Searching Fast for a Target on DNA without Falling to Traps

O. Bénichou,[1] Y. Kafri,[2] M. Sheinman,[2] and R. Voituriez[1]

[1]*UMR 7600, Université Pierre et Marie Curie/CNRS, 4 Place Jussieu, 75255 Paris Cedex 05 France*
[2]*Department of Physics, Technion, Haifa 32000, Israel*

(Received 15 January 2009; published 24 September 2009)

Genomic expression depends critically on both the ability of regulatory proteins to locate specific target sites on DNA within seconds and on the formation of long-lived (many minutes) complexes between these proteins and the DNA. Equilibrium experiments show that indeed regulatory proteins bind tightly to their target site. However, they also find strong binding to other nonspecific sites which act as traps that can dramatically increase the time needed to locate the target. This gives rise to a conflict between the speed and stability requirements. Here we suggest a simple mechanism which can resolve this long-standing paradox.

DOI: 10.1103/PhysRevLett.103.138102 PACS numbers: 87.10.Mn, 05.40.−a

It is commonly believed that three-dimensional diffusion is too slow for proteins to locate their specific target on a DNA molecule for cells to function properly. To resolve this issue Berg and von Hippel suggested, in a series of seminal papers [1,2], that combining periods of one-dimensional diffusion along the DNA (sliding) with periods of three-dimensional diffusion off the DNA (jumping) can speed up the search time by several orders of magnitude. Since then, sliding (or equivalently binding of proteins to nonspecific DNA sequences) has been observed in many experiments [3–5] and is now believed to be a common mechanism [6–11]. On the other hand, as pointed out already in [12], experimental and theoretical works have shown that the binding energies of a protein to different DNA sequences are very large—a direct consequence of the required stability of the protein with its target site. The binding energies can be well fitted by a Gaussian with the strongest binding energies of the order of $\sim 30k_BT$ and a standard deviation of the order of $5k_BT$ [13]. This casts a cloud on the simple facilitated diffusion picture of Berg and von Hippel—the binding energy distribution suggests an unacceptably slow search with very slow sliding and deep traps [10]. This unresolved conflict is called the speed-stability paradox [2].

Here, motivated by direct experimental observations [14–17] and theoretical work by Slutsky and Mirny [10] and Hu *et al.* [18], we consider a model in which the protein, when bound to the DNA, can switch between two conformations separated by a free energy barrier. In one, termed the search state, the protein is loosely bound to the DNA and can slide along it. In the second, recognition mode, it is trapped in a deep energy well. Note that equilibrium measurements of binding energies to the DNA are controlled by the recognition state.

In this Letter we argue that, due to the occurrence of several time scales in the search process, the widely used definition of the reaction rate of a single protein as the inverse of the average search time t^{av} [19] is generally irrelevant as a measure of the efficiency of target location on DNA. When n_p proteins are searching for the target, the relevant quantity is the probability $\mathcal{R}_{n_p}(t)$ for a reaction to occur before time t. We show below that $\mathcal{R}_{n_p}(t)$ can reach values close to 1 in a time scale $t^{\mathrm{typ}}_{n_p}(t)$ which can be orders of magnitude smaller than the value t^{av}/n_p expected from the usual approach.

Our analysis has several important merits. First, it reports a *fast* search time despite a very strong binding of the protein in the recognition state to any site on the DNA. We suggest that the measured binding energies of proteins to the DNA are irrelevant to the kinetics of the search process; the relevant quantities are transition rates (specified below). Second, it shows that in the realistic case of generic disorder in the barrier height the search can be very effective even if the target site is *not* designed. If experimentally verified, the proposed mechanism will resolve the speed-stability paradox.

