Diffusion dynamics, moments, and distribution of first-passage time on the protein-folding energy landscape, with applications to single molecules

Chi-Lun Lee,[1] Chien-Ting Lin,[2] George Stell,[2,*] and Jin Wang[3,2,4,5,*]

[1]*Department of Physics, State University of New York at Stony Brook, Stony Brook, New York 11794*

[2]*Department of Chemistry, State University of New York at Stony Brook, Stony Brook, New York 11794*

[3]*State Key Laboratory of Electroanalytical Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130021, People's Republic of China*

[4]*Department of Physics, National University of Singapore, 2 Science Drive 3, Singapore 117542*

[5]*Global Strategic Analytics Unit, Citigroup, One Huntington Quadrangle, Suite 1N16, Melville, New York 11747*

(Received 2 September 2001; revised manuscript received 17 September 2002; published 17 April 2003)

We study the dynamics of protein folding via statistical energy-landscape theory. In particular, we concentrate on the local-connectivity case with the folding progress described by the fraction of native conformations. We found that the first passage-time (FPT) distribution undergoes a dynamic transition at a temperature below which the FPT distribution develops a power-law tail, a signature of the intermittent nonexponential kinetic phenomena for the folding dynamics. Possible applications to single-molecule dynamics experiments are discussed.

DOI: 10.1103/PhysRevE.67.041905 PACS number(s): 87.15.−v, 05.40.−a, 82.37.Np

The study of diffusion along a complex energy landscape is a very important issue for many fields. In the field of protein folding, the crucial question is how the many possible configurations of a polypeptide chain dynamically converge to a particular folded state [1]. Clearly, a statistical description is needed for a large number of configurational states. According to the energy-landscape theory of protein folding [3–5], there exists a global bias of the energy landscape towards the folded state due to natural evolution selection. Superimposed on this is the fluctuation or roughness of the energy landscape coming from competing interactions of the amino-acid residues. The folding energy landscape is like a funnel, and there are, in general, multiple routes towards the folded state. Discrete paths emerge only when the underlying energy landscape becomes rough, and local traps (minima) start to appear. If the energy landscape is smooth, the average diffusion time is a good parameter to describe such dynamical processes. On the other hand, if the energy landscape is rough, there exist large fluctuations in the energies, and the diffusion time is expected to fluctuate very much around its mean. In this case one needs to know the full distribution of the diffusion time in characterizing the folding process.

It is now possible to measure the reaction and folding dynamics of individual molecules *in vitro* [6]. On complex systems such as biomolecules, reactions in general do not obey exponential laws, and activation processes often do not follow the simple Arrhenius relation. However, these phenomena usually cannot be clarified by typical bulk measurements. The study of statistics for individual molecular reaction events can probe these reaction processes with much more subtlety [7]. The information on the diffusion-time distribution provides a way to help unravel the fundamental mechanism of single molecule reactions. Previously many works have focused on the average rate behavior [8,9], whereas very few physical studies and discussions addressed on the whole distribution of the folding time. In this paper, we concentrate on the statistics and distributions of the first passage time (FPT) for the folding dynamics. A dynamic transition temperature is found above which the FPT distribution is Poisson-like, and below which the distribution develops a power-law tail, where non-self-averaging behavior in kinetics emerges. The physical reasons for the power-law behavior is given in addition to the rigorous derivation. Further discussions will be on the need for cooperative multibody interactions in order for the protein-folding driving force to be consistent with available experiments. In this study, we will briefly summarize the theoretical framework and results, concentrate on the discussion of the underlying physics, and leave the rather lengthy and tedious mathematical details to a long paper [2].

The framework we adopt here was first introduced by Bryngelson and Wolynes [3]. The problem of folding dynamics is characterized by random walks on a funnel-like energy landscape with roughness. In this model, the energy landscape is generated by the random-energy model [10], which assumes that interactions among non-native states are random variables with given probability distributions. For this model there are N residues in a polypeptide chain. For each residue there are $\nu+1$ available conformational states, one being the native state. A simplified version of the polypeptide chain energy is expressed as

$$E=-\sum \epsilon_i(\alpha_i)-\sum J_{i,i+1}(\alpha_i,\alpha_{i+1})-\sum K_{i,j}(\alpha_i,\alpha_j), \tag{1}$$

where the summation indices i and j are labels for amino-acid residues, and α_i is the state of ith residue. The three terms represent the one-body potential, two-body interactions for nearest-neighbor residues in sequence, and interactions for residues close in space but not in sequence, respectively. In this random-energy model the energies for non-native states and interactions are replaced by random

*Author to whom correspondence should be addressed.

variables with Gaussian distributions. Along with the assumption that energies for different configurations are uncorrelated, one can easily generate an energy landscape with roughness tuned by the spreads of these probability distributions. Using a microcanonical ensemble analysis, the average free energy and thermodynamic properties of the polypeptide chain can be obtained [10]. Note that the polymer connectivity is embodied in the entropy calculations.

