Supporting Information

Extension-Dependent Drift-Velocity and Diffusion (DrDiff) Directly Reconstructs the Folding Free Energy Landscape of Atomic Force Microscopy Experiments

Frederico Campos Freitas and Ronaldo Junio de Oliveira*

Laboratório de Biofísica Teórica, Departamento de Física, Instituto de Ciências Exatas, Naturais e Educação, Universidade Federal do Triângulo Mineiro, Uberaba, MG, Brazil.

E-mail: ronaldo.oliveira@uftm.edu.br

DrDiff approach

The idea of study protein folding using diffusive coordinates in rough energy landscape was started by Bryngelson-Wolynes in the late 80's[1,2]. The diffusion equation was derived under the assumption that an arbitrary reaction coordinate Q can only be changed by gradual relatively small steps, which is a general hypothesis. The diffusive dynamics can be described by the Fokker-Planck equation, which is the probability flow in time ($P(Q,t)$) of stochastic motion superimposed with deterministic drift-velocity ($v(Q)$), and is given by[3]

$$\frac{\partial P(Q,t)}{\partial t} = \left[-\frac{\partial}{\partial Q}v(Q) + \frac{\partial^2}{\partial Q^2}D(Q)\right]P(Q,t). \tag{S1}$$

$v(Q)$ and $D(Q)$ are the coordinate-dependent drift-velocity and diffusion coefficients, respectively.

The solution of equation S1 for short-time dynamics is given by

$$P(Q,t) = -\frac{1}{\sqrt{4\pi D(Q_c)t}}\exp\left[-\frac{(Q-Q_c-v(Q_c)t)^2}{4D(Q_c)t}\right] \tag{S2}$$

with the initial condition at $P(Q,t=0) = \delta(Q_c)$. This solution represents a Gaussian distribution centered at Q_c, moving with a velocity of $v(Q_c)$ and widening as a function of the square root of t ($\sigma(t) = \sqrt{2D(Q_c)t}$). The drift-velocity ($v(Q)$) and the diffusion ($D(Q)$) coefficients can be calculated from Gaussian distributions with the shift of centers ($Q_c(t)$) and growths of widths ($\sigma^2(t)$) by the expressions

$$v(Q) = \frac{Q_c(t_2) - Q_c(t_1)}{\Delta t} \tag{S3}$$

and

$$D(Q) = \frac{\sigma^2(t_2) - \sigma^2(t_1)}{2\Delta t}. \tag{S4}$$

with $\Delta t = t_2 - t_1$. Therefore, v and D should be taken in the limit of $\Delta t \to 0$.

The stochastic diffusion-drift (DrDiff) approach is an updated Python script implemented accordingly to the theory described in the previous work[4,5]. This updated version of the algorithm has implemented the folding time (τ_f) and transition path time (τ_{TP}) by using the D and v obtained by the stochastic approach. The DrDiff algorithm was implemented in the Python Programming Language (`https://python.org`) using the NumPy and SciPy native libraries, which makes it faster and automated. The code is freely available to be downloaded at `https://github.com/ronaldolab/DrDiff`.

The only input required by DrDiff is the trajectory ($Q(t)$) file and the required parameters are the number of equilibration steps (number of data values to be ignored from the beginning), the size of Q_{bins} read from the trajectory, the time step and the snapshot used to save $Q(t)$, t_{min} and t_{max}. The key point for the DrDiff approach is to wisely collect histograms by only reading one-dimensional trajectory as a function of time ($Q(t)$). The algorithm that collects histograms from $Q(t)$ and calculates $D(Q)$ and $v(Q)$ is the following:

1. The full one-dimensional trajectory data $Q(t)$ is read with the Q boundaries identified (Q_{min} and Q_{max}). Equilibration steps are discarded from the beginning of the trajectory due to thermal equilibrium.

