How to accelerate protein search on DNA: Location and dissociation

Anatoly B. Kolomeisky and Alex Veksler
Department of Chemistry, Rice University, Houston, Texas 77005, USA

(Received 2 February 2012; accepted 7 March 2012; published online 28 March 2012)

One of the most important features of biological systems that controls their functioning is the ability of protein molecules to find and recognize quickly specific target sites on DNA. Although these phenomena have been studied extensively, detailed mechanisms of protein-DNA interactions during the search are still not well understood. Experiments suggest that proteins typically find their targets fast by combining three-dimensional and one-dimensional motions, and most of the searching time proteins are non-specifically bound to DNA. However these observations are surprising since proteins diffuse very slowly on DNA, and it seems that the observed fast search cannot be achieved under these conditions for single proteins. Here we propose two simple mechanisms that might explain some of these controversial observations. Using first-passage time analysis, it is shown explicitly that the search can be accelerated by changing the location of the target and by effectively irreversible dissociations of proteins. Our theoretical predictions are supported by Monte Carlo computer simulations.
 [http://dx.doi.org/10.1063/1.3697763]

I. INTRODUCTION

In biological systems most processes start when some protein molecules bind to specific target sequences on DNA molecules to initiate a cascade of biochemical reactions.[1] This fundamental aspect of protein-DNA interactions has been studied extensively by various experimental[2–14] and theoretical methods.[5,9,10,15–29] Although a significant progress in explaining protein search phenomena has been made, detailed mechanisms remain not fully understood.[10,26] Furthermore, there are strong theoretical debates on how to explain fast protein search for the targets on DNA, which is also known as a facilitated diffusion.[5,9,10,26]

Large amount of experimental evidences, coming mostly from single-molecule measurements,[6–8,12] suggest that protein search is a complex dynamic phenomenon consisting of three-dimensional (in the solution) and one-dimensional (on the DNA) modes. But the most paradoxical observation is that protein molecules spend most of the search time ($\geq$90 − 99%) on the DNA chain where they diffuse very slowly.[7,8,12] It is not clear then how the fast search can be achieved in this case. Several theoretical ideas that point out to the role of lowering dimensionality,[3–5,10,15,16,21] electrostatic effects,[9] correlations between 3D and 1D motions,[17,26,27] transitions between different chemical states,[12,28] bending fluctuations, and hydrodynamics[25] have been proposed. However, a comprehensive theoretical description is still not available, especially for the case when concentration of proteins is relatively small. In this letter, we propose and investigate two possible mechanisms that might accelerate one-dimensional search of proteins for specific targets on DNA. Using explicit calculations via first-passage analysis, it is argued that optimal location of the target site as well as effectively irreversible dissociations of protein molecules from the DNA segments might strongly lower the overall search time.

II. THEORETICAL MODEL

We consider a simple model for a search where one protein molecule diffuses along the DNA chain while scanning for the target as shown in Fig. 1. As 3D excursions to the solution are very fast, we concentrate here on analyzing only the rate-limiting 1D contributions to the overall facilitated target search. The DNA segment has L binding sites, and one of them at the position m is the target for the protein molecule. At time $t = 0$, the protein molecule starts the searching procedure with equal probability at any site of the DNA chain. The protein molecule can diffuse forward/backward with the rate u, and it also might dissociate irreversibly with the rate k (see Fig. 1). These rates are connected to the protein hopping barriers along and away from the DNA molecule. For *lac* repressor proteins from experimental data,[7,8] one could estimate these rates as following: $u \simeq 10^3 - 10^6$ s^{-1} and $k \simeq 200 - 3000$ s^{-1}. Note that for a given DNA segment, these dissociations are only effectively irreversible since after leaving the DNA segment the protein most likely will bind to other DNA segments. The protein molecule diffuses so fast in the solution that the detachment and reattachment locations can be viewed as non-correlated.[30]

