

Noise reduction in single-molecule fluorescence trajectories of folding proteins

Gilad Haran *

Department of Chemical Physics, Weizmann Institute of Science, Rehovot 76100, Israel

Received 9 February 2004; accepted 19 May 2004

Available online 15 June 2004

Abstract

A method for noise reduction in stochastic trajectories of single molecules is described. The method generalizes a nonlinear filtration technique developed by Chung and Kennedy for single-channel recording [J. Neurosci. Meth. 40 (1991) 71] and applies it to the case of two correlated trajectories, as in a fluorescence resonance energy transfer experiment. Thus it is particularly suitable for single-molecule studies of protein folding. It is shown that the nonlinear filter facilitates the detection of various intermediate conformational states in a noisy trajectory, and can provide better estimates for the temporal positions of jumps between states and dwell times than a standard low-pass filter. It is finally suggested that the filter can also be of use for directly resolving molecular motion through the transition state region on the folding energy landscape.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Protein folding; Single-molecule spectroscopy; Noise reduction

1. Introduction

Single-molecule fluorescence spectroscopy has emerged in recent years as a powerful tool for unraveling the heterogeneous behavior of biological macromolecules during the folding process [1]. Much can be learned from measurements of individual molecules as they freely diffuse through a laser beam [2,3]. A particularly promising form of the single-molecule experiment, however, involves molecules immobilized near a surface, so that each molecule can be studied for an extended period of time. This experiment allows direct visualization of transitions between the various conformational states of a molecule. Extensive experiments on the folding of individual surface-immobilized RNA molecules were carried out by Chu and co-workers [4,5] and by Bokinsky et al. [6]. In the context of protein folding, pioneering work was carried out by Hochstrasser and co-workers [7,8], who looked at the dynamics

of a helical dimer, GCN4-P1. Rhoades et al. [9] have recently shown that single-molecule folding studies can be facilitated by trapping protein molecules within surface-tethered lipid vesicles. This method was devised in order to prevent direct interaction with the surface, yet allow long-time trajectories to be collected. It enabled visualizing folding and unfolding jumps in the protein adenylate kinase.

All of the above experiments used fluorescence resonance energy transfer (FRET) as the observable [10]. The value of FRET stems from its ability to provide information on the evolution of a specific distance within a macromolecule. In a typical FRET experiment donor and acceptor fluorescent probes are specifically attached at particular positions on a macromolecule. Changes in the conformational state of the macromolecule lead to large changes in the distance between the donor and acceptor and hence a large change in the efficiency of FRET. If the molecule is prepared under conditions where the free energies of two states (say the folded and unfolded states) are equal it will readily jump between these two states, and the jumping rate will match the ensemble kinetics. This is the basis for the

* Tel.: +97-289-342-625; fax: +97-289-342-749.

E-mail address: gilad.haran@weizmann.ac.il (G. Haran).

equilibrium version of the single-molecule folding experiment.

What is the novel information that can be gleaned from trajectories of individual folding molecules? First and foremost, it should be possible to enumerate the number of states populated during the folding or unfolding process. Many small globular proteins are believed to populate only two states, folded or unfolded, based on ensemble studies [11]. The two-state behavior was also observed in a recent single-molecule experiment in our lab [12]. However, some proteins exhibit intermediate states, which can be low enough in energy to be populated at equilibrium, or of high energy so that they are not significantly populated at equilibrium but are still transiently populated during a transition between stable states. Intermediate states should be directly observable in single molecule trajectories, as seen by Zhuang et al. [13] in RNA folding, and Rhoades et al. [9] in protein folding.

A second aspect of interest in single-molecule trajectories is the dynamics at various points on the energy landscape. While some knowledge on dynamics can be obtained from analysis of free diffusion experiments [3], where long-time trajectories are not available, a much more detailed look at chain dynamics is available from the latter. For example, if the trajectories are long enough it is possible to test whether the folding/unfolding rate is constant in time (similarly to the experiment of Xie and co-workers [14] who probed the activity of individual enzyme molecules). It is also possible to calculate correlation functions from fluctuations in the trajectories, in order to obtain information on the time scale of chain motion, e.g., in the denatured state of a protein. Finally, as will be discussed more fully below, it should be possible to clock the motion of a protein through the transition state of the folding reaction.

The actual length of trajectories obtained in single-molecule experiments depends on the photobleaching propensity of the fluorescent dyes used, which in turn depends on the laser power used. A rather low laser power of 0.6 μW had to be used in our experiment on the folding of adenylate kinase [9] in order to obtain fluorescence trajectories with an average length of ~ 10 s. The payoff for that was a low photon count rate and thus noisy trajectories. In order to identify various conformational states in the trajectories we had to use a noise reduction method. The practitioners of single ion-channel recording had elaborated a variety of methods to treat the same problem of the identification of several signal levels within a noisy experimental trajectory [15]. We found that one of these methods, the nonlinear forward–backward filter developed by Chung and Kennedy [16], can be used effectively for noise reduction in FRET trajectories [9]. This non-linear filter (NLF) was rather simple to implement and quite effective when

applied judiciously to photon trajectories. In this paper, we describe in some detail the implementation of this filter for single molecule fluorescence data and show its utility and versatility, which should make it of interest to a broad readership. Further, we generalize it to include information about anti-correlated transitions in FRET trajectories, thereby making it useful for single molecule studies of protein folding in particular and protein dynamics in general. We note a different and very interesting noise reduction approach recently described by Talaga and co-workers [17], which is based on hidden Markov models.

