=1}^{N} F_n \\ C_{AB}(t) \approx \hat{G} &= \frac{1}{N}\sum_{n=1}^{N} G_n \end{aligned} \tag{A3}$$

We compute the rate constant for a given observation time t (whose functional dependence we shall suppress) from an

estimate of the second eigenvalue $\hat{\mu}_2$:

$$\hat{k}_{im} = -t^{-1} \ln \hat{\mu}_2 \tag{A4}$$

The second eigenvalue is estimated from Eq. 14:

$$\begin{aligned} \mu_2 &= 1 - \frac{C_{AB}(t)}{\pi_A(1-\pi_A)} \\ \Leftrightarrow \quad \hat{\mu}_2 &= 1 - \frac{\hat{G}}{\hat{F}(1-\hat{F})} \end{aligned} \tag{A5}$$

The variance of the estimate $\hat{k}_{im}(t)$ can be estimated by simple first-order Taylor series expansion propagation of error,

$$\delta^2 \hat{k}_{im} = \left[\frac{\partial \hat{k}_{im}}{\partial \hat{\mu}_2}\right]^2 \delta^2 \hat{\mu}_2 = \frac{\delta^2 \hat{\mu}_2}{t^2\, \hat{\mu}_2^2} \tag{A6}$$

We apply the first-order Taylor series expansion propagation of error to compute the uncertainty in $\delta^2 \hat{\mu}_2$:

$$\delta^2 \hat{\mu}_2 = \left[\frac{\partial \hat{\mu}_2}{\partial \hat{F}}\right]^2 \delta^2 \hat{F} + \left[\frac{\partial \hat{\mu}_2}{\partial \hat{G}}\right]^2 \delta^2 \hat{G} + 2 \left[\frac{\partial \hat{\mu}_2}{\partial \hat{F}}\right] \left[\frac{\partial \hat{\mu}_2}{\partial \hat{G}}\right] \delta \hat{F} \delta \hat{G} \tag{A7}$$

where the required derivatives are given by

$$\frac{\partial \hat{\mu}_2}{\partial \hat{F}} = \frac{\hat{G}(1-2\hat{F})}{\hat{F}^2(1-\hat{F})^2} \; ; \; \frac{\partial \hat{\mu}_2}{\partial \hat{G}} = \frac{1}{\hat{F}(1-\hat{F})} \tag{A8}$$

To estimate $\delta^2 \hat{\mu}_2$, we must estimate the variance and covariance of the estimators $\hat{F}$ and $\hat{G}$:

$$\delta^2 \hat{F} = \frac{\mathrm{var}\, F_n}{N/g} \; ; \; \delta^2 \hat{G} = \frac{\mathrm{var}\, G_n}{N/g} \; ; \; \delta \hat{F} \delta \hat{G} = \frac{\mathrm{cov}\,(F_n, G_n)}{N/g} \tag{A9}$$

where the sample covariances are used to estimate $\mathrm{var}\, A_n$, $\mathrm{var}\, B_n$, and $\mathrm{cov}\,(A_n, B_n)$, and g is the maximum statistical inefficiency for the timeseries F_n and G_n. If the X_n denote independent short trajectories (such as obtained by a computer simulation from uncorrelated initial starting configurations x_0), then $g = 1$; otherwise, the statistical inefficiency g can be estimated via standard means (see Section 5.2 of Chodera *et al.* [14]).

Practically, we take advantage of time-reversibility of dynamics, and use a slightly modified set of trajectory functionals that yield the same expectation but average over more snapshots from the trajectory in the case that the trajectory segments are of length $T > t$:

$$\begin{aligned} \mathcal{G}[X] &= \frac{1}{T-t} \sum_{t_0=0}^{T-\tau} \frac{1}{2} \left[h_A(x_0)\, h_B(x_{t_0+t}) + h_B(x_{t_0})\, h_A(x_{t_0+t})\right] \\ \mathcal{F}[X] &= \frac{1}{T-t} \sum_{t_0=0}^{T-t} \frac{1}{2} \left[h_A(x_0) + h_A(x_{t_0+t})\right] \end{aligned} \tag{A10}$$

This appropriately accounts for the fact that all time origins produce equally valid estimates and, for systems with multiple conformational states, ensures satisfaction of detailed balance.

