How does a protein search for the specific site on DNA: The role of disorder

Tao Hu and B. I. Shklovskii
Department of Physics, University of Minnesota, 116 Church Street SE, Minneapolis, Minnesota 55455, USA
(Received 27 February 2006; revised manuscript received 20 April 2006; published 3 August 2006)

Proteins can locate their specific targets on DNA up to two orders of magnitude faster than the Smoluchowski three-dimensional diffusion rate. This happens due to nonspecific adsorption of proteins to DNA and subsequent one-dimensional sliding along DNA. We call such a one-dimensional route towards the target an "antenna." We studied the role of the dispersion of nonspecific binding energies within the antenna due to a quasirandom sequence of natural DNA. A random energy profile for sliding proteins slows the searching rate for the target. We show that this slowdown is different for macroscopic and mesoscopic antennas.

DOI: 10.1103/PhysRevE.74.021903 PACS number(s): 87.14.Ee, 87.15.Vv

A protein binding to a specific site on DNA, which we call the target, is one of the central paradigms of biology [1]. Well-known examples include the *lac*-repressor in *E. coli*, which regulates a specific gene producing enzyme-consuming lactose and the proper restriction enzyme-destroying genome of invading *E. coli* λ-phage in real-time warfare for bacteria survival. It has been known since the early days of molecular biology that in some cases proteins can find their target sites along a DNA chain one to two orders faster than the maximum rate achievable by three-dimensional diffusion [2,3]. To resolve this paradox, nonspecific binding and subsequent one-dimensional (1D) sliding of proteins along the DNA to the target were suggested as an important component of the searching process [2,3]. This idea was studied in various models proposed by both physicists and biologists [4–8]. A comprehensive study of the interplay between the 1D sliding and 3D diffusion for different DNA conformations on the search rate can be found in Ref. [9].

Some authors calculate the typical time τ needed for the target site to be found by a protein, when a small concentration c of proteins is randomly introduced into the system. Other authors [9] consider the specific site as a sink consuming proteins with the diffusion-limited rate J proportional to the concentration c (which in turn should be supported on a constant level by an influx of proteins into the system). Obviously, then, $\tau=1/J$. Search rate enhancement due to the sliding along DNA may be calculated as the ratio of the rate J to the 3D Smoluchowski rate $J_s=4\pi D_3 cb$ of diffusion to the sphere of radius b modeling the target site on DNA. The central physical idea is that one can define a piece of DNA adjacent to the target for which 1D sliding diffusion dominates over a parallel 3D diffusion channel and which, therefore, serves as a receiving antenna for the 3D Smoluchowski-like diffusion of proteins. Then the key point of the theory is to find the antenna length λ. In the language of stationary flux J, this is done by matching the incoming 3D flux J_3 of proteins to the antenna with the 1D flux J_1 of proteins sliding on the antenna toward the target.

All the works cited above assume that the nonspecific adsorption energy w of protein is sequence independent; i.e., the energy profile experienced by the searching protein away from the target is totally flat. This, however, disagrees with the quasirandom character of the natural sequences of DNA. It is known that the nonspecific protein-DNA adsorption energy can be divided into two parts [10,11]: (i) The sequence-independent Coulomb energy of the attraction between the positively charged domain of the protein surface and the negatively charged phosphate backbone and (ii) the sequence specific adsorption energy due to the formation of hydrogen bonds of the protein with the DNA bases. This is done by the recognition α-helix going deep into the major groove of DNA [1]. Suppose the protein encounters l base pairs between positions i and $i+l$. We call this position of the protein *site* i and characterize it by energy $\epsilon_i<0$, where the energy of the free protein in water is chosen to be 0. Because the sliding protein has a complex nonuniform structure and interacts with a random DNA sequence, the total energy ϵ_i randomly fluctuates along DNA (Fig. 1). One can assume that at nonspecific positions on DNA, the protein exploits the same set of potential hydrogen bonds it forms with the target [12]. Since target recognition is often mediated by hydrogen bonds to some of the four chemical groups on the major groove side of the base pair [13] and the recognition α-helix interacts with several base pairs, many hydrogen bonds contribute to ϵ_i. Therefore the distribution of ϵ_i can be approximated by the Gaussian distribution [12,14,15] with a mean w and standard deviation $\sigma \ll |w|$:

$$g(\epsilon_i)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(\epsilon_i-w)^2}{2\sigma^2}\right]. \tag{1}$$

In this paper we study the role of disorder on the rate enhancement J/J_s assuming that the disorder is strong—i.e., $\sigma>kT$, where k is the Boltzmann constant and T is the ambient temperature.