2. Simulation of FRET trajectories

In order to generate virtual trajectories useful for demonstration of the properties of the NLF we performed full Monte-Carlo simulations of the single molecule FRET experiment. In a typical simulation a set of conformational states of a protein was assumed, e.g., folded and unfolded, as well as rates for interconversion between these states. A particular donor–acceptor distance was assigned to each state, leading to a specific FRET efficiency value. The simulation, starting from a particular molecular state, proceeded in 1 μs steps. At each step of the simulation a stochastic decision was made as to whether a conformational transition was made. A second decision used the FRET efficiency value of a specific state to determine whether an excitation resided on the donor fluorophore or transferred to the acceptor fluorophore. Since we were interested in time scales significantly longer than either singlet-state or triplet-state lifetimes, we did not include the excitation and de-excitation processes in the simulation. Rather, we made a stochastic decision whether a photon was detected at each step of the simulation based on a certain average detection rate. The result of this procedure was a trajectory of photons emitted at various times either from the donor or from the acceptor. These two photon trajectories were then binned in 20 ms bins, unless otherwise noted. Noise was added to each trajectory to simulate background noise, assuming a background count level of ~ 100 Hz per channel. FRET efficiency was calculated from the trajectories using the relation $E = I_a / (I_a + I_d)$, where $I_{a,d}$ are the acceptor and donor intensities, respectively.

3. The nonlinear forward–backward filter

The NLF builds on the familiar running-average filter (RAF), which is a form of a low-pass filter, to include knowledge about the data in order to avoid distorting it. In one possible realization of the RAF the intensity at data point i , $I(i)$, is represented in the averaged trajec-

tory by the average of the N intensities preceding it in the original trajectory

$$\hat{I}^f(i) = \frac{1}{N} \sum_{j=i-N}^{i-1} I(j). \quad (1)$$

This average is termed the “forward predictor of length N ” of the data point i , hence the label f . In a similar manner one can define the “backward predictor of length N ” of the data point i

$$\hat{I}^b(i) = \frac{1}{N} \sum_{j=i+1}^{i+N} I(j). \quad (2)$$

In the Chung and Kennedy scheme a series of forward and backward predictors with various lengths are calculated for each data point. A weighted sum of these K predictors is then used to represent the intensity at that point:

$$\hat{I}(i) = \sum_{k=1}^K [f_k(i)\hat{I}_k^f(i) + b_k(i)\hat{I}_k^b(i)]. \quad (3)$$

The weights $f_k(i)$ and $b_k(i)$ are calculated from the data, thus making $\hat{I}(i)$ a nonlinear function of it. Indeed, the weights are derived in such a way as to ensure that if a jump occurs within the stretch of data used to generate one of the predictors, the weight for that predictor is zeroed, so that it does not contribute to $\hat{I}(i)$. The fact that a series of predictors of different lengths are used allows the filter to work well on both short-time and longer-time fluctuations in the data. Explicit expressions for the weights were derived by Chung and Kennedy [16] in the appendix to their paper, using Bayesian statistics. The forward weight is given by

$$f_k(i) = C \left[\sum_{j=0}^{M-1} (I(i-j) - \hat{I}_k^f(i-j))^2 \right]^{-p} \quad (4)$$

and the backward weight by

$$b_k(i) = C \left[\sum_{j=0}^{M-1} (I(i-j) - \hat{I}_k^b(i-j))^2 \right]^{-p}, \quad (5)$$

where C is a normalization factor obtained from the condition

$$\sum_{k=1}^K [f_k(i) + b_k(i)] = 1. \quad (6)$$

M is a parameter which determines the size of the window over which the predictors are to be compared. p is a weighting factor, and a larger value of this factor more strongly accentuates nonzero weights. Both parameters are determined empirically (see a thorough discussion in Chung and Kennedy), and we have found after performing extensive simulations that for most purposes good smoothing results are obtained when M is varied between 10 and 20 and p is varied between 10 and 100.

All simulations presented below used the following parameters for the filter, unless noted otherwise: windows of 4, 8, 16, 32 and 64 points, $M = 10$, $p = 100$.

These values can serve as good starting points for data analysis but the reader is urged to experiment with them in order to obtain the best result. In particular, longer windows will lead to better noise-reduction in featureless sections of the data. Thus if one is confident that no transitions occur in the data on the time scale represented by, say, 4 or 8 points, one can drop the first two windows and obtain smoother data. A similar approach can be applied to the window size M – a larger value of this parameter will further smooth the data. As noted above, p is important in determining how many of the weights will contribute to the sum in Eq. (4) or (5). Thus in general a large value of p will lead to very sharp transitions in the smoothed data. It might be a good practice therefore to attempt smoothing first with a smaller value of this parameter, especially if the occurrence of slow changes is suspected (see Section 6).

