Important Notice to Authors

Attached is a PDF proof of your forthcoming article in *Physical Review Letters*. The accession code is LE12962.

Your paper will be in the following section of the journal: Soft Matter, Biological, and Interdisciplinary Physics

Figures submitted electronically as separate PostScript files containing color usually appear in color in the online journal. However, all figures will appear as grayscale images in the printed journal unless the color figure authorization form has been received and you have agreed to pay the necessary charges. For figures that will be color online but grayscale in print, please insure that the text and caption clearly describe the figure to readers who view it only in black and white.

No further publication processing will occur until we receive your response to this proof.

Questions & Comments to Address

The numbered items below correspond to numbers in the margin of the proof pages pinpointing the source of the question and/or comment. The numbers will be removed from the margins prior to publication.

1. A check of online databases revealed a possible error in Ref. [1]. The page number has been changed from ’941’ to ’943’. Please confirm this is correct.
2. NOTE: External links, which appear as blue text in the reference section, are created for any reference where a Digital Object Identifier (DOI) can be found. Please confirm that the links created in this PDF proof, which can be checked by clicking on the blue text, direct the reader to the correct references online. If there is an error, correct the information in the reference or supply the correct DOI for the reference. If no correction can be made or the correct DOI cannot be supplied, the link will be removed.
3. In Ref. [3], please note the change from I. T., Jr. to I. Tinoco according to the published reference. Is this correct?
4. In Ref. [6], please provide the name of the publisher and the city and year of publication. If there are editors for this book, please provide those names as well.
5. This query was generated by an automatic reference checking system. Reference [10] could not be located in the databases used by the system. While the reference may be correct, we ask that you check it so we can provide as many links to the referenced articles as possible.
6. A check of online databases revealed a possible error in Ref. [15]. The date has been changed from ’2010’ to ’2009’. Please confirm this is correct.
7. The references have been reordered so that they are cited in the text in numerical order.
8. Please provide a brief description of the Supplemental Material to be included in Ref. [25].
9. The website for the MATLAB code has been moved to the reference section per PRL style.
10. Please note the change to outer curly braces and square brackets in Eq. (6) to avoid nested fences of the same size and type per PRL style. Is this change acceptable?
11. I do not see a red line in Fig. 1. Please verify.
12. In Fig. 3, pmf has been changed to PMF as used in text. Is this correct?

Other Items to Check

- Please check your title, author list, receipt date, and PACS numbers. More information on PACS numbers is available online at http://publish.aps.org/PACS/.
- Please proofread the article very carefully.
- Please check that your figures are accurate and sized properly. Figure quality in this proof is representative of the quality to be used in the online journal. To achieve manageable file size for online delivery, some compression and downsampling of figures may have occurred. Fine details may have become somewhat fuzzy, especially in color figures. The print journal uses files of higher resolution and therefore details may be sharper in print. Figures to be published in color online will appear in color on these proofs if viewed on a color monitor or printed on a color printer.

Ways to Respond

- ***Web:*** If you accessed this proof online, follow the instructions on the web page to submit corrections.
- ***Email:*** Send corrections to apsproofs@beacon.com. Include the accession code LE12962 in the subject line.
- ***Fax:*** Return this proof with corrections to +1.419.289.8923.

If You Need to Call Us

You may leave a voicemail message at +1.419.289.0558 ext. 133. Please reference the accession code and the first author of your article in your voicemail message. We will respond to you via email.

Splitting Probabilities as a Test of Reaction Coordinate Choice in Single-Molecule Experiments

John D. Chodera[1,*] and Vijay S. Pande[2,†]

[1]*California Institute of Quantitative Biosciences (QB3), University of California, Berkeley, California 94720, USA*
[2]*Department of Chemistry, Stanford University, Stanford, California 94305, USA*

(Received 4 May 2011)

To explain the observed dynamics in equilibrium single-molecule measurements of biomolecules, the experimental observable is often chosen as a putative reaction coordinate along which kinetic behavior is presumed to be governed by diffusive dynamics. Here, we invoke the splitting probability as a test of the suitability of such a proposed reaction coordinate. Comparison of the observed splitting probability with that computed from the kinetic model provides a simple test to reject poor reaction coordinates. We demonstrate this test for a force spectroscopy measurement of a DNA hairpin.