2. Each Q in the range $[Q_{min}, Q_{max}]$ within a bin size (Q_{bin}) is indexed from the trajectory.

3. From each of these Q's, histogram distributions ($P(Q,t)$) are collected at elapsed times in the range $[t_{initial}, t_{final}]$.

4. The histograms collected are then fitted to Gaussian distributions given by equation S2. Standard deviations ($\sigma(t)$) and distribution centers ($Q_c(t)$) are recorded for the shooting times in the range $[t_{initial}, t_{final}]$.

5. In sequence, it is performed linear regressions in the variance $\sigma^2(t)$ and centers $Q_c(t)$ data-points. D is the slope divided by 2 of the $\sigma^2(t)$ fitting and v is exactly the slope of $Q_c(t)$ fitting.

6. This process is iterated in Q within the range $[Q_{min}, Q_{max}]$ in order to obtain $D(Q)$ and $v(Q)$ along with the reaction coordinate.

Time steps in the range $[t_{initial}, t_{final}]$ of the data points are chosen small enough in order to be valid the short-time approximation.

Figure S1 shows an illustration of the algorithm for $Q = 0.1$ applied in a $Q(t)$ trajectory. $Q(t)$ was read and all $Q = 0.1$ were indexed in Figure S1A. Figure S1B shows the Qs landed in the range $t_{initial} = 0.01$ and $t_{final} = 0.06$ for six time windows. In Figure S1C, the histograms collected starting from $Q = 0.1$ were fitted with Gaussian functions. $\sigma^2(t)$ and $Q_c(t)$ parameters were recorded from each Gaussian in the time range and linear regressions were performed to obtain v and D at $Q = 0.1$ as in the equations S3 and S4.