To demonstrate the utility of the NLF in noise reduction in single-molecule data we show in Fig. 1(a) a simulated FRET trajectory of a single protein molecule

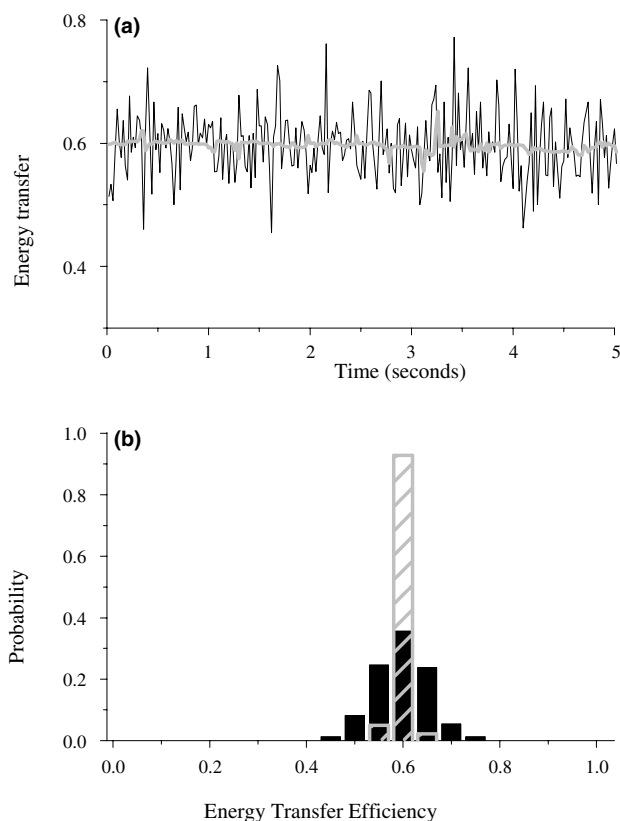


Fig. 1. Noise reduction by the NLF. (a) Simulated trajectory, with average FRET efficiency of 0.6 before (black line) and after (gray line) treatment by the filter. (b) Black bars – distribution of the noise in the original trajectory. Gray patterned bars – noise distribution after filtration.

in an unfolded state, where the average FRET efficiency is ~ 0.6 . Fig. 1(b) presents the distribution of energy transfer values calculated from the trajectory (black bars). The standard deviation of this distribution is 0.054. The gray line in Fig. 1(a) is the trajectory after treatment by the NLF, and the gray bars in Fig. 1(b) show the distribution of the filtered data. The standard deviation of this distribution is 0.013, which is smaller by almost a factor of 4 than the standard deviation of the unfiltered data. The main strength of the NLF is not, however, in noise reduction per se, but in its ability to preserve sharp jumps in the data. In order to make full use of this property for the analysis of FRET data we first need to modify the original scheme.

4. Generalizing the nonlinear filter for single-molecule FRET experiments

In the case of single-molecule FRET experiments one has the prior knowledge that each jump in the donor trajectory which is caused by a change in the donor/acceptor distance should be accompanied by an anti-correlated jump in the acceptor trajectory. It is possible to significantly improve the performance of the NLF by including this information in the analysis. This is done here by constructing new weights which are based both on the donor and on the acceptor trajectories. These weights can be derived using a similar procedure to that outlined in the appendix of [16]. The forward weight is given by

$$\bar{f}_k(i) = \left\{ \sum_{j=0}^{M-1} \left[\left(I_d(i-j) - \hat{I}_{d,k}^f(i-j) \right)^2 + \left(I_a(i-j) - \hat{I}_{a,k}^f(i-j) \right)^2 \right] \right\}^{-p} \quad (7)$$

The subscripts d and a now refer to the donor- and acceptor-related data or predictors. A similar expression is used for the backward weight. Both donor and acceptor trajectories are filtered using the same weights, $\bar{f}_k(i)$ and $\bar{b}_k(i)$. For example, the filtered donor signal is given by

$$\hat{I}_d(i) = \sum_{k=1}^K \left[\bar{f}_k(i) \hat{I}_{d,k}^f(i) + \bar{b}_k(i) \hat{I}_{d,k}^b(i) \right]. \quad (8)$$

The new form for the weights ensures that jumps in donor and acceptor data will occur at exactly the same position in time. This leads to sharper transitions in the FRET efficiency functions calculated from filtered data. Fig. 2 compares a portion of a folding simulation filtered using either the original form or the new form of the weights. It clearly demonstrates the better performance of the latter in preserving the position and sharpness of a FRET transition. The transitions at ~ 0.5