The model consists of n_p proteins which can each be in three states: (i) an unbound state u, in which it performs three-dimensional diffusion (jumping), (ii) a search state s, where it is weakly bound to the DNA, performing one-dimensional diffusion (sliding), and (iii) a recognition state r, where it is tightly bound to the DNA. We assume, for simplicity, that in the recognition state the protein is trapped in a deep energy well (as justified by the experimentally measured strong binding energies) and is unable to move [10]. The transition rates, λ_s^i, λ_r^i, λ_b, and λ_u, between the different states are defined in Fig. 1. To model sliding, in the s state the protein can move with rate $\lambda_0/2$ to neighboring sites on the DNA. Note that the rates λ_r^i and λ_s^i are expected in general to depend on the location $i = 1, \ldots, N$ along the DNA. In principle λ_0 and λ_u also have a dependence on i. As justified later this will have a weaker effect on our results and we omit it for clarity. We consider a DNA molecule of N sites, with a centered target site (labeled 0), and finally assume that after a jump the protein relocates to a random position on the DNA due to its packed conformation [20]. A similar model was pre-

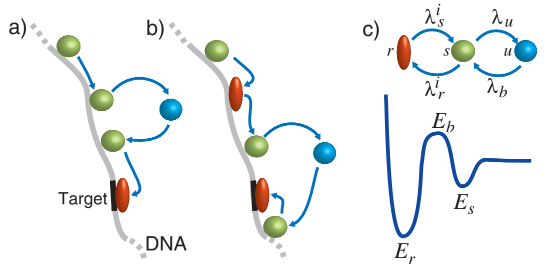

FIG. 1 (color online). An illustration of the model. (a) A time sequence of a protein sliding in the s mode [light gray (green) circle], diffusing off the DNA [dark gray (blue) circle], and entering the target site in the r mode (red oval). (b) A protein finding the target after entering the r state. (c) An illustration of the rates and the energy landscape which governs them at each location, $i = 1, \ldots, N$, along the DNA. Here $\lambda_r^i \propto e^{-(E_b^i - E_s^i)/k_BT}$, $\lambda_s^i \propto e^{-(E_b^i - E_r^i)/k_BT}$, and $\lambda_u \propto e^{-E_s^i/k_BT}$, while λ_b depends on details of the three-dimensional diffusion process.

viously studied in [18] in the case of site independent transition rates. There, however, it was assumed that the protein equilibrates with DNA segments far from the target site. As we show below, this assumption can fail.

Nondisordered case.—To gain an understanding of the difference between the two time scales $t_{n_p}^{\rm typ}$ and $t^{\rm av}/n_p$ we first consider $n_p = 1$ in a simplified model where λ_r^i and λ_s^i are independent of i except at the target site $\mathcal{T}$ where $\lambda_r^{\mathcal{T}} = \infty$ and $\lambda_s^{\mathcal{T}} = 0$. The disorder of the DNA sequence is neglected and the target is designed such that a reaction takes place at the first visit of the target site. These assumptions will be relaxed later. As stated above, we are interested in the probability $\mathcal{R}(t) = \int_0^t P(t')dt'$ that a reaction occurs before time t, where $P(t)$ is the distribution of the first-passage time (FPT) [21] to the target (we drop the subscript when $n_p = 1$).

The Laplace transform, $\tilde{P}(s) = \int_0^\infty e^{-st}P(t)dt$, of $P(t)$, can be obtained exactly. Consider the joint probability density for a protein to find the target at time $t = t_s + t_r$ starting from a location x_0 at $t = 0$ before unbinding from DNA. Here t_s is the total time spent in the s state and t_r is the total time spent in the r state. If exactly n transitions occurred from the s state to the r state, this is given by

$$P_n(t_s, t_r|x_0) = \lambda_s \mathcal{P}(n-1, \lambda_s, t_r)\mathcal{P}(n, \lambda_r, t_s) j(t_s|x_0)e^{-\lambda_u t_s}, \qquad (1)$$

where $\mathcal{P}(n, \mu, t) = (\mu t)^n e^{-\mu t}/n!$ is the Poisson distribution [with the convention $\mathcal{P}(-1, \mu, t) \equiv \delta(t)/\mu$], and $j(t|x_0)$ is the FPT density at the target $x = 0$ for a usual random walk starting from x_0 [22]. The FPT density before unbinding starting from x_0 then reads:

$$J(t|x_0) = \sum_{n=0}^{\infty} \int_0^\infty \int_0^\infty dt_s dt_r \delta(t_s + t_r - t) P_n(t_s, t_r|x_0). \quad (2)$$