DOI: PACS numbers: 87.15.Cc, 87.10.Mn, 87.64.Dz, 87.64.kv

A variety of new experimental techniques have made it possible to monitor the conformational fluctuations of single biological macromolecules under both equilibrium and nonequilibrium conditions. These experiments aim to probe the statistical dynamics and conformational substates relevant to folding and function. In a typical experiment, such as observation of the resonant energy transfer efficiency between two fluorophores incorporated into an RNA molecule [1], fluctuations of a spectroscopic observable in the absence of an external field are monitored. Other experiments allow the effect of an external biasing potential on the dynamics to be observed, as in an optical trap [2–4].

To describe the observed dynamics of the system, it is tempting to identify the observable with a reaction coordinate and construct a model in which the dynamics evolves by a diffusion process in an effective potential, such as by overdamped Langevin (also called ''Brownian'') dynamics [5],

$$\dot{x}(t) = -\beta \frac{\partial}{\partial x} F(x) + \sqrt{2D(x)} R(t). \tag{1}$$

Here, $x(t)$ is the time-dependent motion along the resolved coordinate, $D(x)$ is the diffusion constant (often assumed to be a constant independent of x), $\beta \equiv (k_B T)^{-1}$ is the inverse temperature, $F(x) \equiv -k_B T \ln \pi(x)$ is the potential of mean force (PMF) defined in terms of the observed equilibrium probability density $\pi(x)$, and $R(t)$ is a Gaussian process with zero mean satisfying $\langle R(t)R(t')\rangle = \delta(t - t')$.

Many physical systems such as biomolecules exhibit strong metastabilities in the conformational degrees of freedom, resulting in the presence of two or more discrete conformational states in which the system remains for a long time before transitioning to another metastable state [6]. While it is often easy to find an observable x that is a suitable order parameter that allows these metastable states to be discriminated to some degree, it is generally difficult to find a good reaction coordinate so that dynamics along the resolved coordinate are well described by Eq. (1).

For data collected in a given single-molecule experiment, how can we determine whether the resolved coordinate provides a good reaction coordinate? Recent work on tests of reaction coordinate suitability in computer simulations has focused on the calculation of the committor or splitting probabilities, a concept dating back to Onsager [7]. This quantity, now extensively used in simulation studies of protein folding [8], represents the probability that a trajectory first encounters one absorbing boundary placed along the reaction coordinate before another, given an initial microscopic state of the system. For suitable choices of reaction coordinate, the distribution of committor probabilities along an equilibrium ensemble of configurations restricted to a given value of the reaction coordinate will be closely grouped about a characteristic value [8–13]; indeed, ideal reaction coordinates organize committor isosurfaces in an ordered fashion along the reaction coordinate [14].

Unfortunately, tests based on evaluating distributions of committor values along cuts of the putative reaction coordinate are impossible to apply in a physical experiment, since there is no way to prepare the system in precisely the same microscopic configuration to probe the statistics of committor probabilities. Instead, we propose a simple alternative that is readily computable from observed equilibrium trajectories of the resolved coordinate: comparison of the average committor along each value of the reaction coordinate evaluated by Eq. (1) with the empirical average committor from the observed trajectory.

Theory.—Consider the placement of absorbing boundaries at a and b near the periphery of the observed range of the resolved coordinate x. For a diffusion process in one dimension governed by Eq. (1), the probability of first encountering a before b starting from $x \in [a, b]$ can be shown to be [5,14,15]

$$p_A(x) = \frac{\int_x^b dx' D(x')^{-1} e^{\beta F(x')}}{\int_a^b dx' D(x')^{-1} e^{\beta F(x')}}. \tag{2}$$

The PMF along the resolved coordinate x, $F(x)$, can be estimated from a single-molecule trajectory of sufficient

length [4,16] or from multiple trajectories under different equilibrium [17] or nonequilibrium [18,19] conditions. The diffusion profile $D(x)$ can be estimated in a number of ways (such as the Bayesian scheme of Best and Hummer that allows simultaneous computation of both PMF and diffusion constant [15]), though it is commonly assumed to be constant, in which case it cancels from both numerator and denominator.