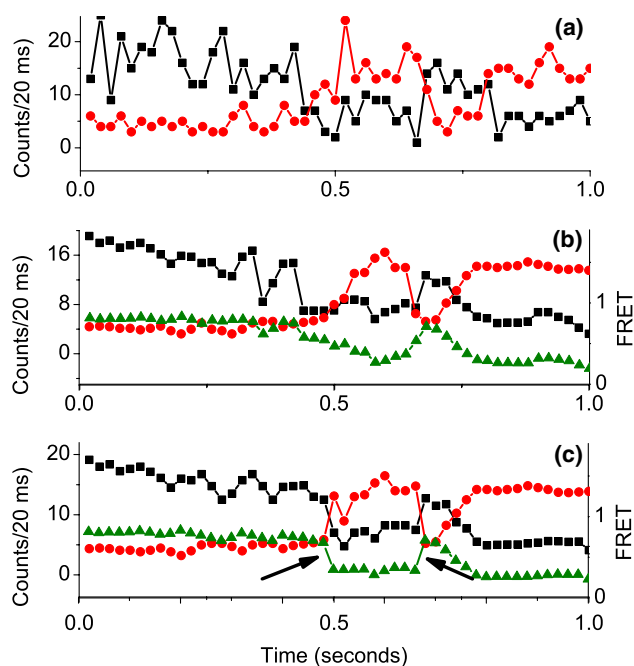


Fig. 2. Comparison of the original and generalized NLF. A stretch of simulated data (shown in (a), donor in red circles, acceptor in black squares) was first filtered with the original form of the NLF, and the data as well as the FRET efficiency calculated from it are shown in (b) (donor in red circles, acceptor in black squares, FRET efficiency in green triangles). The same stretch was filtered with the generalized filter (c, donor in red circles, acceptor in black squares, FRET efficiency in green triangles) and shows significantly sharper transitions (marked with black arrows). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and ~ 0.7 s, which are rather smeared in the energy transfer function calculated using the standard filter (Fig. 2(b), green line), appear as single-point jumps (as they should be) in the energy transfer function calculated with the generalized filter (Fig. 2(c), green line, transitions marked by black arrows).

5. Reduction of noise in single molecule FRET trajectories by the NLF

In order to show the usefulness of the NLF in the analysis of single-molecule folding experiments, we simulate a trajectory in which the FRET efficiency can change stochastically between the values 0.5 and 0.7, mimicking a two-state folding protein jumping between unfolded and folded conformations. Note the small difference between the values selected for the two states, presenting the NLF with a stringent test. The total photon emission rate is kept at a low value of 1000 Hz in order to obtain a high-noise trajectory. The donor and acceptor signals obtained in a typical simulation, after binning into 20 ms bins, are shown in Fig. 3(a), and the FRET efficiency calculated from these signals is shown