After Laplace transform and using $\tilde{\mathcal{P}}(n, \mu, s) = \mu^n/(s+\mu)^{n+1}$, we find $\tilde{J}(s|x_0) = \tilde{j}(u(s)|x_0)$ with $u(s) = \frac{s(s+\lambda_r+\lambda_s+\lambda_u)+\lambda_s\lambda_u}{s+\lambda_s}$. Averaging over x_0 and following [6,23], we then obtain

$$\tilde{P}(s) = \tilde{j}(u(s))\left\{1 - \frac{\lambda_b\lambda_u}{s+\lambda_b}\frac{1-\tilde{j}(u(s))}{u(s)}\right\}^{-1}, \qquad (3)$$

where $\tilde{j}(s) \equiv \langle \tilde{j}(s|x)\rangle_x \sim \frac{1}{N}\sqrt{(1+e^{-s/\lambda_0})/(1-e^{-s/\lambda_0})}$ for large N [22].

An analysis of the pole structure of Eq. (3) shows that in the regime $\lambda_s \ll \lambda_r \ll \lambda_u, \lambda_b, \lambda_0$ (with λ_u, λ_b, λ_0 of comparable order) the reaction probability simplifies to

$$\mathcal{R}(t) \simeq 1 - qe^{-t/\tau_1} - (1-q)e^{-t/\tau_2}, \qquad (4)$$

with $q = [1 + \lambda_r/(\lambda_u\kappa/N)]^{-1}$, $\kappa = \sqrt{\coth(\lambda_u/2\lambda_0)}$, $\tau_1 = (\lambda_b + \lambda_u)/[\lambda_b(\lambda_r + \kappa\lambda_u/N)]$ and $\tau_2 = (\lambda_r + \kappa\lambda_u/N)/(\lambda_s\kappa\lambda_u/N)$. The short time scale τ_1 characterizes searches where the protein never enters the r state and is therefore independent of the binding energy E_r (and hence of λ_s). The time scale τ_2 characterizes searches where the protein enters the r state, and is therefore much larger than τ_1 in the case of strong binding (λ_s small). In turn, q is the probability of an event where the target is found without falling into the trap.

Expression (4) enables an explicit determination of $t^{\rm ave} = q\tau_1 + (1-q)\tau_2$ and $t^{\rm typ}$, which can be defined, for example, through

$$\int_0^\infty e^{-t/t^{\rm typ}} P(t)dt = \tilde{P}(1/t^{\rm typ}) = 1/2. \qquad (5)$$

We stress that experimentally the relevant time, where almost all search processes end, is $t^{\rm typ}$ and not $t^{\rm av}$. In the regime $\lambda_r \gg \lambda_u\kappa/N$, one has $t^{\rm typ} \simeq t^{\rm av} \simeq \tau_2$. A difference between $t^{\rm typ}$ and $t^{\rm av}$ emerges as λ_r is decreased, and in the limit $\lambda_r \ll \lambda_u\kappa/N$ we find that $t^{\rm typ} \simeq \tau_1/(2q-1)$ (with $q \simeq 1$) is independent of λ_s. This shows that for DNA lengths $N \leq \lambda_u\kappa/\lambda_r$ the typical search time is significantly smaller than the average even in the presence of deep traps (λ_s small). This is a direct result of the two time scales, τ_1 and τ_2.