An empirical estimate of the splitting probability $\hat{p}_A(x)$ can also be computed directly from an observed equilibrium trajectory $x(t)$, $t \in [0, \mathcal{T}]$ (also recently noted in Ref. [20]),

$$\hat{p}_A(y) = \frac{\int_0^{\mathcal{T}} dt \delta(y - x(t)) c_A(t)}{\int_0^{\mathcal{T}} dt \delta(y - x(t))}, \tag{3}$$

where we have defined the hitting function $c_A(t)$ in terms of $x(t)$ as

$$c_A(t) = \begin{cases} 1 & \text{if } \tau_A(t) < \tau_B(t) \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where auxiliary functions $\tau_A(t)$ and $\tau_B(t)$ are defined as

$$\begin{aligned} \tau_A(t) &= \inf\{t' > t : x(t) < a\}, \\ \tau_B(t) &= \inf\{t' > t : x(t) > b\}. \end{aligned} \tag{5}$$

The hitting function $c_A(t)$ simply keeps track of whether $x(t)$ will hit boundary a before b immediately following time t, and assumes the value of unity if so, and zero otherwise. In practice, the delta function $\delta(y - x(t))$ is replaced by some kernel function of finite width, such as a histogram bin. An estimate of $\hat{p}_A(x)$ from multiple equilibrium trajectories can be produced by averaging the trajectories weighted by their lengths.

Our proposed test is simple: By comparing the splitting probability estimated from the PMF, $p_A(x)$, with the empirical estimate of the splitting probability from the trajectory, $\hat{p}_A(x)$, we can judge whether these quantities are obviously discrepant over the range $x \in [a, b]$, which would indicate that x is a poor reaction coordinate. Note that this test is necessary, but not sufficient, for x to be a good reaction coordinate; agreement does not mean that the putative reaction coordinate is a true reaction coordinate. Nevertheless, the test may be sufficiently exacting to reject poor choices of reaction coordinate that are not immediately obvious by eye yet fail this comparison.

When the observed coordinate is determined to be a poor reaction coordinate, the consequences of assuming it to be an adequate reaction coordinate depend on the precise nature of the information extracted from the single-molecule data. The consequences could be as simple as underestimating the rate constant for a two-state process or as subtle as inferring an erroneous mechanism for more complex processes. The most obvious consequence is mistaking the location of the transition state—the point where the splitting probability $p_A = 0.5$—to be displaced from the free energy barrier in the potential of mean force. For systems like DNA hairpins and proteins, this can have consequences for the interpretation of how "brittle" or "compliant" the conformational states are perceived to be. Notably, similar tests have been found to be useful in validating putative reaction coordinate choices in computer simulations, despite the ability to inspect the atomic coordinates directly [21,22].

Model system.—As an illustrative example, we consider the two-dimensional model system previously studied by Rhee and Pande [14], 10

$$\begin{aligned} U(x, y) = {} & [1 - 0.5 \tanh(y - x)](x + y - 5)^2 \\ & + 0.2\{[(y - x)^2 - 9]^2 + 3(y - x)\} \\ & + 15 e^{-(x-2.5)^2-(y-2.5)^2} - 20 e^{-(x-4)^2-(y-4)^2}, \end{aligned} \tag{6}$$

pictured here in the upper-left hand panel of Fig. 1. Two stable states are present, located roughly at (x, y) coordinates (4, 1) and (1, 4). At $k_B T = 5$, the PMFs along both x and y clearly show two distinct wells separated by a barrier, and yet these coordinates are expected to be poor reaction coordinates individually; the coordinate $q = (x - y)/\sqrt{2}$ (Fig. 1, upper-left hand panel, red line), however, which 11 connects the two stable basins more directly, is known to be a good reaction coordinate at this temperature [14].