In this study, we use the fraction of native conformations ρ as an order parameter to represent the folding progress. The system is assumed to be in quasiequilibrium with respect to ρ, and the states are kinetically locally connected. In this way the dynamics evolves continuously with ρ. With this assumption we exclude from our current study the class of fast-folding proteins that do not have local connectivity and can fold in discrete steps. The kinetic process is approximated via the use of Metropolis dynamics. Therefore, the transition rate from one conformation state to a neighboring state is determined by the energy difference of these two states, and an overall constant R_0 characterizes the time scale of residue interactions. Readers are referred to Ref. [2,11] for detailed derivations of the dynamic equations. In brief, the statistical energy landscape is first categorized by the order parameter ρ, along which an energy distribution function is derived via Eq. (1). With the use of Metropolis dynamics, one can obtain expressions for the waiting-time distribution function and also the rate distribution $P(R,\rho)$ for transitions between successive ρ's. Finally, using continuous-time random walks (CTRW), the following generalized Fokker-Planck equation in the Laplace-transformed space can be obtained [3]:

$$s\tilde{G}(\rho,s)-n_i(\rho)=\frac{\partial}{\partial\rho}\left\{D(\rho,s)\left[\tilde{G}(\rho,s)\frac{\partial}{\partial\rho}U(\rho,s)+\frac{\partial}{\partial\rho}\tilde{G}(\rho,s)\right]\right\}, \tag{2}$$

where $U(\rho,s)\equiv F(\rho)/T+\log[D(\rho,s)/D(\rho,0)]$, and

$$F(\rho)=N\left[-\left(\delta\epsilon-\frac{\Delta\epsilon^2}{2T}\right)\rho-\left(\delta L-\frac{\Delta L^2}{2T}\right)\rho^2+T\rho\log\rho+T(1-\rho)\log\frac{1-\rho}{\nu}\right]. \tag{3}$$

In Eq. (2) s is the Laplace transform variable over time τ. $\tilde{G}(\rho,s)$ is the Laplace transform of $G(\rho,\tau)$, the probability density function. $G(\rho,\tau)d\rho$ gives the probability for a polypeptide chain to stay between ρ and $\rho+d\rho$ at time τ. $n_i(\rho)$ is the initial condition for $G(\rho,\tau)$. $F(\rho)$ is the average free energy for the polypeptide chain. T is a scaled temperature, $\nu+1$ is the number of conformational states of each residue, and $\delta\epsilon$ and δL are energy differences between the native and average non-native states for one- and two-body interactions, respectively. $\Delta\epsilon$ and ΔL are energy spreads of one- and two-body non-native interactions. Note that the two-body energies δL and ΔL include contributions from the second and third term in Eq. (1). $D(\rho,s)$ is the frequency-dependent diffusion coefficient [3]:

$$D(\rho,s)\equiv\left(\frac{\lambda(\rho)}{2N^2}\right)\left\langle\frac{R}{R+s}\right\rangle_R(\rho)\Big/\left\langle\frac{1}{R+s}\right\rangle_R(\rho), \tag{4}$$

where $\lambda(\rho)\equiv 1/\nu+(1-1/\nu)\rho$. The average $\langle\rangle_R$ is taken over $P(R,\rho)$, the probability distribution function of transition rate R from one state with order parameter ρ to its neighboring states, which may have order parameters equal to $\rho-1/N$, ρ, or $\rho+1/N$. The explicit expression of $P(R,\rho)$ can be found in Ref. [2]. The boundary conditions for Eq. (2) are set as a reflecting one at $\rho=0$ and an absorbing one at $\rho=\rho_f$. The choice of an absorbing boundary condition at $\rho=\rho_f$ facilitates our calculation for the first passage-time distribution. One can also rewrite Eq. (2) in its integral-equation representation by integrating it twice over ρ:

$$\tilde{G}(\rho,s)=-\int_\rho^{\rho_f}d\rho'\int_0^{\rho''}d\rho''[s\tilde{G}(\rho'',s)-n_i(\rho'')]\times\frac{\exp[U(\rho',s)-U(\rho,s)]}{D(\rho',s)}. \tag{5}$$