A. Target position

First, we will consider the role of the target position on the mechanisms of protein search by neglecting dissociations, i.e., for the case when $k = 0$. Let us define a function $F_n(t)$ as a first-passage probability to reach the target, if at $t = 0$, the protein was at the site n ($n = 1, 2, \ldots, L$). Temporal evolution of this quantity can be described by backward master equations,[31]

$$\frac{dF_n(t)}{dt} = u[F_{n+1}(t) + F_{n-1}(t)] - 2uF_n(t), \qquad (1)$$

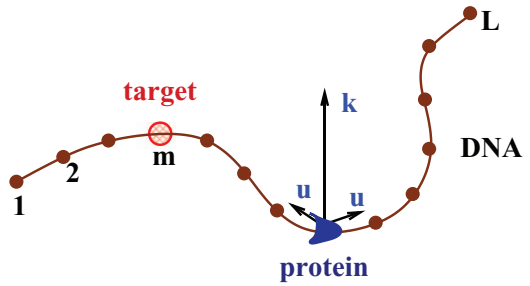

FIG. 1. A general scheme for one-dimensional model of the protein search. DNA chain has $L-1$ non-specific binding sites and one specific site that is a target of the search. A protein molecule can diffuse along the DNA segment with rates u in both directions and it can also dissociate with the rate k. These rates are directly connected with hopping over the barriers along and away from the DNA. The search is finished when the protein binds to the target site at the position m.

for $1 < n < L$, and we assume reflecting boundaries at the ends,

$$\frac{dF_1(t)}{dt} = u[F_2(t) - F_1(t)], \quad \frac{dF_L(t)}{dt} = u[F_{L-1}(t) - F_L(t)]. \tag{2}$$

In addition, we have $F_m(t) = \delta(t)$ and $F_{n\neq m}(t=0) = 0$ because the target is occupying the site m. Introducing Laplace transform, $\widetilde{F_n(s)} \equiv \int_0^\infty e^{-st} F_n(t)dt$, backward master equations can be written as

$$s\widetilde{F_n(s)} = u[\widetilde{F_{n+1}(s)} + \widetilde{F_{n-1}(s)}] - 2u\widetilde{F_n(s)}, \tag{3}$$

$$s\widetilde{F_1(s)} = u[\widetilde{F_2(s)} - \widetilde{F_1(s)}], \quad s\widetilde{F_L(s)} = u[\widetilde{F_{L-1}(s)} - \widetilde{F_L(s)}]. \tag{4}$$

Using boundary and initial conditions, these equations can be solved producing a simple expression,

$$\widetilde{F_n(s)} = \frac{y^{1+L-n} + y^{n-L}}{y^{1+L-m} + y^{m-L}}, \tag{5}$$

for $n > m$, while for $n < m$, we have

$$\widetilde{F_n(s)} = \frac{y^{1-n} + y^{n}}{y^{1-m} + y^{m}}, \tag{6}$$

with $y = (s + 2u - \sqrt{s^2 + 4us})/2u$. Explicit formulas for Laplace transforms of the first-passage probability distribution functions allow us to obtain full description of the search process. The mean first-passage time τ_n to reach the target if the starting point of the protein is at the position n can be calculated from $\tau_n = -\frac{d\widetilde{F_n(s)}}{ds}|_{s=0}$, yielding

$$\tau_n = \frac{[(L-m)^2+(L-m+1)^2]-[(L-n)^2+(L-n+1)^2]}{4u} \tag{7}$$

for $n > m$, while for $n < m$, it can be shown that

$$\tau_n = \frac{[m^2 + (m-1)^2] - [n^2 + (n-1)^2]}{4u}. \tag{8}$$

In order to obtain the mean time $T_m(L)$ for the protein, which starts with equal probability anywhere on the DNA segment of length L, to reach the target at the site m, we must average over the initial positions of the protein molecule, $T_m(L) = \frac{\sum_{n=1}^{L} \tau_n}{L}$. It leads to the following expression:

$$T_m(L) = \frac{(L-m)(L-m+1)(L-m+1/2)+m(m-1)(m-1/2)}{3uL}. \tag{9}$$

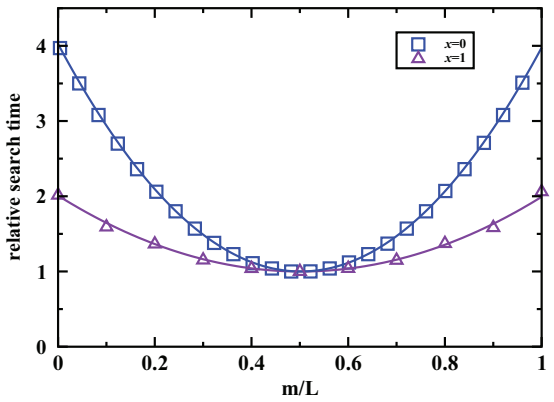

FIG. 2. Relative search time, $T_m/T_{m=(L+1)/2}$, as a function of the position along the DNA segment for $L = 251$. Solid curves are analytical results, while symbols are from Monte Carlo computer simulations. Note that results depend only on ratio of rates $x = k/u$.