A Brownian dynamics trajectory of 10^6 steps was generated using the discretization of Eq. (1) by Ermak

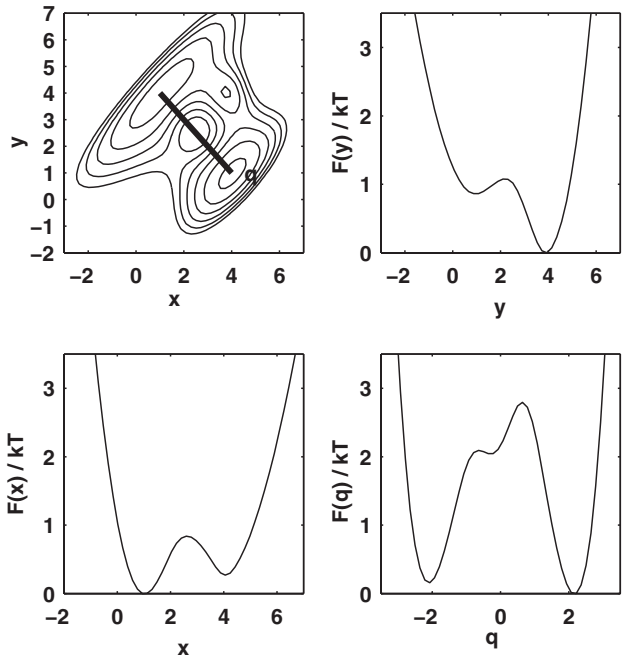

FIG. 1. Two-dimensional model system and potentials of mean force. Upper left: Potential for the two-dimensional model system, with contours drawn every 5 $k_B T$. x and y are poor reaction coordinates, while $q = (x - y)/\sqrt{2}$ (thick black line) is a good reaction coordinate. Other panels: Potentials of mean force in units of $k_B T$ for projections onto x, y, and q.

2

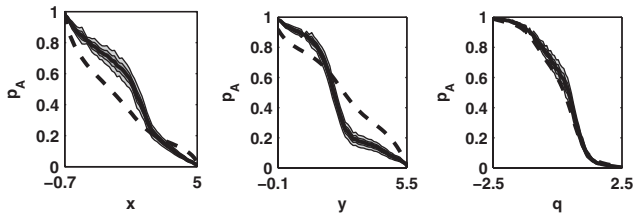

FIG. 2. Splitting probability tests for two-dimensional model system. For each choice of projected coordinate shown in Fig. 1, both the trajectory-derived empirical splitting probability $\hat{p}_A$ (solid black line) and the PMF-derived splitting probability p_A (dashed black line) are shown. Dark shading represents a 68% confidence interval about $\hat{p}_A$, and light shading a 95% confidence interval.

and Yeh [23,24], with a diffusion constant of $D = 1$ and time step $\Delta t = 0.1$. This trajectory was projected onto either poor choices of reaction coordinate x and y or good reaction coordinate q (see Fig. 1 in the Supplemental Material [25]). For each projection, the potential of mean force was estimated from an empirical histogram, e.g., $F(x) \approx -k_B T \ln p(x)$ for the projection onto x, using 100 equally sized bins. The PMF-derived splitting probability $p_A(x)$ was computed from $F(x)$ using Eq. (2), and the empirical splitting probability $\hat{p}_A(x)$ according to Eq. (3). To judge whether disagreement between these estimates was statistically meaningful, the statistical uncertainty in the empirical $\hat{p}_A(x)$ was estimated by using time-correlation analysis (see Supplemental Material [25]).

The results of this comparison assuming a uniform diffusion constant are shown in Fig. 2. The poor suitability of x and y as reaction coordinates is easily seen by the large discrepancy between the splitting probability p_A computed from the PMF (dashed line) and the empirical splitting probability $\hat{p}_A$ estimated from the trajectories (solid line). However, the coordinate $q = (x - y)/\sqrt{2}$, previously identified by Rhee and Pande as being well aligned with the true reaction coordinate at this temperature by sophisticated means not available to single-molecule experiments [14], agrees to within statistical error (shaded region).