The results, compared with numerics which were performed using a standard continuous time Gillespie algorithm, are shown in Fig. 2. We use realistic ranges of parameters (from available experimental data summarized in [24]) which are specified in the caption. We assume the barrier height for different DNA sequences to be of the same order of magnitude as the experimentally measured binding energies [13]. This conforms well with measurements of transition rates of about 0.1 s^{-1} for a papillomavirus E2 protein-DNA complex, which gives barriers of the order of 20–30k_BT [14]. It is found that $\mathcal{R}(t)$ reaches a plateau close to 1 on a typical time scale $t^{\rm typ}$ which, for $N = 10^6$, is much shorter than $t^{\rm av}$.

This interesting regime where $t^{\rm typ} \ll t^{\rm av}$ requires a rather large barrier between the s and r state in the case of long DNA molecules (namely, $\lambda_r \leq \lambda_u\kappa/N$). Although this condition may not hold for all proteins, we now argue

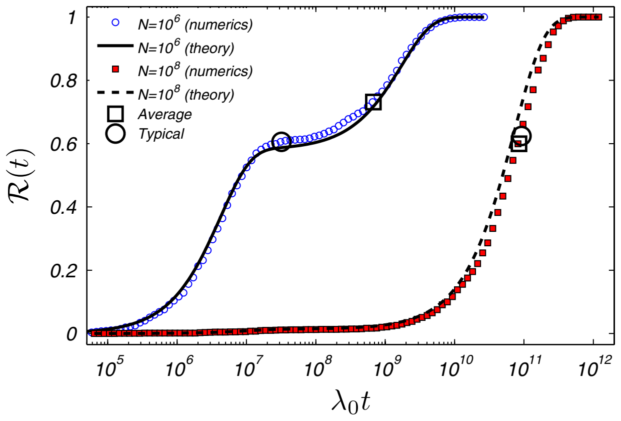

FIG. 2 (color online). A plot of $\mathcal{R}(t)$ for $N = 10^6$ (empty circles) and $N = 10^8$ (filled squares). Lines correspond to Eq. (4), with τ_1, τ_2, and q derived analytically. Here $\lambda_u = 10^{-2}\lambda_0$, $\lambda_b = 0.1\lambda_0$, $\lambda_r = 10^{-7}\lambda_0$, and $\lambda_s = 10^{-9}\lambda_0$, in agreement with [24]. These correspond to energies, measured relative to the energy of the unbound state, of $E_s = -4.6k_BT$, $E_b = 11.5k_BT$, and $E_r = -9.2k_BT$. Experiments suggest $\lambda_0 \simeq 10^6\ \mathrm{s}^{-1}$ for the Lac repressor [5].

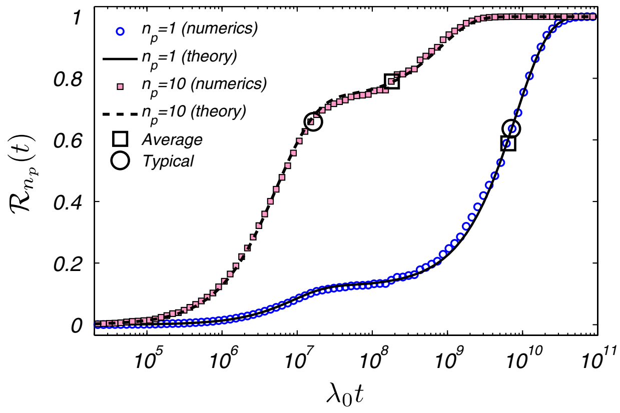

FIG. 3 (color online). A plot of $\mathcal{R}_{n_p}(t)$ for $n_p = 1$ (empty circles) and $n_p = 10$ (filled squares). Here $N = 10^6$, $\lambda_u = 10^{-4}\lambda_0$, $\lambda_b = 0.1\lambda_0$, $\lambda_r = 10^{-7}\lambda_0$, and $\lambda_s = 10^{-9}\lambda_0$ (see [24]). These correspond to energies, measured relative to the unbound state, of $E_s = -9.2k_BT$, $E_b = 6.9k_BT$, and $E_r = -13.8k_BT$. Lines correspond to Eq. (4) with calculated values of τ_1, τ_2, and q. Note that here λ_u is different from Fig. 2.