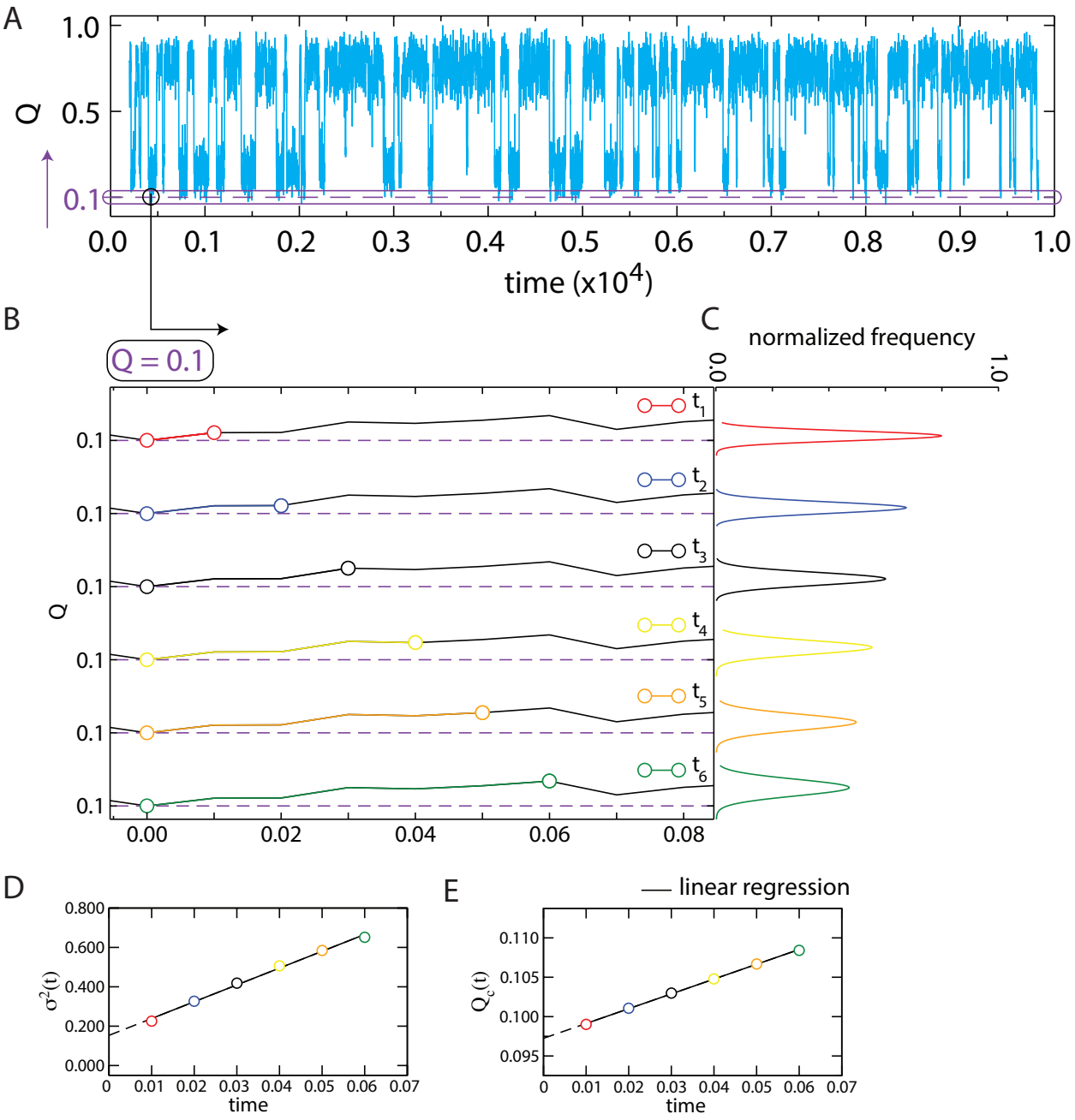

Figure S1: A) One-dimensional trajectory as a function of time ($Q(t)$) is read for $Q = 0.1$. B) Q's landed in the range $t_{initial} = 0.01$ and $t_{final} = 0.06$ for six time windows. C) Gaussian functions fitting the histograms collected for the reached Q's starting from $Q = 0.1$. D) Standard deviations $\sigma^2(t)$ and E) centers $Q_c(t)$ extracted from each of the Gaussian fitting in C. Diffusion ($D(Q = 0.1)$) and drift ($v(Q = 0.1)$) coefficients are obtained by the slope of the linear regression of $\sigma^2(t)$ divided by 2 in D and $Q_c(t)$ in E, respectively.

1D Numerical Langevin Dynamics of a Single Particle

The Langevin equation of the stochastic process was chosen to computationally generate long time trace trajectories with given free-energy ($F(x)$) and diffusion coefficient ($D(x)$)[3,6] and recover them with the diffusive models.

The Langevin equation that corresponds to the stochastic process of the Fokker-Planck equation S1) is given by

$$\frac{dQ(t)}{dt} = v(Q) + \eta(Q,t) \tag{S5}$$

where $v(Q)$ is the drift-velocity and $\eta(Q,t)$ is a Gaussian white noise, related to stochastic processes following a normal distribution with zero mean, and with variance related to the fluctuation-dissipation theorem

$$\langle \eta(Q,0)\eta(Q,t)\rangle = 2D(Q)\delta(t) \tag{S6}$$

where D given by $D = k_bT/\gamma$ (γ is the damping coefficient) if one uses the Itô interpretation of the Langevin equation[7]. The Gaussian-type white noise distribution of η can be approximated to

$$P[\eta] \propto \exp\left(-\int^{\delta t} \frac{\eta^2(t)}{4D}dt\right) \overset{\delta t\to 0}{\approx} \exp\left(-\frac{\eta^2\delta t}{4D}\right). \tag{S7}$$

By defining $\tilde{\eta} = \eta\sqrt{\delta t}$, the Langevin equation can be numerically solved by

$$x(t+\delta t) = x(t) + v(x)\delta t + \tilde{\eta}(x)\sqrt{\delta t} \tag{S8}$$

with $\tilde{\eta}$ being the redefined Gaussian random number distribution with zero mean and standard deviation $\sigma_{\tilde{\eta}} = \sqrt{2D}$. Equation S8 can be numerically solved to generate dynamical trajectories $x(t)$ with $D(x)$ and $v(x)$ (or $F(x)$) plugged in. The numerical algorithm that integrates equation S8 was implemented with the Perl Programming Language (`https://perl.org`) with native libraries in a in-house script.