DNA hairpin force spectroscopy.—To demonstrate the utility of our proposed splitting probability test in a real laboratory measurement, we performed the same analysis on a single-molecule trajectory of a DNA hairpin in a double optical trap, previously reported by Woodside *et al.* [4]. The hairpin, referred to as 30R50/T4 due to the content of a 30 bp stem-forming sequence, is attached by means of dsDNA handles to two polystyrene beads held in a passive all-optical constant-force clamp [2] at an external force that encourages hopping among closed and open conformations over the course of the experiment. Bead displacements in the trap were recorded with a sampling frequency of 25 kHz [4], and the bead-to-bead extension trajectory was analyzed here.

Figure 3 shows the observed trajectory of the molecular extension coordinate and corresponding splitting probability analysis for a uniform diffusion constant. From this analysis, it is evident there is poor agreement between $p_A(x)$ estimated from the PMF and the empirical $\hat{p}_A(x)$ estimated from the trajectory in the region of extensions between 535 and 545 nm. This suggests that, at this external force, dynamics would be poorly described by Brownian dynamics along the total molecular extension coordinate using Eq. (1) and a uniform diffusion constant.

Nonuniform diffusion.—Recently, it has been suggested that nonuniformity of the diffusion constant along the resolved coordinate may have important ramifications for single-molecule biophysical experiments [15]. Could strong position dependence of the diffusion constant $D(x)$ be responsible for the observed discrepancy in Fig. 3? To judge whether nonuniform diffusion significantly impacted our test of reaction coordinate suitability, we used the Bayesian inference scheme proposed by Best and Hummer [15] to simultaneously compute position-dependent diffusion constant $D(x)$ and potential of mean

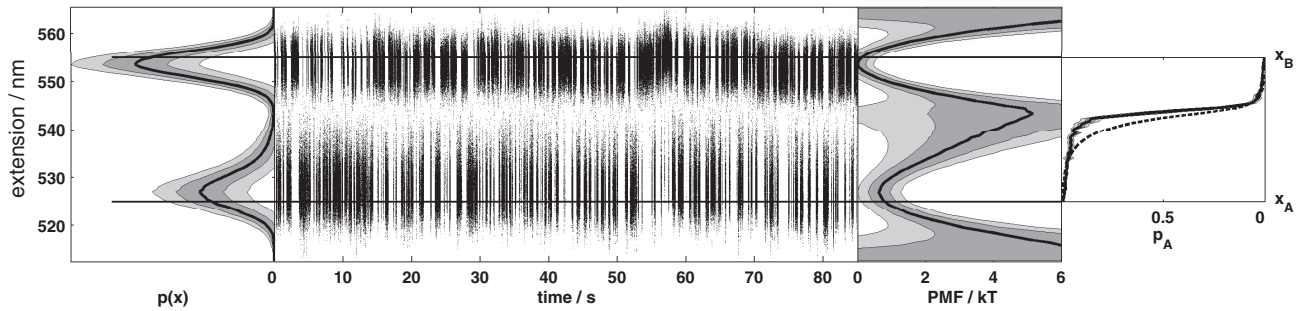

FIG. 3. Splitting probability analysis for a DNA hairpin in a passive all-optical constant-force double trap. From left to right: Histogram of observed values of the extension coordinate; complete observed trajectory of extension coordinate over experimental time course; potential of mean force along extension coordinate estimated from histogram; splitting probabilities estimated directly from trajectory (solid line) and computed from the potential of mean force (dashed line) using Eq. (2). Dark shaded regions around solid lines represent a 68% symmetric confidence interval, and light shaded regions 95% confidence interval. Note that the bead-to-bead DNA hairpin extension coordinate (along the ordinate) is the same throughout all panels.