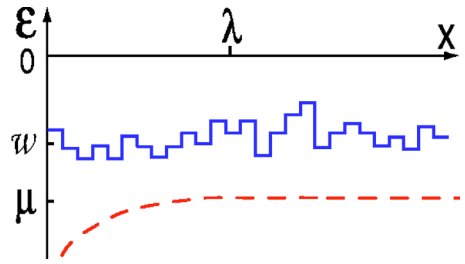

FIG. 1. (Color online) Distribution of the nonspecific adsorption energy ϵ and chemical potential $\mu(x)$ along a DNA molecule. The target site is located at $x=0$; λ is the antenna length.

Similar to the the case of the flat energy profile [9], we assume that transport outside the antenna is *mainly* due to the 3D diffusion, while inside the antenna transport is *dominated* by sliding, or 1D diffusion, along DNA and we equate the fluxes J_1 and J_3 to find λ. The rate J_3 is given by the Smoluchowski formula for the target size λ and for the concentration of "free" (not adsorbed) proteins, c_3; it is $J_3 \sim D_3 c_3 \lambda$. The flux on antenna J_1 strongly depends on σ and also, generally speaking, on the DNA sequence in the finite antenna. We show that there is a characteristic length of the antenna, $\lambda = \lambda_c(\sigma, T)$, such that at $\lambda > \lambda_c$ flux J_1 self-averages and becomes sequence independent. Such a "macroscopic" antenna determines J/J_s for moderate disorder. In this case, the ratio J/J_s decreases exponentially fast with the growth of disorder. At stronger disorder we deal with a mesoscopic antenna with $\lambda < \lambda_c$ and strictly speaking J/J_s depends on a random DNA sequence. In this paper, we concentrate only on the most probable value of J/J_s. In order to calculate it, we estimate the most probable value of J_1. We show that in such a mesoscopic situation disorder leads to a weaker reduction of J/J_s.

We assume that within some volume v there is a straight, immobile (double helical) DNA with length L smaller than $v^{1/3}$, but much larger than any antenna length. For a dilute DNA solution, $1/v$ stands for the concentration of DNA. We also assume that all the microscopic length scales such as the length of a base pair, the size of the target site, the diameter of the DNA, etc., are of the same order b. We are mainly interested in the scaling dependence of the rate enhancement J/J_s on major system parameters, such as σ, w, L, and v. This means that all numerical coefficients are dropped in our scaling estimates.

To estimate J_1, we assume that at each site i on DNA, the protein has some probabilities of hopping to nearest-neighboring sites j. We write the probability for the hopping from an occupied site i to an empty site j as

$$\gamma_{ij} = \nu_0 \exp\left(-\frac{\epsilon_j - \epsilon_i + |\epsilon_j - \epsilon_i|}{2kT}\right) = \begin{cases} \nu_0 \exp\left(-\dfrac{\epsilon_j - \epsilon_i}{kT}\right) & \text{if } \epsilon_j > \epsilon_i, \\ \nu_0 & \text{if } \epsilon_j < \epsilon_i, \end{cases} \tag{2}$$

where $\nu_0 \sim D_1/b^2$ is the effective attempt frequency. In Eq. (2) we neglected the activation barriers separating two states in comparison with $\epsilon_j - \epsilon_i$. The number of proteins making such a transition from site i to j per unit time can be estimated by $\Gamma_{ij} = \gamma_{ij} f_i (1 - f_j)$, where the function f_i is the average occupation number of sites, i. At small enough c, all $f_i \ll 1$ and thus $\Gamma_{ij} \simeq \gamma_{ij} f_i$. The function f_i is given then by

$$f_i = \exp[-(\epsilon_i - \mu_i)/kT], \tag{3}$$

where μ_i is the chemical potential. Using Γ_{ij} and Γ_{ji}, we can write the net flux from site i to j in the form

$$J_{ij} = \Gamma_{ij} - \Gamma_{ji} \simeq \nu_0 e^{-\epsilon_{ij}/kT} (e^{\mu_i/kT} - e^{\mu_j/kT}), \tag{4}$$

where $\epsilon_{ij} = \max\{\epsilon_i, \epsilon_j\}$.