Figure S2 shows numerical simulations of the Langevin equation of a particle diffusing in double-well free-energy profiles ($F(x)$) with sinusoidal $D(x)$ (left panels) and constant $D = 2$ (right panels).

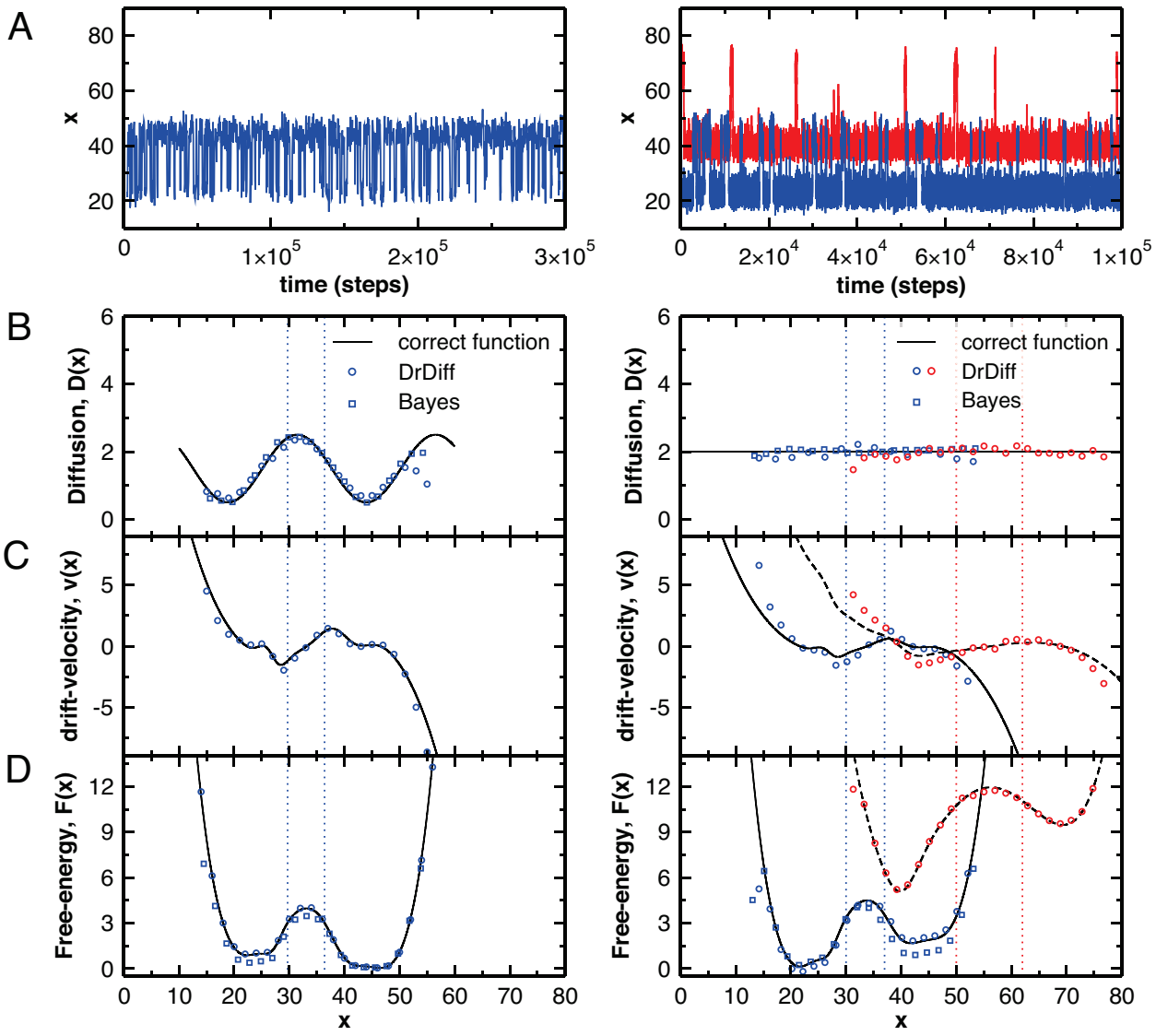

Figure S2: A single particle diffusing in numerical Langevin simulations. A) One-dimensional time-serie trajectory ($x(t)$) generated by numerical integration of the Langevin equation S8 with inputted $F(x)$ (continuous curves in D) and $D(x)$ (continuous curve in B), the termed correct functions. Right panel in A) contains two trajectories (blue and red curves) generated with the same $D = 2$, yet with two different Fs shown in the right panel in B) and D), respectively. B) Diffusion ($D(x)$), C) drift-velocity ($v(x)$) and D) free-energy profile ($F(x)$) as a function of the position x. The inputted F and D were accurately recovered by the DrDiff approach (circles), and also by the Bayesian analysis[8] (squares), used here for comparison. The recovered $v(x)$ in C) by DrDiff (in blue circles) is superposed with minus the gradient of the inputted free-energy profile $F(x)$ from D). Energy is in arbitrary units of k_BT so that $k_BT = 1$. Dashed vertical lines delimitate defined transition states (TS). The Bayesian analysis was performed with the same previous protocol[4,5].

2D Numerical Brownian Dynamics of a Single Particle

The anisotropic 2-dimensional (2D) Brownian dynamics simulations were performed by Cossio and collaborators[9,10] and analyzed here with the DrDiff methodology. The 2D numerical simulation is presented for the sake of completion. In this 2D system, x is the molecular (hidden) extension and q is the total (observable) extension. Trajectories were generated along q (probe coordinate) and x (molecule coordinate) by numerical integration of[9]