In this work we mainly study the behavior of the FPT for the order parameter to reach ρ_f. This FPT is analogous to the folding time in single-molecule experiments when the temperature is below the folding transition. One can calculate the FPT distribution function $P_{\mathrm{FPT}}(\tau)$ via

$$P_{\mathrm{FPT}}(\tau)=\frac{d}{d\tau}(1-\Sigma)=-\frac{d\Sigma}{d\tau}, \tag{6}$$

where $\Sigma(\tau)\equiv\int_0^{\rho_f}d\rho\,G(\rho,\tau)$. The nth moment of the FPT distribution function can be calculated via $\langle\tau^n\rangle=n!(-1)^{n-1}\int_0^{\rho_f}d\rho\,G_{n-1}(\rho)$, where $\tilde{G}(\rho,s)=G_0(\rho)+sG_1(\rho)+s^2G_2(\rho)+\cdots$. By series expanding Eq. (5) with respect to s, we can solve for $G_n(\rho)$ and therefore $\langle\tau^n\rangle$ iteratively by matching the coefficients of s^n. One can also solve for $\tilde{G}(\rho,s)$ directly from Eq. (5), using a linear inversion technique. Observing that $\tilde{P}_{\mathrm{FPT}}(s)=1-s\tilde{\Sigma}(s)$, where $\tilde{P}_{\mathrm{FPT}}(s)$ and $\tilde{\Sigma}(s)$ are the Laplace transforms of $P_{\mathrm{FPT}}(\tau)$ and $\Sigma(\tau)$, respectively, one can calculate $\tilde{P}_{\mathrm{FPT}}(s)$ and therefore investigate the behavior of $P_{\mathrm{FPT}}(\tau)$.

From this model, the resulting energy landscape is funnel-like, with numerous configurations categorized by the degree of folding, and also with built-in roughness and a bias toward the native state, both of which can be controlled using parameters $\Delta\epsilon$, ΔL, $\delta\epsilon$, and δL. These features render it possible to mimic the energy landscape qualitatively for a large class of proteins. We start the numerical calculations by setting $R_0=10^9s^{-1}$, $N=100$, and $\nu=10$ to match the physical scales. In this work we only study single-domain model proteins. For proteins of larger size ($N>100$) they tend to form multiple domains where spatial variations need to be taken into account. This is beyond the scope of the current mean-field approximation. For simplicity we assume $\delta\epsilon=\delta L$ and

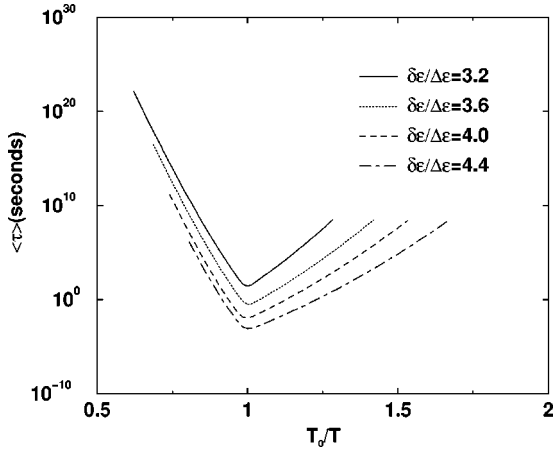

FIG. 1. MFPT vs reduced inverse temperature T_0/T for various $\delta\epsilon/\Delta\epsilon$.