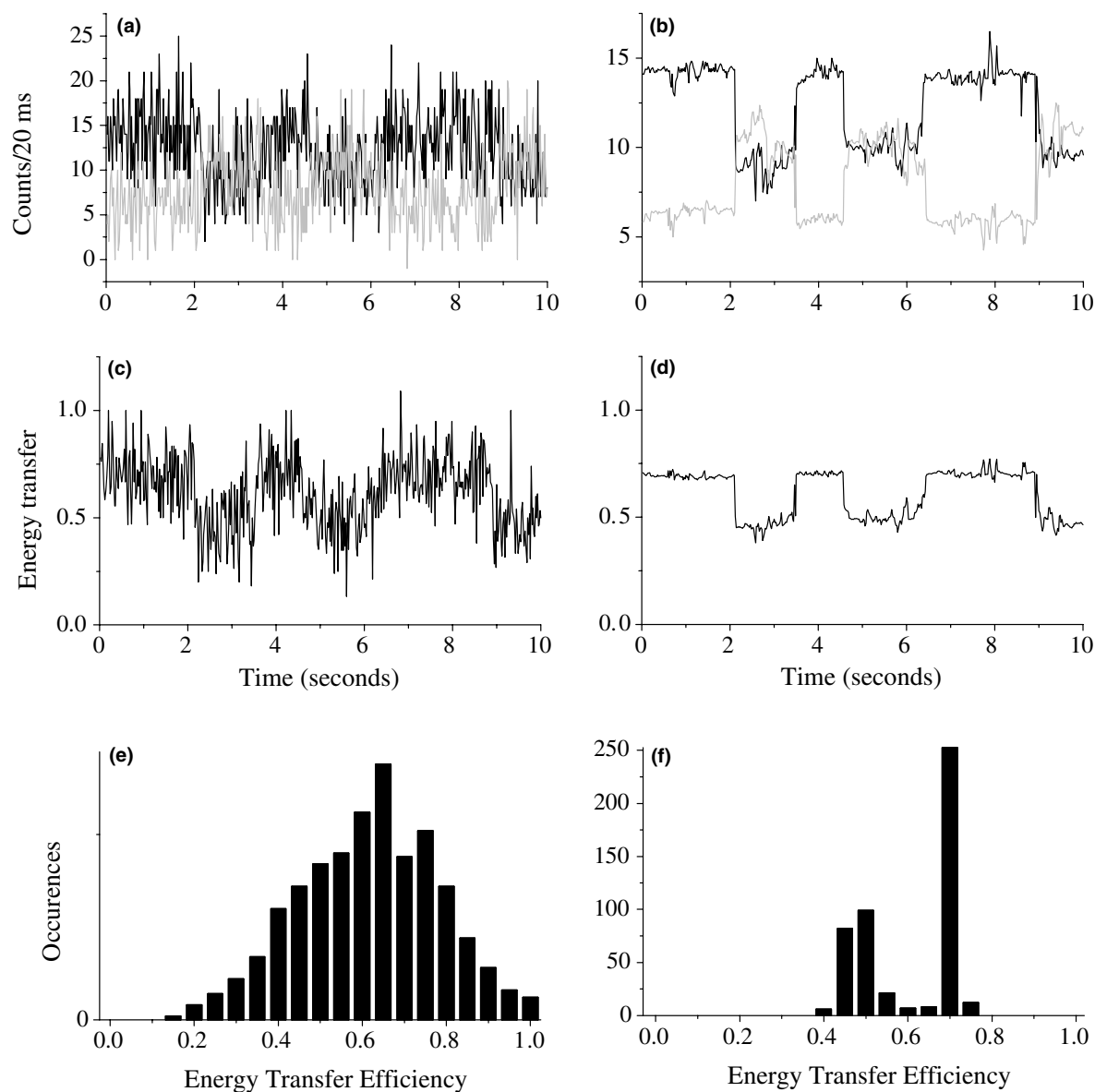


Fig. 3. Simulation of a two-state single-molecule protein folding experiment, in which the FRET value changes abruptly between 0.5 and 0.7. The overall count rate is 1000 Hz. (a) Simulated data. Acceptor in black and donor in gray. (b) Simulated data after filtration with the NLF. (c) FRET efficiency calculated from (a). (d) FRET efficiency calculated from (b). (e) Histogram of the FRET efficiency values of (c). (f) Histogram of the FRET efficiency values of (d).

in Fig. 3(c). Even with the noise level of this data set it is possible to establish the position of jumps between the two states. However, an attempt to histogram the data in order to look at the relative contribution of the two states (Fig. 3(e)) leads to a featureless distribution. After operating on the data with the NLF (Fig. 3(b) and (d)) it is much easier to locate transitions. The histogram in Fig. 3(f) clearly shows the two states at FRET values of 0.5 and 0.7 as two narrow peaks in the distribution, with only minimal ‘contamination’ in the region between them.

A similar and even more striking result is obtained when the FRET trajectory involves more than two

states. Fig. 4 shows a simulated trajectory with three states, at FRET values of 0.3, 0.5 and 0.7. Again, the simulation is carried out with an average count rate of 1000 Hz. It is very difficult to find out the number of states from the FRET efficiency calculated directly from the data. However, after filtration the three states are very obvious, and the histogram clearly shows three peaks.

Finally, it is instructive to directly look at the advantage of the NLF over a standard RAF. To this effect we plot in Fig. 5 the FRET efficiency curve obtained from a three-state simulation, this time with a photon count of 2500 Hz, after filtration with the NLF (black