$$\begin{aligned} q(t+\Delta t) &= -\beta \partial_q G(q,x) D_q \Delta t + \sqrt{2 D_q \Delta t} R_q(t) \\ x(t+\Delta t) &= -\beta \partial_x G(q,x) D_x \Delta t + \sqrt{2 D_x \Delta t} R_x(t) \end{aligned} \tag{S9}$$

with $R_{\{q,x\}}$ being the independent Gaussian random numbers with zero mean and unit variance, Δt the time step, $G(x,q)$ the free-energy profile, $\beta = 1/k_B T$ with k_B being the Boltzmann constant and T is the simulation temperature. The time step is such that $D_x \Delta t = 5 \times 10^{-4}$. The diffusion coefficient of the molecule was kept constant ($D_x = 0.1$), while the diffusion coefficient of the pulling device (molecule + linker + apparatus) D_q were varied in decades from 10^{-4} to 1. $q(t)$ for different D_q were recorded for analysis.

The 2D free-energy surface for a constant force exerted on the system was given by

$$G(q,x) = G_o(x) + \frac{\kappa_l}{2}(x-q)^2 \tag{S10}$$

with $G_o(x)$ been the molecule free-energy subjected to the force and the second term been the coupling due to a harmonic linker with spring constant κ_l. Five trajectories along the measured extension q of the 2D Brownian dynamics simulations were generated with similar surface parameters extracted from the 20TS06/T4 DNA hairpin[11], $\Delta G_0^{\ddagger} = 8.1 k_B T$, $\Delta x^{\ddagger} = 1.5[x]$, and $\kappa_l = 2.6 k_B T/[q]^2$, where [q] = [x] denotes units of length for the extension.