that it can, to a large extent, be relaxed when n_p proteins are searching for the target simultaneously. In this case, even when for a single protein $t^{\mathrm{av}} \simeq t^{\mathrm{typ}}$ the typical search time $t_{n_p}^{\mathrm{typ}}$ of n_p proteins can be significantly shorter than t^{av}/n_p even for relatively small values $n_p \approx 10$. Here, again, t^{av} is the average search time of a single protein and $t_{n_p}^{\mathrm{typ}}$ is defined as in Eq. (5) where for n_p proteins the first-passage distribution $P_{n_p}(t)$ is deduced from the cumulative distribution

$$\mathcal{R}_{n_p}(t) = 1 - [1 - \mathcal{R}(t)]^{n_p}. \tag{6}$$

In Fig. 3 we show the results of $\mathcal{R}_{n_p}(t)$ for $n_p = 10$. Note that, as claimed above, $t_{n_p}^{\mathrm{typ}} \ll t^{\mathrm{av}}/n_p$, whereas t^{typ} is close to t^{av} for one protein. This can be understood as follows. Using Eq. (4) in Eq. (6), it is obvious that when $\tau_2 \gg \tau_1$, the decay of $\mathcal{R}_{n_p}(t)$ is dominated by τ_1 as long as $(1 - q)^{n_p} \ll 1$. In essence, since only one protein needs to find the target, the probability of a catastrophic event where the search time is of the order of τ_2 is $p_{\mathrm{cat}} = (1 - q)^{n_p}$, which decays exponentially fast with n_p. For large enough values of n_p the short time scale τ_1 controls the behavior of $\mathcal{R}_{n_p}(t)$, even if it is insignificant for the one protein search time. The typical search time is $t_{n_p}^{\mathrm{typ}} = \tau_1/m$, where m is of the order of n_p, and is therefore again widely independent of the binding energy of the r mode. This makes fast searches possible even in the presence of deep traps—enabling both speed and stability.

Disordered case.—We now argue that this mechanism of fast search can still be at play when the binding energy of the protein to the DNA is strongly disordered, as observed in experiments. To account for this we consider the case where the barrier height is drawn from a Gaussian distribution: $p(E_b^i) = e^{-(E_b^i - E_0)^2/2\sigma^2}/\sqrt{2\pi\sigma^2}$. Importantly, in the presence of disorder we can propose an intrinsic definition of the target as the site with the lowest barrier with no specifically designed properties. Indeed, our previous assumption $\lambda_r^{\mathcal{T}} = \infty$ at the target site and λ_r^i small everywhere else is a rather strong demand. Since the target sequence is of the order of 10 base pairs, many sequences with similar properties are likely to exist, unless the DNA sequence is tailored.

To analyze this model we combine numerics with a mean-field analysis. For simplicity, we consider the extreme case where all recognition sites are infinitely long-lived $\lambda_s = 0$, which obviously fulfills the stability requirement. Note that t^{av} is then infinite. In real systems one expects this transition rate to also be disordered. In particular, for sites very different from the target site the recognition state may not be present at all. Since such features would only make the search quicker, our model gives an overestimate of the search time.

Within the mean-field approach we replace the different quantities by their disorder average and account for the barrier at the target site. We first compute the disorder averaged probability of crossing the barrier at the target at each visit. Knowing the distribution of the minimum of the barrier [25], this is given by $p_1 = \int_{-\infty}^{\infty} dE[e^{-E/k_BT}/(1 + \lambda_u/\lambda_0 + e^{-E/k_BT})](d/dE)\{\frac{1}{2}\,\mathrm{erfc}[(E - E_0)/\sqrt{2}\sigma]\}^N$. Here we set the time scale of the activation process across the barrier to be λ_0. Note that even in the case where p_1 is small, the search is still controlled by transport since it is governed by the statistics of return times to the target. We finally assume that the expression for $u(s)$ of the non-disordered model holds with λ_r replaced by $\bar{\lambda}_r = \lambda_0 \int_{-\infty}^{\infty} e^{-E/k_BT}(e^{-(E-E_0)^2/2\sigma^2}/\sqrt{2\pi}\sigma)dE$ and $\tilde{j}$ replaced by