As expected, for this case of the unbiased diffusion, the average search time to find the target has a quadratic scaling with DNA length L, as shown in Fig. 3. Its dependence on the target position, m, is non-monotonic: it is minimal when the target is in the middle of the chain, and maximal if the target is positioned at the ends of the DNA segment, as shown in Fig. 2. From Eq. (9), it can be easily shown that

$$T_{\min} = T_{m=(L+1)/2} = \frac{(L+1)(L-1)}{12u}, \qquad T_{\max} = T_{m=1} = T_{m=L} = \frac{(L-1)(L-1/2)}{3u}. \tag{10}$$

One can see that moving the target closer to the center of the DNA chain might decrease the search time significantly, up to four times for large L. This result can be easily explained: when the target is near the ends of the DNA segment, the average distance to the specific site from the starting protein is $\sim L/2$, while for the target in the center of the chain this distance is smaller, $\sim L/4$. Thus the protein molecule, on average, makes shorter scans if the target is in the middle of the DNA segment. The ratio $T_{\max}/T_{\min}$ is intimately related to the scaling of T with L. This can be seen from the following properties of the problem: (i) search time is symmetric with respect to the middle of the chain, $m = (L+1)/2$, i.e., $T_m(L) = T_{L-m+1}(L)$; (ii) $T_m(L)$ also depends on the relative position of the target, m/L, rather than on m alone; and (iii) a protein which starts on one side of the target cannot pass to the other side without finding the target first, therefore $T_m(L) = [mT_m(m) + (L-m+1)T_1(L-m+1)]/(L+1)$. It leads to $T_{(L+1)/2}(L) = T_1((L+1)/2)$, which combined with the scaling $T_m(L) \propto L^\alpha$ (α is a scaling exponent) for $L \gg 1$ yields

$$T_{L/2}(L) \approx 2^{-\alpha} T_1(L), \quad T_{\max}/T_{\min} \approx 2^\alpha. \tag{11}$$

The fact that the searching time depends on the position of the target on DNA has been predicted before.[30] However,

the suggested mechanism relies on combination of 3D and 1D motions, while in our case the mechanism is different since it is a purely 1D effect. It is also important to note that for DNA lengths much longer than the average sliding length of a protein (e.g., 100–500 base pairs for transcription factors), the target position makes only a small effect on the mean search time, as shown below in this paper. In a simple bacteria, the length of DNA is of order of 10^6 bps, while typical sliding length for transcription factor proteins is much smaller, namely 10^2–10^3 bps. In this case, the target positioning is probably not relevant for acceleration of the search. However, when these two lengths are comparable the effect might be significant. For example, this might be the case for compactified DNA molecules in prokaryotes and for nucleosome-bound and tightly wrapped DNA molecules in eukaryotic cells. Our results suggest that this effect might be observed in *in vitro* experiments by changing the ionic strength of the solution: for low-salt conditions, nonspecific protein-DNA interactions are large leading to longer 1D searches.

B. Effectively irreversible detachments

Since in real systems the protein molecule cannot be bound infinitely long to the DNA chain it will eventually dissociate. In this paper, we consider a situation when this dissociation is effectively irreversible. In biological systems, it might correspond to the case of fast intersegment rates, or when the concentration of proteins in the solution is quite small so the associations to the given DNA segment are rare events, or when the protein diffusion in the solution is much faster than binding to DNA and it leads to uncorrelated locations of dissociation and rebinding sites, but proteins do not disappear and they still participate in the search process on other DNA segments. Our goal here is to evaluate the effect of irreversible detachments on the search dynamics. For $k > 0$, the corresponding backward master equations for first-passage distributions are modified as compared to the case without dissociations,

$$\frac{dF_n(t)}{dt} = u[F_{n+1}(t) + F_{n-1}(t)] - (2u + k)F_n(t), \tag{12}$$

$$\begin{aligned} \frac{dF_1(t)}{dt} &= uF_2(t) - (u + k)F_1(t), \\ \frac{dF_L(t)}{dt} &= uF_{L-1}(t) - (u + k)F_L(t). \end{aligned} \tag{13}$$