We now argue that as long as the antenna is only a small part of the DNA molecule, every protein adsorbs to DNA and desorbs many times before it locates the target. Therefore, outside the antenna there is statistical equilibrium between adsorbed and desorbed proteins, and hence proteins have a uniform chemical potential $\mu_i = \mu = kT \ln(c_3 b^3)$. Within the antenna, μ_i decreases when the site approaches the target and reaches $-\infty$ at the target site (see Fig. 1). If we label the border of the antenna as site 1 and the target as site $\lambda/b + 1$, using Eq. (4), we can write

$$\sum_{i=1}^{\lambda/b} J_{ij} e^{\epsilon_{ij}/kT} = \nu_0 (e^{\mu/kT} - e^{-\infty/kT}) = \nu_0 c_3 b^3, \tag{5}$$

where $j = i + 1$. Since the 1D current J_1 towards the target is the same at any antenna site—i.e., $J_{ij} = J_1$—we can find it as

$$J_1 = \frac{\nu_0 c_3 b^3}{\sum_{i=1}^{\lambda/b} \exp(\epsilon_{ij}/kT)} \simeq \frac{\nu_0 c_3 b^3 \sqrt{2\pi\sigma^2}}{(\lambda/b) \int_{-\infty}^{0} d\epsilon_{ij} R(\epsilon_{ij})}, \tag{6}$$

where $R(\epsilon_{ij})$ is given by

$$R(\epsilon_{ij}) = \sqrt{2\pi\sigma^2}\, g(\epsilon_{ij}) \exp(\epsilon_{ij}/kT) = \exp\left\{\frac{\sigma^2}{2(kT)^2} + \frac{w}{kT} - \frac{[\epsilon_{ij} - (w + \sigma^2/kT)]^2}{2\sigma^2}\right\}. \tag{7}$$

One can interpret Eq. (6) as Ohm's law, where the numerator plays the role of the voltage applied to the antenna and the denominator is the sum of resistances of all pairs (i,j) which are similar to Miller-Abrahams resistances for the hopping transport of electrons [16].

The sharp maximum value of the function $R(\epsilon_{ij})$ determining the sum of Eq. (6) is reached when $\epsilon_{ij} = \epsilon_{opt} = w + \sigma^2/kT$ and $R(\epsilon_{opt}) \sim \exp[\sigma^2/2(kT)^2 + w/kT]$. Thus

$$J_1 \sim \frac{D_3 c_3 b^2}{\lambda} \exp\left[\frac{|w|}{kT} - \frac{\sigma^2}{2(kT)^2}\right], \tag{8}$$

where we assumed for simplicity that $D_3 = D_1 \sim b^2 \nu_0$.

Before we move forward, we emphasize the crucial assumption already made in the above derivation. We assumed that λ is so long that within the antenna the sliding protein encounters sites with energy ϵ_{opt} more than once and, therefore, the sum in Eq. (6) can be replaced by the integral with limits from $-\infty$ to 0. We call such an antenna macroscopic. For a short antenna, the probability for such a site to appear inside is very small. Thus the sum in Eq. (6) is determined by the largest value of $R(\epsilon_{ij})$ typically available within the antenna. We call such an antenna mesoscopic.

Macroscopic antenna. We study the macroscopic antenna first. Using J_1 and J_3, our main *balance* equation for the rate J reads

$$J \sim D_3 c_3 \lambda \sim \frac{D_3 c_3 b^2}{\lambda} \exp\left[\frac{|w|}{kT} - \frac{\sigma^2}{2(kT)^2}\right]. \tag{9}$$