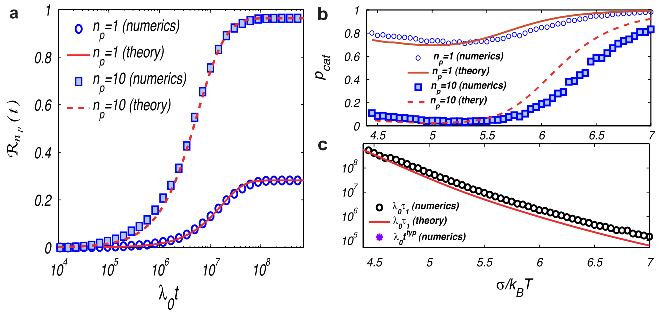

FIG. 4 (color online). Disordered case for $n_p = 1$ (empty circles) and $n_p = 10$ (filled squares), where $N = 10^6$, $\lambda_u = 10^{-2}\lambda_0$ ($E_s = -4.6k_BT$), $\lambda_b = 0.1\lambda_0$, and $E_0 = 25.4k_BT$. (a) Plot of $\mathcal{R}_{n_p}(t)$ for $\sigma = 5.3k_BT$. The lines were obtained by fitting the form $1 - [qe^{-t/\tau_1} + (1-q)]^{n_p}$ to the numerical simulations with $q = 0.2817$, $\lambda_0\tau_1 = 1.7 \times 10^7$, and $\tau_2 = \infty$. These are close to the mean-field prediction $q = 0.2827$, $\lambda_0\tau_1 = 1.1 \times 10^7$. The average height of the barrier at the target site is then $6.25k_BT$, which corresponds to a transition rate of $2 \times 10^3\ \mathrm{s}^{-1}$. (b) p_{cat} as a function of σ for $n_p = 1$ and $n_p = 10$. (c) t^{typ} for $n_p = 10$ and τ_1 are plotted as a function of σ. Using $\lambda_0 = 10^6\ \mathrm{s}^{-1}$ [5] for $n_p = 10$ at the minimal p_{cat} we find $t^{\mathrm{typ}} \simeq 10$ s. Note that by moderate changes in E_0 similar results can be obtained for longer DNA sequences.

$$\tilde{j}_{p_1} = \frac{p_1\tilde{j}(z)}{1 - (1 - p_1)\tilde{j}_0(z)}, \tag{7}$$

where $\tilde{j}_0(s)$ is the generating function of the first return time to site 0 [22].

First, we show that the two scales scenario described above still holds. Indeed, Fig. 4(a) shows that $\mathcal{R}(t)$ is well fitted by Eq. (4) for realistic values of parameters. This implies that for n_p large enough the only relevant time scale is τ_1, and the typical search time again takes the form $t_{n_p}^{\mathrm{typ}} \simeq \tau_1/m$ with m of the order of n_p. This enables a fast search even in the presence of infinitely deep traps.

The regime of a fast search with $t_{n_p}^{\mathrm{typ}}$ independent of the trap depth E_r also requires, as above, a small p_{cat}. We now show that this condition holds in a wide range of disorder parameters. To illustrate this, the dependence (holding all other variables constant) of p_{cat} and $t_{n_p}^{\mathrm{typ}}$ on σ, obtained from numerics and the mean-field treatment, is shown in Figs. 4(b) and 4(c) for realistic values of parameters. Notably, the value of p_{cat} can be minimized as a function of σ. This reflects the fact that for small values of σ the DNA sequence has to be scanned many times before the target is entered in the r mode. Increasing σ lowers the barrier at the target and therefore reduces the number of scans needed, which diminishes p_{cat}. For larger σ the chance of falling into a trap increases due to lower secondary minima of the barrier, which leads to an increase of p_{cat}. As expected, p_{cat} is dramatically decreased when n_p is increased, even by a few units, and can remain small for a wide range of values of σ. For larger σ, p_{cat} increases and $t_{n_p}^{\mathrm{typ}}$ rises quickly as it starts to depend on τ_2.