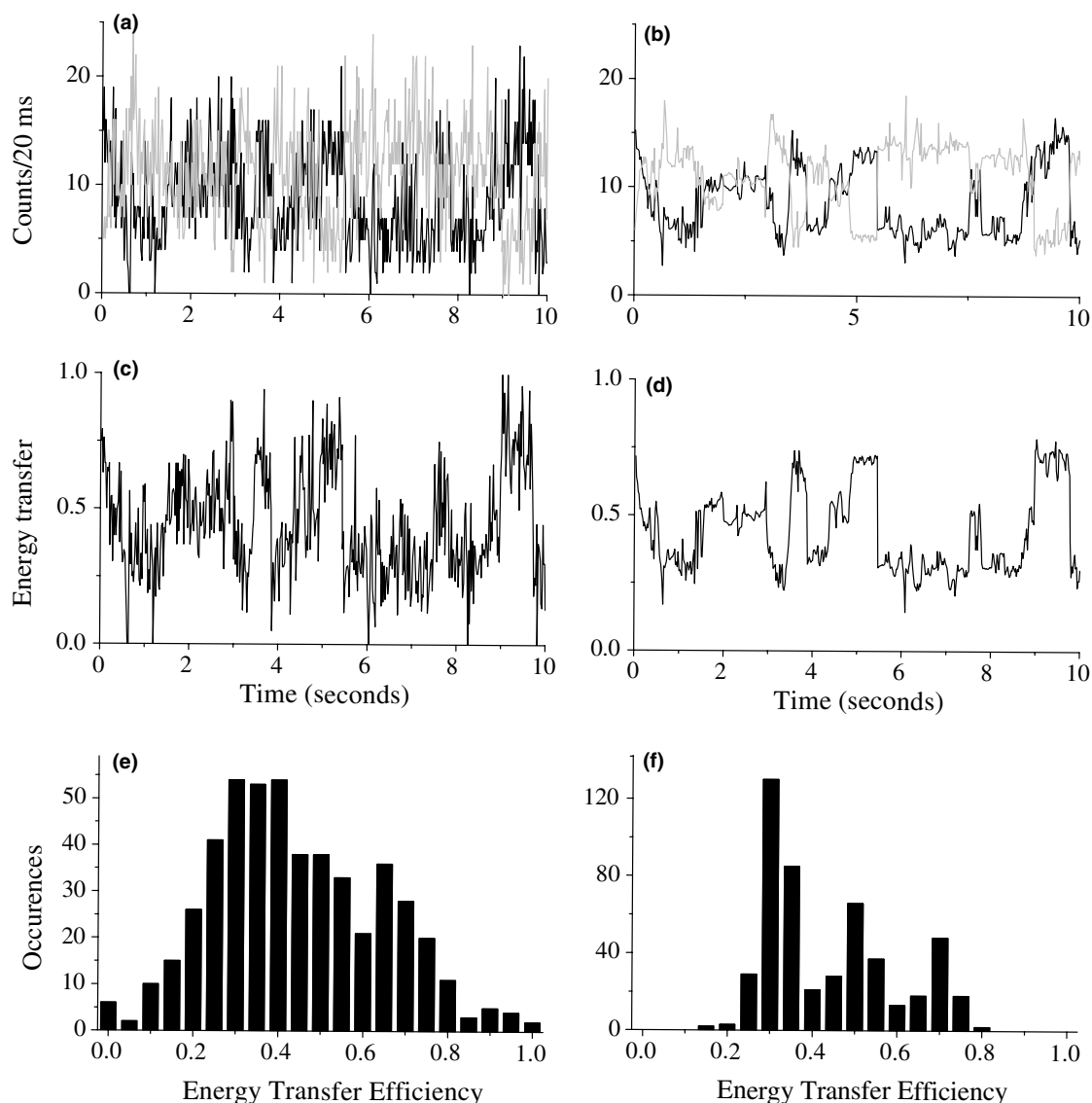


Fig. 4. Simulation of a three-state single-molecule protein folding experiment, in which the FRET value changes abruptly between 0.3, 0.5 and 0.7. The overall count rate is 1000 Hz. (a) Simulated data. Acceptor in black and donor in gray. (b) Simulated data after filtration with the NLF. (c) FRET efficiency calculated from (a). (d) FRET efficiency calculated from (b). (e) Histogram of the FRET efficiency values of (c). (f) Histogram of the FRET efficiency values of (d).

line) and after filtration with an RAF (green line). The size of the averaging window used in the latter, 16 points, is four times shorter than the longest window of the former. The green line is somewhat smoother than the black line. However, all transitions are significantly smeared by the RAF, and in many cases their accurate position cannot be inferred from treated data. Further, some cases might involve an erroneous assignment of the state (see the signal at ~ 2.5 s). These errors are particularly likely to occur when the dwell times in various states become shorter and shorter. The NLF seems to be significantly more immune towards such errors, and allows an accurate determination of dwell

times in the various states. Moreover, the NLF incorporates several window sizes at once and therefore makes the search for the optimal window size faster and easier.

6. Obtaining information about transition-state passage times

A key piece of information which is of high interest in current protein folding studies is the shape of the transition state region on the energy landscape. While computer simulations might shed some light on this is-

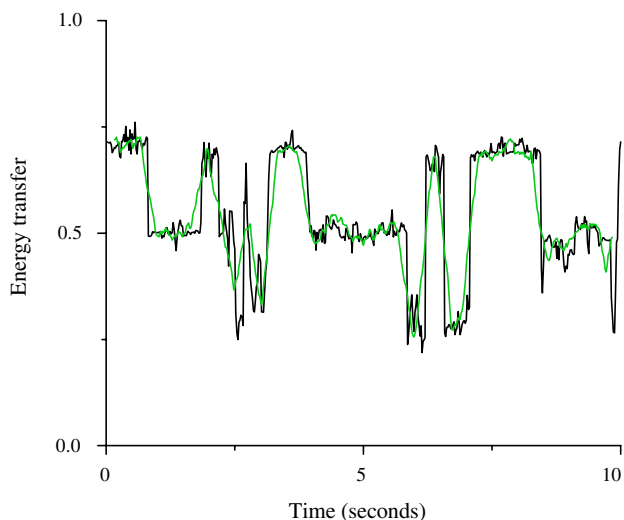


Fig. 5. Comparison of the NLF with a running-average filter. Simulated data was generated with three FRET values, 0.3, 0.5 and 0.7 and an average count rate of 2500 Hz. The FRET efficiency curve calculated from the data was averaged with the NLF with the same parameters as above (black curve) or with a running-average filter with a window of 16 points (green curve). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sue, it is very difficult to directly probe it experimentally. In fact, it can be argued that in principle only single-molecule techniques can confront this problem. This is because molecular motion through the transition state is totally uncorrelated in the ensemble, so that each molecule traverses the folding barrier at a different time. Following the trajectory of a single molecule, on the other hand, should allow probing the time of passage through the transition state. A major obstacle towards achieving this goal is the fact that this time might be very short, of the order of microseconds or less. Single-molecule measurements cannot directly sustain the time resolution required for viewing such fast dynamics. However, we suggest here that by using the NLF it is possible in principle to determine an upper limit for the transition state crossing time.

We simulate FRET trajectories with a high photon count, 50,000 Hz. This high count is achievable, although it requires a high laser power, which will lead to fast photobleaching and short overall trajectories. Nevertheless, if the folding/unfolding rate is large enough, transitions will occur at least in a fraction of the trajectories, enough for the analysis suggested here. The simulated trajectories involve either very fast folding transitions (Fig. 6) or transitions in which the FRET efficiency changes with an exponential time dependence involving a time-constant of 0.4 ms (Fig. 7). The trajectories are binned in 0.01 ms bins. The average photon count number with such short bins is just 0.5 photons/bin. However, after treating the trajectories with the