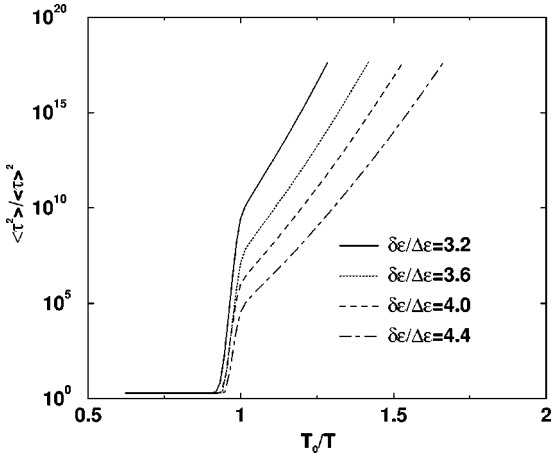

FIG. 2. $\langle\tau^2\rangle/\langle\tau\rangle^2$ vs reduced inverse temperature T_0/T for various $\delta\epsilon/\Delta\epsilon$.

$\Delta\epsilon=\Delta L$. Therefore, the ratio of the energy gap between the native state and the average of non-native states over the spread of non-native states, $\delta\epsilon/\Delta\epsilon$, becomes an appropriate parameter, representing the importance of gap bias towards the folded state relative to the roughness of the landscape. One can show that only the relative ratios among $\delta\epsilon$, $\Delta\epsilon$, and T are the controlling parameters in this problem. We set the initial distribution of the polypeptide chain molecules to be $n_i(\rho)=\delta(\rho-\rho_i)$, where ρ_i is set to be 0.05. In our calculations we set $\rho_f=0.9$. This means that 90% of the amino-acid residues are in their native states.

The mean first passage time (MFPT) $\langle\tau\rangle$ for the folding process versus a scaled inverse temperature T_0/T is plotted in Fig. 1 for various settings of the parameter $\delta\epsilon/\Delta\epsilon$. We have an inverted bell-like curve for each fixed $\delta\epsilon/\Delta\epsilon$, and the MFPT reaches its minimum at a temperature T_0. At high temperatures, the MFPT is large although the diffusion process itself is fast [i.e., $D(\rho,s)$ is large]. This long-time folding behavior is due to the instability of the folded state. The MFPT is also large at low temperature, which indicates that the polypeptide chain is trapped in low-energy non-native states. This is in agreement with simulation studies [8].

By comparing the MFPT minimum for various $\delta\epsilon/\Delta\epsilon$, one finds that this minimum becomes smaller when the ratio of the energy gap versus roughness increases. This can be easily understood. As the bias towards the native state increases, it is easier to overcome the local traps to reach the folded state. On the other hand, if the barrier for local traps is shallower, then the folding process towards the native state will be faster. This suggests that a possible kinetic criterion for selecting the subset of the whole sequence space leading to well-designed fast folding protein is to maximize $\delta\epsilon/\Delta\epsilon$. In other words, one has to choose the sequence subspace such that the global bias overwhelms the roughness of the energy landscape [4,9,14].

We also calculate higher-order moments for the FPT distribution. In Fig. 2 we show the behavior of the reduced second moment, $\langle\tau^2\rangle/\langle\tau\rangle^2$. We find that the reduced second moment starts diverging at a temperature around and below T_0, where the MFPT is at its minimum. This is an indication of a long tail in the FPT distribution. The divergence of the second moment also shows that the dynamics exhibits non-self-averaging behavior. This can be explained by the fact that at high temperature the folding process is less sensitive to local traps, and the energy landscape is smoothed out. On the other hand, as the temperature becomes lower, local traps start to make significant contributions, resulting in non-self-averaging behavior for different paths towards the folded state. This non-self-averaging behavior can be observed through large fluctuations in the FPT distribution.

From the study of higher moments, we find $\langle\tau^n\rangle \approx n!\langle\tau\rangle^n$ when $T>T_0$. This indicates a Poisson distribution in the FPT distribution with an exponential tail at large time. On the other hand, when the temperature is well below T_0, we find a stretch-exponential form in $\tilde{P}_{\mathrm{FPT}}(s)$ over several orders of magnitude by solving Eq. (5) directly. This indicates that at low temperature $P_{\mathrm{FPT}}(\tau)$ can be approximated by a Lévy distribution, which has $P_{\mathrm{FPT}}(\tau)\sim\tau^{-(1+\alpha)}$ for large τ. In Fig. 3 we plot the exponent α versus T_0/T for the case $\delta\epsilon/\Delta\epsilon=4.0$. We find that α is decreasing when the temperature is lowered. On the other hand, as T gets closer to