Utilizing again the Laplace transformations, we obtain for $n > m$ again [see Eqs. (5) and (6)],

$$\widetilde{F_n(s)} = \frac{y^{1+L-n} + y^{n-L}}{y^{1+L-m} + y^{m-L}}, \tag{14}$$

and for $n < m$,

$$\widetilde{F_n(s)} = \frac{y^{1-n} + y^{n}}{y^{1-m} + y^{m}}, \tag{15}$$

but now with $y = (s + 2u + k - \sqrt{(s + 2u + k)^2 - 4u^2})/2u$. It will be also useful to consider a function $\bar{y} \equiv y(s = 0) = (2u + k - \sqrt{k^2 + 4uk})/2u$. It can be roughly interpreted as a quantity that is proportional to the survival probability for the protein over a single step. Defining $x = k/u$, it can be easily shown that $\bar{y}$ varies between $\bar{y} \simeq 1 - \sqrt{x} + O(x)$ at $x \ll 1$ and $\bar{y} \simeq x^{-1} + O(x^{-2})$, at $x \gg 1$.

Since the protein molecule that started at the site n can dissociate before finding the target at the site m, the probability to reach the special binding site, $\Pi_n < 1$, can be explicitly evaluated via $\Pi_n = \widetilde{F_n(s = 0)}$ producing,

$$\Pi_n = \frac{\bar{y}^{m-n}(\bar{y}^{1+2L} + \bar{y}^{2n})}{\bar{y}^{1+2L} + \bar{y}^{2m}}, \tag{16}$$

for $n > m$ and

$$\Pi_n = \frac{\bar{y}^{m-n}(\bar{y} + \bar{y}^{2n})}{\bar{y} + \bar{y}^{2m}} \tag{17}$$

for $n < m$. In the limit $x \gg 1$, we obtain $\Pi_n = x^{-|m-n|}$. The average probability to reach the target is given then by

$$P_m \equiv \frac{\sum_{n=1}^{L} \Pi_n}{L} = \frac{(1 + \bar{y})(1 - \bar{y}^{2L})}{L(1 - \bar{y})(1 + \bar{y}^{1+2(L-m)})(1 + \bar{y}^{1+2(m-1)})}. \tag{18}$$

P_m varies between $P_m \simeq 1 - [m(2L + 1)(L + 1) - m(L + 1 - m)/6]x^2$ at $\bar{y} \approx 1$ and

$$P_m \simeq \frac{1}{L}(1 + 2\bar{y} - \bar{y}^{2(L-m)+1} - \bar{y}^{2m-1}) + O(\bar{y}^2) \tag{19}$$

for $\bar{y} < 1$. Equation (19) is valid not only in the limit of large x. Instead, the crossover between the two regimes for P_m can be obtained by testing the limit of validity of expanding the expression $\bar{y}^L \approx 1 - L\sqrt{x}$ for $x \ll 1$. Then one finds that the crossover is observed for $x_c \sim L^{-2}$. In addition, analysis of Eq. (18) suggests that the probability to reach the target is higher when this special site is in the middle of the DNA segment. Again, this can be easily understood in terms of average distance between the starting position of the proteins and the target. The closer the target, the higher the probability to survive and to reach the target site. As suggested in our discussion after Eq. (11), for $L \gg 1$, $x > x_c$ and m far from the ends of DNA, the dependence of the probability on the target position becomes extremely weak.

In the model with irreversible dissociations, the search times are associated with the conditional mean-first passage times that can be calculated by utilizing the expression $\tau_n = -\frac{d\widetilde{F_n(s)}}{ds}|_{s=0}/\Pi_n$, which leads to

$$\tau_n = \frac{1}{\sqrt{k^2 + 4uk}\ln \bar{y}} \left.\frac{\partial(I(n, a) - I(m, a))}{\partial a}\right|_{a=1} \tag{20}$$

with an auxiliary function, $I(z, a)$, given by $I(z, a) = \ln\left(\bar{y}^{(1+L-z)a} + \bar{y}^{(z-L)a}\right)$ for $n > m$ and $I(z) = \ln\left(\bar{y}^{(1-z)a} + \bar{y}^{za}\right)$ for $n < m$. The explicit expression for the average search time, T_m is quite complex. But it is useful again to consider separately the limiting cases of extremely slow detachments, $x < x_c$ (i.e., $\bar{y} \to 1$), and fast dissociations, $x > x_c$ (i.e., $\bar{y} < 1$). For small x, we obtain

$$T_m(L) \simeq (L^2 - 3Lm + 3m^2)/3u. \tag{21}$$

For large x, Eq. (20) can be simplified (for $1 < m$, $n < L$ with $L \gg 1$) to yield $\tau_n \simeq |m - n|/\left(u\sqrt{x^2 + 4x}\right)$, which