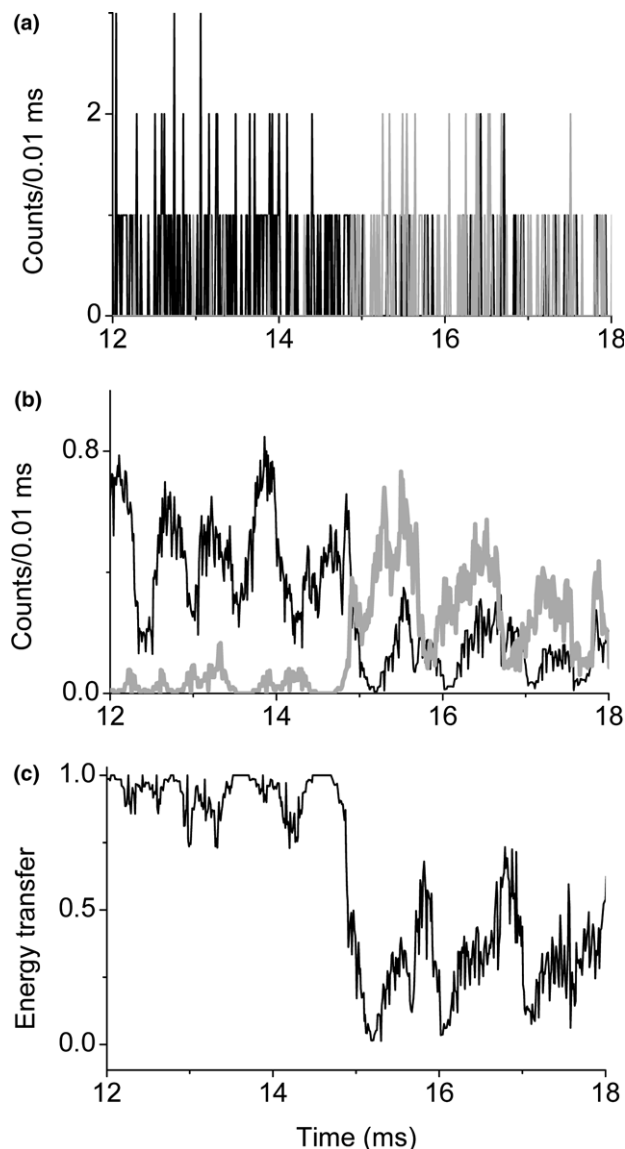


Fig. 6. NLF analysis of an “instantaneous” unfolding transition in a simulated trajectory binned with 0.01 ms bins. (a) Binned data, acceptor in black and donor in gray. (b) Data after operation with the NLF with the following parameters: windows of 4, 8 and 16 points, $p = 1$, $M = 20$. (c) FRET efficiency calculated from the data in (b). The transition appears as a sharp drop in the efficiency.

NLF the difference between the two types of transitions becomes evident. While the fast jump appears as a single-point change in the filtered FRET efficiency function of Fig. 6(c), a gradual change is seen in the equivalent function in Fig. 7(c). One can conclude that the crossing time in Fig. 6 is of the order of the binning time, 0.01 ms, or shorter. A refined version of this procedure might use the accurate values of transition times obtained from the NLF-treated trajectories to superimpose pieces of experimental trajectories showing transitions and average them to obtain more accurate estimates for barrier crossing times.

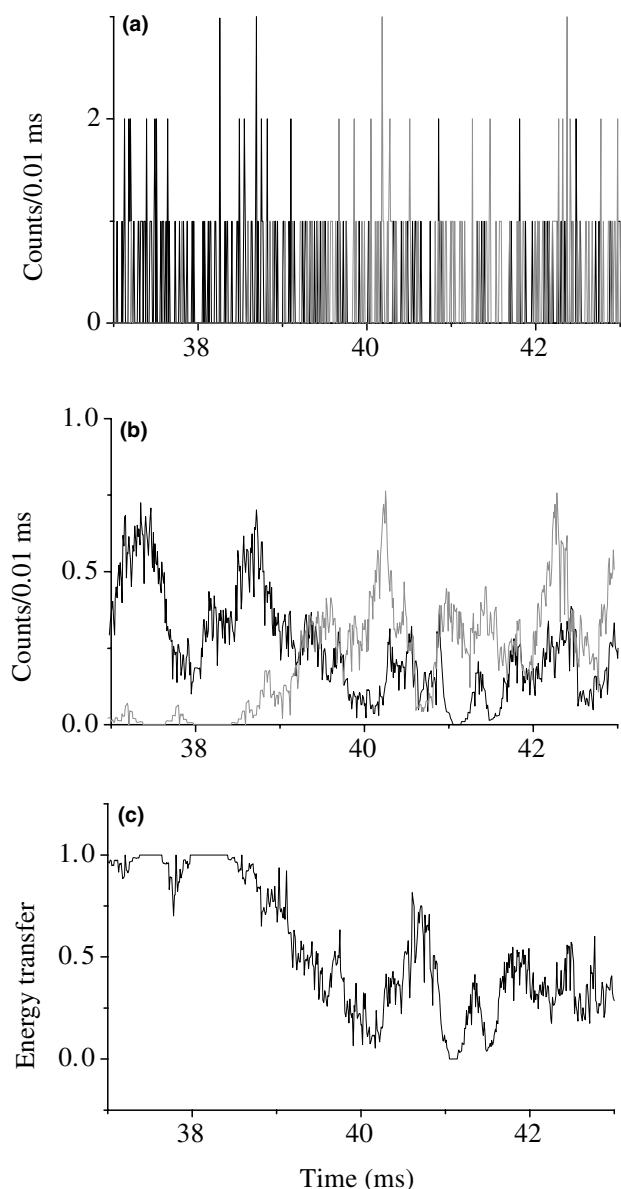


Fig. 7. NLF analysis of a gradual unfolding transition in a simulated trajectory binned with 0.01 ms bins. The gradual transition was obtained by assuming that the FRET efficiency decayed exponentially from one value to another, with a time constant of 0.4 ms. (a) Binned data, acceptor in black and donor in gray. (b) Data after operation with the NLF with the same parameters as in Fig. 6. (c) FRET efficiency calculated from the data in (b). The gradual character of the transition is recovered.