Appendix B: Model system

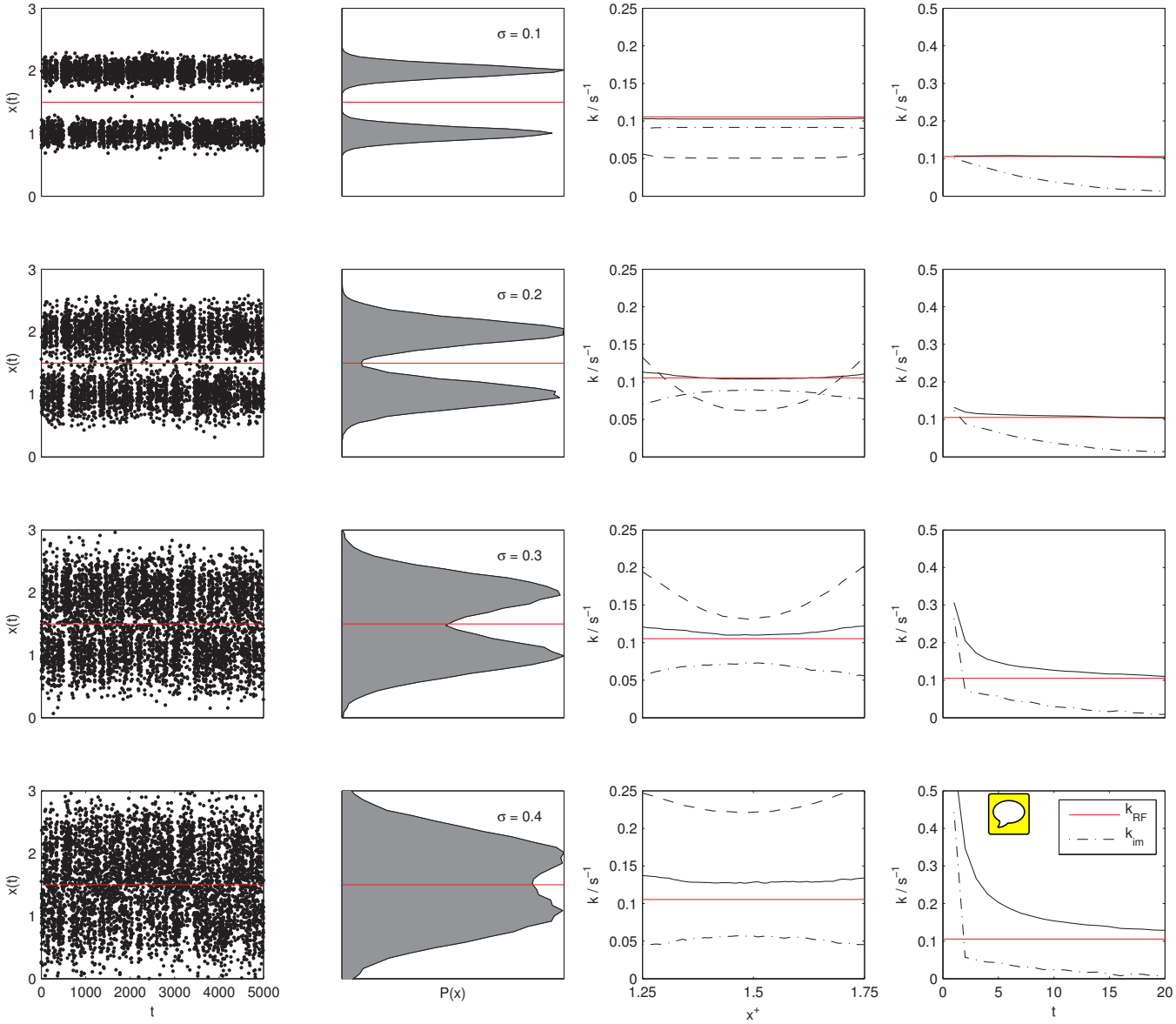

FIG. 4. **Rate estimates for model two-state system.** A two-state discrete state trajectory between two states centered on $x = 1$ and $x = 2$ was generated, on top of which was superimposed Gaussian noise with standard deviation σ of 0.1 (top), 0.2 (upper middle), 0.3 (lower middle), or 0.4 (bottom). *Left:* A portion of the resulting observed trajectory $x(t)$ is shown, with the optimal dividing surface $x^{\ddagger} = 1.5$ drawn in red. *Left middle:* Histogram of the 50 000 observation trajectory. *Right middle:* Rate estimate from crossing rate k_{crossing} (dashed line), reactive flux rate estimated $k_{\mathrm{RF}}(\tau)$ near plateau time of $\tau = 2$ (dash-dotted line), and implied rate $k_{\mathrm{im}}(\tau)$ evaluated at $\tau = 20$ (solid line), estimated as a function of dividing surface $x^{\ddagger}$ choice the the vicinity of the transition region. *Right:* Rate estimates from crossing rate k_{crossing} (dashed line), reactive flux rate estimated $k_{\mathrm{RF}}(\tau)$ (dash-dotted line), and implied rate $k_{\mathrm{im}}(\tau)$ (solid line), evaluated at optimal dividing surface $x^{\ddagger} = 1.5$.