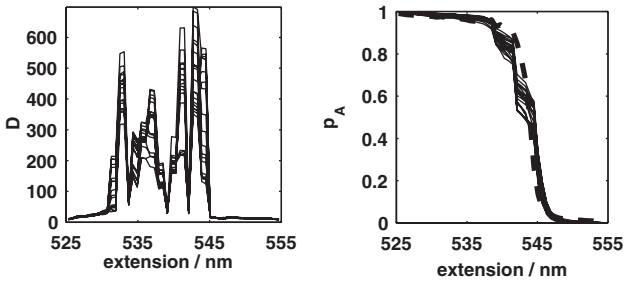

FIG. 4. Position-dependent diffusion constant and splitting probability test incorporating position-dependent diffusion for DNA hairpin. Left: Position-dependent diffusion constant, in nm/s^2. Right: Splitting probability test incorporating position-dependent diffusion constant, with empirical splitting probability $\hat{p}_A$ shown as a thick dashed line. Because the Bayesian scheme of Best and Hummer [15] was used to compute potentials of mean force and diffusion constants, the estimated diffusion constant $D(x)$ and PMF-derived splitting probabilities p_A are shown as thin solid lines representing 20 samples from the Bayesian posterior.

force $F(x)$ for the systems considered here (see Figs. 2 and 3 in the Supplemental Material [25]). Notably, the diffusion constant varies markedly with the bead-to-bead extension (Fig. 4, left), and the agreement of the PMF-derived p_A and empirical $\hat{p}_A$ (Fig. 4, right) improves substantially. In contrast, repeating the reaction coordinate test for the 2D model system allowing for a position-dependent diffusion constant reveals only relatively minor variations in the estimated diffusion constant that result in no substantial change in which reaction coordinates are rejected by the test (see Fig. 2 in the Supplemental Material [25]). Taken together, these data suggest a significant role for position-dependent diffusion in the DNA hairpin system under force, in agreement with the theoretical findings of Best and Hummer [15].

Discussion.—We note that the reaction coordinate test presented here only allows us to test a condition that is necessary, but not sufficient, for Brownian dynamics to appropriately describe the observed dynamics on a one-dimensional landscape determined by the PMF. This does not rule out the possibility of pathological cases where poor reaction coordinates go unnoticed because the average splitting probability at a particular value of the resolved coordinate matches the PMF-derived model, but the splitting probability distribution is not tightly peaked about its average value. Additionally, if multiple reactive channels exist that are otherwise indistinguishable by this test, differences between the channels will not be resolvable.

Despite this, our test was able to discern good from poor choices of reaction coordinate in a model system and reject the extension coordinate as a good choice of coordinate for a DNA hairpin unless a strongly position-dependent diffusion constant is permitted. Even then, there are statistically significant discrepancies between the observed splitting probability and the PMF-derived splitting probability that indicate this reaction coordinate choice is not ideal. We note that the presence of ~1 kb dsDNA handles tethering the DNA hairpin to the laser-trapped polystyrene beads is one potential source of the incomplete alignment of the extension coordinate with the reaction coordinate for hairpin unzipping. Shorter dsDNA handles have recently been suggested as a way to improve the signal-to-noise ratio [26], and may also improve the reaction coordinate quality. For proteins, techniques that allow the attachment of tethers at specific attachment points can be exploited to probe for improved reaction coordinate should the experimenter find that the current pulling coordinate under study is unsuitably poor [27]. Finally, we note that though this test is able to test the suitability of the extension coordinate for a polymer under force, we cannot determine from the present analysis whether a good reaction coordinate in the presence of external force would also be a good reaction coordinate in the absence of force, or even under different biasing forces; this concern is still the subject of active study [28,29].

The authors thank Michael Woodside (University of Alberta and National Institute for Nanotechnology, NRC), Phillip Elms and David Chandler (U. Berkeley), Gerhard Hummer and Attila Szabo (NIH), Steven Block and Imran Haque (Stanford University), Felix Ritort (University of Barcelona), and the anonymous referees for their helpful feedback on this work. The authors are grateful to Michael Woodside and Steven M. Block (Stanford University) for kindly providing original single-molecule data. J. D. C. acknowledges support through a NSF grant for Cyberinfrastructure (NSF CHE-0535616) and a California Institute of Quantitative Biosciences (QB3) Distinguished Postdoctoral Fellowship. V. S. P. acknowledges support from NIH R01-GM062868, NSF-DMS-0900700, NSF-MCB-0954714, and NSF EF-0623664. For the MATLAB code implementing the analysis procedure described here, see Ref. [30].