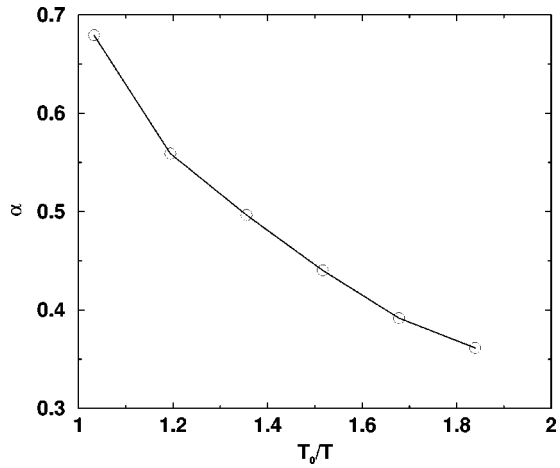

FIG. 3. The exponent α vs T_0/T in the case $\delta\epsilon/\Delta\epsilon=4.0$ when $T<T_0$. Below T_0 the FPT distribution starts to develop a power-law tail $P_{\mathrm{FPT}}(\tau)\sim\tau^{-(1+\alpha)}$ at large τ.

T_0, the power-law feature starts to break down and $P_{\mathrm{FPT}}(\tau)$ gradually resumes exponential behavior. The power-law behavior seen here can be understood [11] by noticing that as the temperature is close to that of a partially frozen energy state, we can expand the energy around this partially frozen energy. Since the density of states in terms of energy has a Gaussian distribution in our study, the expansion is equivalent to the linearization on the exponentials of density of states. In other words, the density of states approaches an exponential distribution $\exp(E/T_c)$, where T_c is the partially frozen temperature close to T_0. Since the transition state for folding is exponentially related to the energy barrier, $\exp(-E/T)$, this results in a power-law distribution in the transition rate as well as the folding time $f(\tau) \sim 1/\tau^{(T/T_c)+1}$. As we can see, $\alpha \sim T/T_c$ giving a physical reason to the explanation of the numerical results obtained above. As the temperature is lowered, α decreases and therefore the power-law tail is even fatter.

From the results above, we find that for a fixed-energy landscape, there exists a dynamic transition temperature T_0. When the temperature is above T_0, the FPT distribution is Poisson, indicating exponential kinetics, and in random-walk language we have normal diffusion on the energy landscape. Below T_0, the variance and higher moments diverge, and the FPT distribution shows a power-law decay behavior, exhibiting signs of anomalous diffusion. This indicates the process is non-self-averaging and distinct folding pathways emerge at various time scales. As a comparison, we have calculated the folding-transition temperature T_f by identifying the crossing point of folding and unfolding rate versus temperature. We found that T_f is close to and slightly higher than T_0 for various settings of $\delta\epsilon/\Delta\epsilon$. This indicates that the equilibrium and dynamic properties in proteins are strongly correlated. Recent simulation and experimental results [12,13] also indicate that T_0 is close to and slightly smaller than T_f. But the turnaround near T_0 has been shown to be smoother than that observed in our current work. This is partly due to the insufficiency of cooperative interactions in our study [12]. Since the dominant driving force towards folded proteins is believed to be the hydrophobic force, which has a cooperative multibody-interaction nature, one has to include the multibody interactions in order to improve the current model. When we included the three-body as well as the two-body interactions, we found that T_0 is still close to but slightly smaller than the folding temperature. Furthermore, the sharp turnover behavior near T_0 in the MFPT versus temperature is smoothed out. Instead of the sharp V shape in the absence of the multibody interactions (three-body or higher), the MFPT has a smoother U shape versus temperature around T_0 in the presence of the multibody force (three). This implies the necessity of the cooperative nature for the folding driving force. More details on this will be discussed in a future publication.

In single-molecule folding experiments, it is now possible to measure not only the mean but also the fluctuation and moments as well as the distribution of folding time [15]. Under different experimental and sequence conditions, one can observe different behavior for the folding time and its distributions. A well-designed fast folding sequence with suitable experimental condition exhibits self-averaging and simple rate behavior. Multiple routes are parallel and lead to folding. A less well-designed sequence (with larger $\Delta\epsilon$) folds slowly and often exhibit non-self-averaging nonexponential rate behavior, indicating the existence of intermediate states or local traps. In this case, the folding process is sensitive to which kinetic path it takes, since a slight change in a folding pathway may cause large fluctuation in the folding time, which indicates intermittency. From the single-molecule experiments the fundamental mechanisms and intrinsic features of the folding process may be unravelled. In typical bulk-molecule measurements, it is often hard to observe and analyze intermittent phenomena, because the dynamics is averaged over numerous molecules. And furthermore, one cannot tell if the bulk phenomena are either from the intrinsic features of individual molecules or the inhomogeneous average over the molecules.