Intrinsic dynamical property of force spectroscopy

The measured dynamics of molecular folding transitions in single-molecule force spectroscopy assays are affected by the hydrodynamic drag on the pulling instrument. The intrinsic molecular diffusion coefficient (D) is then affected by this hydrodynamic drag of the pulling setup (tip/cantilever plus the DNA linker in the case of AFM experiments) resulting in a relatively lower apparent diffusion coefficient (D_{app}). The ratio between the apparent measured and the intrinsic diffusion coefficients is given by[12]

$$\alpha = \frac{D_{app}}{D}. \tag{S11}$$

The scaling factor α is calculated by

$$\alpha = \frac{-(\tau + k - 1) + \sqrt{(\tau + k - 1)^2 + 4\tau}}{2} \tag{S12}$$

with $\tau = \tau_i/\tau_s$ and $k = k_s/k_i$. Here, τ_i and τ_s are the intrinsic molecular and pulling setup relaxation times, respectively. k_i and k_s are related to the intrinsic molecular and pulling setup spring constant stiffness, respectively. τ_i and k_i are associated with the molecular friction coefficient γ_i, $\tau_i = \gamma_i/k_i$, where γ_i is connected to D through the Einstein-Smoluchowski relationship $\gamma_i D = k_B T$. A more detailed description of the theory was given by Makarov[12].

Estimations of the apparent scaling factor (α) were conducted for the RNA hairpin and the 3-aa BR protein with the parameters presented in table S1.

Table S1: Parameters of the two systems used to estimate α in equation S12.

	RNA hairpin[13]	3-aa of BR[14]
τ_i (ms)	6.3	0.14
τ_s (μs)	41	1
k_i (pN/nm)^a	1 – 5	1 – 5
k_s (pN/nm)	4	58
α	0.98 – 0.99	0.70 – 0.93

The parameters were obtained from the original works in each case, except for k_i.
$^a k_i$ was estimated for protein and nucleic acid unfolding and refolding experiments in optical tweezers studies (references in Makarov[12]), which resulted in the lower and upper values of the estimated α.

Memory effects in the single-molecule time-series trajectory

In general, the folding process is modeled as a Markov process, without memory, along one-dimensional (1D) reaction coordinates. However, if the molecule is probed in a single-molecule force spectroscopy experiment, theory predicts that memory can be induced by the probe and, in this case, dynamics along the 1D coordinate becomes a non-Markovian process[15]. Using the generalized Langevin equation (GLE), the memory signal can be extracted from an experimental signal ($q(t)$)[16,17]. The dynamics of both, the probe (q coordinate) and the molecule (x coordinate), is described with a two-dimensional (2D) overdamped Langevin equation and the x coordinate is eliminated to obtain a GLE in terms of the measured q coordinate. The memory effects induced by the apparatus are extracted from a normalized autocorrelation function (ACF), which results as a solution of the GLE[15]. The autocorrelation function ($\chi_q(t)$) is given by

$$\chi_q(t) = C\exp(-t/\tau_1) + (1-C)\exp(-t/\tau_2). \tag{S13}$$

The memory induced by linking the molecule to a probe is given by the relaxation time (τ_{mem}) obtained after fitting equation S13 to the computed ACF from the trajectory and it is given by

$$\tau_{mem} = \frac{C(s_2 - s_1) - s_2}{s_1 s_2} \tag{S14}$$

and the probe relaxation time (τ_p) is obtained by

$$\tau_p = \frac{1}{C(s_2 - s_1) - s_2}. \tag{S15}$$

Both parameters S14 and S15 are evaluated with the characteristic times found in the fitting: $\tau_1 = -s_1^{-1}$ and $\tau_2 = -s_2^{-2}$.

In practice, the ACF is calculated directly from the observed time-series trajectory of the probe position ($q(t)$), then it is fitted to equation S13 to extract the fitting parameters C, τ_1 and τ_2. According to the theory, the condition $\tau_{mem}/\tau_p > 1$ indicates that the probe responds faster than the molecule and then its motion reflects the dynamics of the molecule. A more detailed description of the theory is given by Pyo and Woodside[15].