*jchodera@berkeley.edu
†Corresponding author.
pande@stanford.edu

[1] G. J. Smith, K. T. Lee, X. Qu, Z. Xie, J. Pesic, T. R. Sosnick, T. Pan, and N. F. Scherer, J. Mol. Biol. **378**, 943 (2008).
[2] W. J. Greenleaf, M. T. Woodside, E. A. Abbondanzieri, and S. M. Block, Phys. Rev. Lett. **95**, 208102 (2005).
[3] P. T. X. Li, D. Collin, S. B. Smith, C. Bustamante, and I. Tinoco, Biophys. J. **90**, 250 (2006).
[4] M. T. Woodside, P. C. Anthony, W. M. Behnke-Parks, K. Larizadeh, D. Herschlag, and S. M. Block, Science **314**, 1001 (2006).
[5] C. W. Gardiner, *Handbook of Stochastic Methods* (Springer, New York, 2003), 3rd ed.

4

[6] C. Schütte and W. Huisinga, in *Computer Simulations in Condensed Matter: Systems: From Materials to Chemical Biology. Volume I*.

[7] L. Onsager, Phys. Rev. **54**, 554 (1938).

[8] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).

[9] P. L. Geissler, C. Dellago, and D. Chandler, J. Phys. Chem. B **103**, 3706 (1999).

[10] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Adv. Rev. Phys. Chem. **53**, 291 (2002).

[11] A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).

[12] B. Peters, J. Chem. Phys. **125**, 241101 (2006).

[13] B. Peters, Chem. Phys. Lett. **494**, 100 (2010).

[14] Y. M. Rhee and V. S. Pande, J. Phys. Chem. B **109**, 6780 (2005).

[15] R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **107**, 1088 (2009).

[16] J. C. M. Gebhardt, T. Bornschlögl, and M. Rief, Proc. Natl. Acad. Sci. U.S.A. **107**, 2013 (2010).

[17] M. R. Shirts and J. D. Chodera, J. Chem. Phys. **129**, 124105 (2008).

[18] D. D. L. Minh and A. B. Adib, Phys. Rev. Lett. **100**, 180602 (2008).

[19] D. D. L. Minh and J. D. Chodera, J. Chem. Phys. **131**, 134110 (2009).

[20] G. Morrison, C. Hyeon, M. Hinczewski, and D. Thirumalai, Phys. Rev. Lett. **106**, 138102 (2011).

[21] B. Peters, G. T. Beckham, and B. L. Trout, J. Chem. Phys. **127**, 034109 (2007).

[22] W. Lechner, J. Rogal, J. Juraszek, B. Ensing, and P. G. Bolhuis, J. Chem. Phys. **133**, 174110 (2010).

[23] D. L. Ermak and Y. Yeh, Chem. Phys. Lett. **24**, 243 (1974).

[24] D. L. Ermak, J. Chem. Phys. **62**, 4189 (1975).

[25] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.000.000000 for [brief description].

[26] N. Forns, S. de Lorenzo, M. Manosas, K. Hayashi, J. M. Huguet, and F. Ritort, Biophys. J. **100**, 1765 (2011).

[27] C. Cecconi, E. A. Shank, F. W. Dahlquist, S. Marqusee, and C. Bustamante, Eur. Biophys. J. **37**, 729 (2008).

[28] J. Nummela and I. Andricioaei, Biophys. J. **93**, 3373 (2007).

[29] R. B. Best, E. Paci, G. Hummer, and O. K. Dudko, J. Phys. Chem. B **112**, 5968 (2008).

[30] MATLAB code, https://simtk.org/home/splitting

5