It is worth mentioning that although we focus on the study of the protein-folding problem in this paper, the approach we use here is very general for treating problems with barrier crossings on a complex energy landscape. In fact, several experiments on glasses, spin glasses, viscous liquids, and conformational dynamics already show the existence of nonexponential distributions at low temperature. In particular, a recent experiment on single-molecule enzymatic dynamics [16] shows explicitly the Lévy-like distribution for the on-time relaxation of the underlying complex protein energy landscape. The essence of this model lies in building an energy landscape with a probabilistic approach, which is often applicable for complex, disordered systems. To improve the model, it is necessary to gain better knowledge for the statistical distributions on the energy landscape. Finally, an interesting study on anomalous diffusion has been made recently, using a fractional Fokker-Planck equation (FFPE) [17] to describe dynamic processes characterized by the Lévy distributions. Our results show that the approach we use here can serve as a microscopic basis for the use of such a FFPE.

The authors thank X. S. Xie for useful discussions. C.-L. Lee and G. Stell gratefully acknowledge the support of the Division of Chemical Sciences, Office of Basic Energy Sciences, and Office of Energy Research, U.S. Department of Energy.

[1] C. Levinthal, in *Proceedings in Mossbauer Spectroscopy in Biological Systems*, edited by P. Debrunner, J. Tsubris, and E. Munck (University of Illinois Press, Urbana, 1969), p. 22.

[2] C.-L. Lee, G. Stell, and J. Wang, J. Chem. Phys. **118**, 959 (2003).

[3] J.D. Bryngelson and P.G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).

[4] V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich, J. Chem. Phys. **101**, 6052 (1994).

[5] H.S. Chan and K.A. Dill, J. Chem. Phys. **100**, 9238 (1994);

J.D. Bryngelson, J.O. Onuchic, N.D. Socci, and P.G. Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995); J. Wang, J.O. Onuchic, and P.G. Wolynes, Phys. Rev. Lett. **76**, 4861 (1996).

[6] W.E. Moerner, Accounts of Chemical Research **29**, 563 (1996); H.P. Lu, L. Xun, and X.S. Xie, Science **282**, 1877 (1998).

[7] J.N. Onuchic, J. Wang, and P.G. Wolynes, Chem. Phys. **247**, 175 (1999); J. Wang and P.G. Wolynes, Phys. Rev. Lett. **74**, 4317 (1995); J. Chem. Phys. **110**, 4812 (1999).

[8] A.M. Gutin, V.I. Abkevich, and E.I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996); M. Cieplak, T.X. Hoang, and M.S. Li, *ibid.* **83**, 1684 (1999).

[9] F. Seno, C. Micheletti, A. Maritan, and J.R. Banavar, Phys. Rev. Lett. **81**, 2172 (1998); D.K. Klimov and D. Thirumalai, J. Chem. Phys. **109**, 4119 (1998).

[10] B. Derrida, Phys. Rev. B **24**, 2613 (1981).

[11] M. Mezard, E. Parisi, and M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific Press, Singapore, 1986).

[12] H. Kaya and H.S. Chan, Phys. Rev. Lett. **85**, 4823 (2000); J. Mol. Biol. **315**, 899 (2002).

[13] B. Kuhlman *et al.*, J. Mol. Biol. **284**, 1661 (1998).

[14] R.A. Goldstein, Z.A. Luthey-Schulten, and P.G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **89**, 4918 (1992).

[15] Y. Jia *et al.*, Chem. Phys. **247**, 69 (1999); X. Zhuang *et al.*, Science **288**, 2048 (2000); A.A. Deniz *et al.*, Proc. Natl. Acad. Sci. U.S.A. **97**, 5179 (2000).

[16] H. Yang and X.S. Xie, Chem. Phys. **284**, 4231 (2002); J. Chem. Phys. **117**, 10965 (2002).

[17] E. Barkai, R. Metzler, and J. Klafter, Phys. Rev. E **61**, 132 (2000).