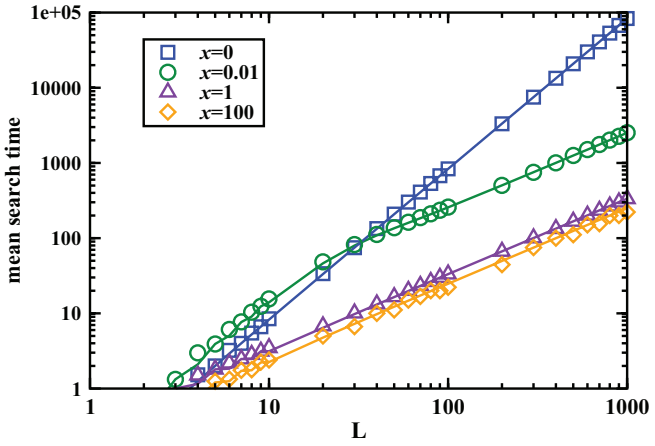

FIG. 3. Average search time as a function of the DNA chain length. The target is always in the middle of the chain. Solid curves are analytical results, while symbols are from Monte Carlo computer simulations. For simplicity, $u = 1\ s^{-1}$ is assumed for all calculations.

leads to

$$T_m(L) = \frac{m(m-1) + (L-m+1)(L-m)}{2Lu\sqrt{x^2+4x}}. \qquad (22)$$

For large DNA segments, these results give us the quadratic scaling, $T \propto L^2$, for $x < x_c$ and the linear scaling, $T \propto L$, for $x > x_c$. The fastest search is again achieved when the target is in the middle of DNA, although the effect of target positioning on the search time is now weaker, as expected from Eq. (11) (see Fig. 2).

III. DISCUSSIONS AND SUMMARY

The computer simulation results and analytical curves for average search times are presented in Figs. 2 and 3. The most surprising result from our theoretical analysis is a linear scaling of T_m as a function of DNA length L for the system with irreversible detachments (see Fig. 3). This observation is counter-intuitive since in the system with the unbiased diffusion the L^2 scaling for the search times is expected. However, in our system the simple diffusion is modified by irreversible dissociations that effectively remove slow hopping molecules. Only proteins that move fast enough (or start close enough) will reach the target. For a given protein molecule, there are many trajectories to reach the target. But since the lifetime of the protein on DNA is limited, only short-time trajectories with the biased motion in the direction of the target will mostly contribute to the mean search time. The dissociations work here as an effective potential that drives the protein molecules away from the starting position, and the system can be described better as diffusion in this effective field. Therefore, we have a driven diffusion motion that leads to the linear scaling, as observed in our case. For cellular DNA lengths ($L \sim 10^6 - 10^9$ bases), it might lead to 6–9 orders of magnitude acceleration over the purely diffusive scanning mechanism, although the probability of reaching the target will be much smaller. It is also interesting to note that for the fixed $x = k/u$, the linear scaling is found only when the length of the DNA chain is long enough ($L > 1/\sqrt{x}$) to observe dissociations. The crossover behavior for $x = 0.01$, when the slope in the log-log plot changes from 2 to 1 (see Fig. 3) clearly illustrates this point. Another important point here is that the linear scaling is still observed when the dissociations become reversible. It is also interesting to note that similar linear scaling has been observed in unrelated processes of formation of signaling molecules profiles during biological development processes.[34]

In conclusion, we have investigated theoretically two possible mechanisms of acceleration of single protein search for specific targets on DNA molecules. Using first-passage analysis for simplified discrete-state stochastic models, it has been shown that putting the target site in the middle of the DNA segment might accelerate the search up to four times in the case without detachments, and up to two times for the situation with dissociations. Much stronger effect is predicted for the protein molecule that might irreversibly dissociate from the DNA chain. Detachments eliminate slow moving molecules and create an effective potential that drives the surviving proteins away from the starting positions. This effective field leads to unexpected linear scaling in the protein motion on DNA, significantly accelerating the overall search process. Our theoretical results are fully supported by Monte Carlo computer simulations. Although the presented theoretical models probably capture some physical/chemical features of the protein search for targets on DNA, they are oversimplified, hence neglecting many important properties of protein search such as sequence dependence, electrostatic effects and the role of protein, and DNA conformational changes. The relevance of the presented mechanisms to *in vivo* systems is also unclear. Recent data on positioning of nucleosomes and other DNA-binding proteins in yeast indicate preference for specific positions on DNA,[32] and in many cases binding sites for transcription factors are in the middle of DNA segments between bound nucleosomes.[33] But it is not known if this specificity is related to speeding up the protein search on DNA. It will be important to test these ideas in more advanced experiments and in more microscopic theoretical calculations.