Most important, as advertised above, these results show that small values of $t_{n_p}^{\mathrm{typ}}$ and p_{cat} can be obtained with realistic values of the parameters (see Fig. 4). Reasonable search times (in the range of seconds) are obtained for a large range of σ as long as n_p is of the order of 10 or more proteins, even in the extreme case of infinitely deep traps suggesting a possible resolution of the speed and stability requirements.

We thank E. Braun and D. Levine for comments and the support of the High Council for Scientific and Technological Cooperation between France and Israel. Y. K. and M. S. were also supported by the Israeli Science Foundation, and O. B. and R. V. by ANR grant "Dyoptri."

[1] O. G. Berg and P. H. von Hippel, J. Biol. Chem. **264**, 675 (1989).
[2] R. B. Winter, O. G. Berg, and P. H. von Hippel, Biochemistry **20**, 6961 (1981).
[3] I. Bonnet *et al.*, Nucleic Acids Res. **36**, 4118 (2008); C. Loverdo *et al.*, Phys. Rev. Lett. **102**, 188101 (2009).
[4] J. Elf, G.-W. Li, and X. S. Xie, Science **316**, 1191 (2007).
[5] Y. M. Wang *et al.*, Phys. Rev. Lett. **97**, 048302 (2006).
[6] M. Coppey *et al.*, Biophys. J. **87**, 1640 (2004).
[7] I. Eliazar, T. Koren, and J. Klafter, J. Phys. Condens. Matter **19**, 065140 (2007).
[8] T. Hu, A. Y. Grosberg, and B. I. Shklovskii, Biophys. J. **90**, 2731 (2006).
[9] M. A. Lomholt, T. Ambjornsson, and R. Metzler, Phys. Rev. Lett. **95**, 260603 (2005).
[10] M. Slutsky and L. A. Mirny, Biophys. J. **87**, 4021 (2004).
[11] B. van den Broek *et al.*, Proc. Natl. Acad. Sci. U.S.A. **105**, 15 738 (2008).
[12] O. G. Berg and P. H. von Hippel, J. Mol. Biol. **193**, 723 (1987).
[13] U. Gerland, J. D. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **99**, 12 015 (2002).
[14] D. U. Ferreiro and G. de Prat-Gay, J. Mol. Biol. **331**, 89 (2003).
[15] C. G. Kalodimos *et al.*, Science **305**, 386 (2004).
[16] A. Pingoud and W. Wende, Structure **15**, 391 (2007).
[17] S. A. Townson *et al.*, Structure **15**, 449 (2007).
[18] L. Hu, A. Y. Grosberg, and R. Bruinsma, Biophys. J. **95**, 1151 (2008).
[19] P. Hanggi, P. Talkner, and M. Borkovec, Rev. Mod. Phys. **62**, 251 (1990).
[20] M. Sheinman and Y. Kafri, Phys. Biol. **6**, 016003 (2009).
[21] S. Redner, *A Guide to First Passage Time Processes* (Cambridge University Press, Cambridge, England, 2001).
[22] E. W. Montroll, J. Math. Phys. (N.Y.) **10**, 753 (1969).
[23] O. Benichou *et al.*, Phys. Chem. Chem. Phys. **10**, 7059 (2008).
[24] Z. Wunderlich and L. A. Mirny, Nucleic Acids Res. **36**, 3570 (2008).
[25] L. de Haan and A. Ferreira, *Extreme Value Theory: An Introduction* (Springer, New York, 2006).