7. Conclusions

We have described here a method to significantly reduce the noise in single-molecule trajectories, while retaining sharp features in the data which are related to transitions between states of a molecule. This property of the method is unique, since all the familiar linear filters (such as the RAF discussed above) will tend to smear sharp features. The NLF is thus particularly ap-

pealing for the analysis of folding trajectories of individual protein molecules, where fast jumps from folded to unfolded states are expected. As seen, the NLF provides a route to easily and accurately obtain the number of states populated in a trajectory, even when the difference in signal (FRET efficiency) between them is rather small. This advantage of the filter has been exploited by Rhoades et al. [9].

An obvious disadvantage of the NLF, which is shared by many other filters, is that it does reduce the bandwidth in stretches of data that do not contain sharp state-to-state transitions. One should therefore be careful in selecting the right parameters for the filter depending on which data features are to be retained. The availability of several averaging windows at the same time facilitates this task. If it is suspected that the data contains fast dynamics, perhaps of a continuous nature [8] (i.e., not between a discrete set of states) it is probably better to resort to correlation function analysis. Such analysis can in principle also yield information about transition rates if enough transitions occur in each trajectory. However, as in ensemble relaxation studies, it will not provide the separate rates but a sum of them, as opposed to a direct analysis of dwell times in the various states in a trajectory [18]. Finally, filtration methods might seem of little use when data of higher signal-to-noise ratio is available. However, as shown in the previous section, more ambitious applications, such as the unraveling of transition state dynamics, might still make use of such techniques.

Acknowledgements

Paul Barbara and his colleagues are gratefully acknowledged for exciting discussions during a recent visit to Austin. This work was supported by a grant from the Israel Science Foundation. G.H. is the incumbent of the Benjamin H. Swig and Jack D. Weiler career development chair.

References

- [1] G. Haran, *J. Phys.: Condens. Matter* 15 (2003) R1.
- [2] A.A. Deniz, T.A. Laurence, G.S. Beligere, M. Dahan, A.B. Martin, D.S. Chemla, P.E. Dawson, P.G. Schultz, S. Weiss, *Proc. Natl. Acad. Sci. USA* 97 (2000) 5179.
- [3] B. Schuler, E.A. Lipman, W.A. Eaton, *Nature* 419 (2002) 743.
- [4] X. Zhuang, H. Kim, M.J.B. Pereira, H.P. Babcock, N.G. Walter, S. Chu, *Science* 296 (2002) 1473.
- [5] L.E. Bartley, X. Zhuang, R. Das, S. Chu, D. Herschlag, *J. Mol. Biol.* 328 (2003) 1011.
- [6] G. Bokinsky, D. Rueda, V.K. Misra, M.M. Rhodes, A. Gordus, H.P. Babcock, N.G. Walter, X. Zhuang, *Proc. Natl. Acad. Sci. USA* 100 (2003) 9302.

- [7] Y. Jia, D.S. Talaga, W.L. Lau, H.S.M. Lu, W.F. DeGrado, R.M. Hochstrasser, *Chem. Phys.* 247 (1999) 69.
- [8] D.S. Talaga, W.L. Lau, H. Roder, J.Y. Tang, Y.W. Jia, W.F. DeGrado, R.M. Hochstrasser, *Proc. Natl. Acad. Sci. USA* 97 (2000) 13021.
- [9] E. Rhoades, E. Gussakovsky, G. Haran, *Proc. Natl. Acad. Sci. USA* 100 (2003) 3197.
- [10] P.R. Selvin, *Nat. Struct. Biol.* 7 (2000) 730.
- [11] S.E. Jackson, *Fold. Des.* 3 (1998) r81.
- [12] E. Rhoades, M. Cohen, B. Schuler, G. Haran (submitted).
- [13] X.W. Zhuang, L.E. Bartley, H.P. Babcock, R. Russell, T.J. Ha, D. Herschlag, S. Chu, *Science* 288 (2000) 2048.
- [14] H.P. Lu, L. Xun, X.S. Xie, *Science* 282 (1998) 1877.
- [15] B. Sakmann, E. Neher, *Single Channel Recording*, Plenum Press, New York, 1995.
- [16] S.H. Chung, R.A. Kennedy, *J. Neurosci. Meth.* 40 (1991) 71.
- [17] M. Andreć, R.M. Levy, D.S. Talaga, *J. Phys. Chem. A* 107 (2003) 7454.
- [18] R. Verberk, M. Orrit, *J. Chem. Phys.* 119 (2003) 2214.