Figure S3 shows the autocorrelation functions from pulling measurements of the HIV RNA hairpin and the 3-aa segment of the bacteriorhodopsin (BR) membrane protein.

Estimations of the memory effects induced by the apparatus given by equations S14 and S15 for the RNA hairpin and the 3-aa BR protein are in table S1.

Table S2: Memory and probe relaxation times of the two molecules.

		RNA hairpin	3-aa of BR
τ_{mem} (μs)	U	20	87
	F	20	127
τ_p (μs)	U	3	8
	F	0.4	7
$\frac{\tau_{mem}}{\tau_p}$	U	7	10
	F	43	16

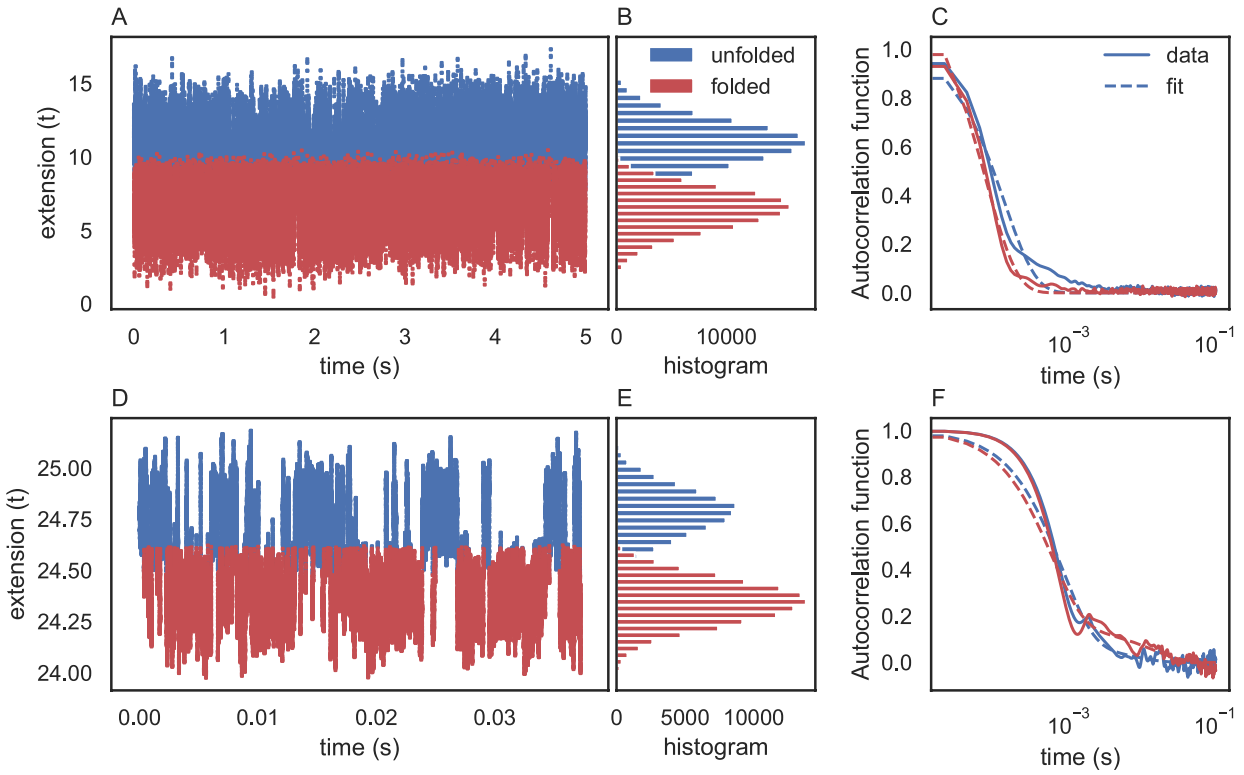

Figure S3: Autocorrelation function (ACF) analyses from measurements of the RNA hairpin (top panels) and the partial membrane protein (bottom panels) of this study. A) and D) are the time-series ($q(t)$) analysed by separating the unfolded (blue) and folded (red) sampling states. B) and E) represent the histograms of each populated state. C) and F) show the computed ACF from the trajectories (continues lines) and the $\chi_q(t)$ function after fitting to equation S13 (dotted lines).