ACKNOWLEDGMENTS

We would like to acknowledge the support from the Welch Foundation (Grant No. C-1559).

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell* (Garland Science, 2002).
[2] A. Riggs, S. Bourgeois, and M. Cohn, J. Mol. Biol. **53**, 401 (1970).
[3] O. G. Berg, R. B. Winter, and P. H. von Hippel, Biochemistry **20** 6929 (1981).
[4] R. B. Winter and P. H. von Hippel, Biochemistry **20** 6948 (1981).
[5] S. Halford and J. Marko, Nucleic Acids Res. **32**, 3040 (2004).
[6] D. M. Gowers, G. G. Wilson, and S. E. Halford, Proc. Natl. Acad. Sci. U.S.A. **102**, 15883 (2005).
[7] Y. M. Wang, R. H. Austin, and E. C. Cox, Phys. Rev. Lett. **97**, 048302 (2006).
[8] J. Elf, G.-W. Li, and X. S. Xie, Science **316**, 1191 (2007).
[9] S. E. Halford, Biochem. Soc. Trans. **37**, 343 (2009).
[10] L. Mirny, M. Slutsky, Z. Wunderlich, A. Tafvizi, J. Leith, and A. Kosmrlj, J. Phys. A: Math. Theor. **42**, 434013 (2009).
[11] D. C. Rau and N. Y. Sidorova, J. Mol. Biol. **395**, 408 (2010).
[12] A. Tafvizi, F. Huang, A. R. Fersht, L. A. Mirny, and A. M. van Oijen, Proc. Natl. Acad. Sci. U.S.A. **108**, 563 (2011).

[13]M. L. Hegde, T. K. Hazra, and S. Mitra, Cell Mol. Life Sci. **67**, 3573 (2010).
[14]I. Bonnetm A. Biebricher, P.-L. Porte, C. Loverdo, O. Benichou, R. Voituriez, C. Escude, W. Wende, A. Pingoud, and P. Desbiolles, Nucleic Acids Res. **36** 4118 (2008).
[15]M. Slutsky and L. Mirny, Biophys. J. **87**, 4021 (2004).
[16]T. Hu, A. Grosberg, and B. Shklovskii, Biophys. J. **90**, 2731 (2006).
[17]A. G. Cherstvy, A. B. Kolomeisky, and A. A. Kornyshev, J. Phys. Chem. B **112**, 4741 (2008).
[18]V. Dahirel, F. Paillusson, M. Jardat, M. Barbi, and J.-M. Victor, Phys. Rev. Lett. **102**, 228101 (2009).
[19]R. Murugan, J. Phys. A: Math. Theor. **44**, 505002 (2011).
[20]A. Afek, I. Sela, N. Musa-Lempel, and D. B. Lukatsky, Biophys. J. **101** 2465 (2011).
[21]C. Loverdo, O. Benichou, R. Voituriez, A. Biebricher, I. Bonnet, and P. Desbiolles, Phys. Rev. Lett. **102**, 188101 (2009).
[22]O. Benichou, Y. Kafri, M. Sheinman, and R. Voituriez, Phys. Rev. Lett. **103**, 138102 (2009).
[23]D. Vuzman, M. Polonsky, and Y. Levy, Biophys. J. **99**, 1202 (2010).
[24]P.-W. Fok, C.-L. Guo, and T. Chou, J. Chem. Phys. **129**, 235101 (2008).
[25]Y. von Hansen, R. R. Netz, and M. Hinczewski, J. Chem. Phys. **132**, 135103 (2010).
[26]A. B. Kolomeisky, Phys. Chem. Chem. Phys. **13**, 2088 (2011).
[27]H.-X. Zhou, Proc. Natl. Acad. Sci. U.S.A. **108**, 8651 (2011).
[28]J. Reingruber and D. Holcman, Phys. Rev. E **84**, 02901(R) (2011).
[29]A. G. Cherstvy, J. Phys. Chem. B **113**, 4242 (2009).
[30]M. Coppey, O. Benichou, R. Voituriez, and M. Moreau, Biophys. J. **87**, 1640 (2004).
[31]S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, 2001).
[32]E. Segal and J. Widom, Trends Genet. **25**, 335 (2009).
[33]G.-C Yuan, Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando, Science **309**, 626–630 (2005).
[34]A. B. Kolomeisky, J. Phys. Chem. Lett. **2**, 1502 (2011).