References

(1) Bryngelson, J.; Wolynes, P. Spin-glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 7524–7528.

(2) Bryngelson, J.; Wolynes, P. Intermediates and Barrier Crossing in a Random Energy-Model (with Applications to Protein Folding). *J. Phys. Chem.* **1989**, *93*, 6902–6915.

(3) Risken, P. D. H. *The Fokker-Planck Equation*; Springer Series in Synergetics 18; Springer Berlin Heidelberg, 1984; pp 63–95.

(4) de Oliveira, R. J. Stochastic diffusion framework determines the free-energy landscape and rate from single-molecule trajectory. *J. Chem. Phys.* **2018**, *149*, 234107.

(5) Freitas, F. C.; Lima, A. N.; Contessoto, V. d. G.; Whitford, P. C.; Oliveira, R. J. d. Drift-diffusion (DrDiff) framework determines kinetics and thermodynamics of two-state folding trajectory and tunes diffusion models. *J. Chem. Phys.* **2019**, *151*, 114106.

(6) Tomé, T.; de Oliveira, M. J. *Stochastic Dynamics and Irreversibility*; Graduate Texts in Physics; Springer International Publishing: Cham, 2015; DOI: 10.1007/978-3-319-11770-6.

(7) Kopelevich, D. I.; Panagiotopoulos, A. Z.; Kevrekidis, I. G. Coarse-grained kinetic computations for rare events: Application to micelle formation. *J. Chem. Phys.* **2005**, *122*, 044908.

(8) Hummer, G. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.* **2005**, *7*, 34.

(9) Cossio, P.; Hummer, G.; Szabo, A. On artifacts in single-molecule force spectroscopy. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 14248–14253.

(10) Cossio, P.; Hummer, G.; Szabo, A. Transition paths in single-molecule force spectroscopy. *J. Chem. Phys.* **2018**, *148*, 123309.

(11) Neupane, K.; Woodside, M. T. Quantifying Instrumental Artifacts in Folding Kinetics Measured by Single-Molecule Force Spectroscopy. *Biophys. J.* **2016**, *111*, 283–286.

(12) Makarov, D. E. Communication: Does force spectroscopy of biomolecules probe their intrinsic dynamic properties? *141*, 241103.

(13) Walder, R.; Van Patten, W. J.; Ritchie, D. B.; Montange, R. K.; Miller, T. W.; Woodside, M. T.; Perkins, T. T. High-Precision Single-Molecule Characterization of the Folding of an HIV RNA Hairpin by Atomic Force Microscopy. *Nano Lett.* **2018**, *18*, 6318–6325.

(14) Yu, H.; Siewny, M. G. W.; Edwards, D. T.; Sanders, A. W.; Perkins, T. T. Hidden dynamics in the unfolding of individual bacteriorhodopsin proteins. *Science* **2017**, *355*, 945–950.

(15) Pyo, A. G. T.; Woodside, M. T. Memory effects in single-molecule force spectroscopy measurements of biomolecular folding. *21*, 24527–24534.

(16) Medina, E.; Satija, R.; Makarov, D. E. Transition Path Times in Non-Markovian Activated Rate Processes. *122*, 11400–11413.

(17) Satija, R.; Makarov, D. E. Generalized Langevin Equation as a Model for Barrier Crossing Dynamics in Biomolecular Folding. *123*, 802–810.