Thus the antenna length λ is obtained as

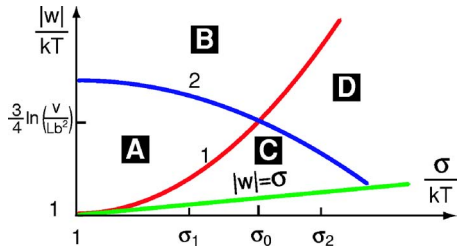

FIG. 2. (Color online) The phase diagram of the scaling regimes for $|w|>\sigma>kT$. Each line marks a smooth crossover between scaling regimes. The red line $|w|=3\sigma^2/2kT$ marks border 1 between macroscopic regimes (A,B) and mesoscopic regimes (C,D). The blue line $|w|=kT\ln(v/Lb^2)-\sigma^2/2kT$ marks border 2 between weak and strong adsorption regimes. They intersect at $\sigma_0=kT[(1/2)\ln(v/Lb^2)]^{1/2}$, $|w|=kT(3/4)\ln(v/Lb^2)$.

$$\lambda \sim b \exp\left[\frac{|w|}{2kT} - \frac{\sigma^2}{4(kT)^2}\right]. \tag{10}$$

Next we calculate the free protein concentration c_3. Suppose the one-dimensional concentration of nonspecifically adsorbed proteins is c_1. Assuming the antenna is only a small part of the DNA and remembering that adsorbed proteins are confined within a distance of the order of b from the DNA, we can write down the equilibrium condition as

$$\frac{c_1}{c_3 b^2} \sim \int f(\epsilon) e^{-\epsilon/kT} d\epsilon \sim \exp\left[\frac{|w|}{kT} + \frac{\sigma^2}{2(kT)^2}\right], \tag{11}$$

which must be complemented by the particle counting condition $c_1 L + c_3(v - Lb^2) = cv$. Since the volume fraction of DNA is always small, $Lb^2 \ll v$, standard algebra then yields

$$c_3 \simeq \frac{cv}{yLb^2 + v} \sim \begin{cases} c & \text{if } y < v/Lb^2, \\ cv/Lb^2 y & \text{if } y > v/Lb^2 \end{cases} \tag{12}$$

where y is $\exp[|w|/kT + \sigma^2/2(kT)^2]$. Equation (12) leads to two different scaling regimes, which are denoted as A and B in Fig. 2. In regime A, the nonspecific adsorption is relatively weak, $c_3 \sim c$; we arrive at

$$\frac{J}{J_s} \sim \exp\left[\frac{|w|}{2kT} - \frac{\sigma^2}{4(kT)^2}\right] \quad \text{(regime } A\text{)}. \tag{13}$$

In regime B, most proteins are adsorbed. Using the lower line of Eq. (12), we obtain

$$\frac{J}{J_s} \sim \frac{v}{Lb^2} \exp\left[-\frac{|w|}{2kT} - \frac{3\sigma^2}{4(kT)^2}\right] \quad \text{(regime } B\text{)}. \tag{14}$$

In both regimes, $|w|>\sigma^2/kT$; thus the σ term of $\ln(J/J_s)$ constitutes a correction. The size of the antenna grows with $|w|$; however, unproductive nonspecific adsorption of proteins on distant pieces of DNA grows with $|w|$ too and can slow down the transport to the specific target. These two effects compete; as a result, the rate enhancement J/J_s grows with w in regime A and declines in regime B. On the other hand, a growing σ reduces the antenna size and promotes nonspecific adsorption. Therefore, J/J_s decreases with σ in both regimes.

The above theory deals with a macroscopic antenna. To be macroscopic, the antenna has to contain at least one site with energy around ϵ_{opt}. The number of sites $n(\epsilon)$ with energy ϵ within the antenna is of the order of $\sim(\lambda/b)\exp[-(\epsilon - w)^2/2\sigma^2]$. Thus a macroscopic antenna requires $n(\epsilon_{opt}) > 1$, which gives $\lambda > \lambda_c = b\exp[\sigma^2/2(kT)^2]$. Since we know λ from Eq. (10), this condition can be written explicitly as $|w| > 3\sigma^2/2kT$. Hence, $|w| = 3\sigma^2/2kT$ is the border between the macroscopic regimes (A,B) and mesoscopic regimes (C,D) in Fig. 2. We can check that when $|w| > 3\sigma^2/2kT$, the condition $\epsilon_{opt} < 0$ is satisfied for the case of a macroscopic antenna. Now we are ready to switch to the case of a mesoscopic antenna and explain regimes C and D.

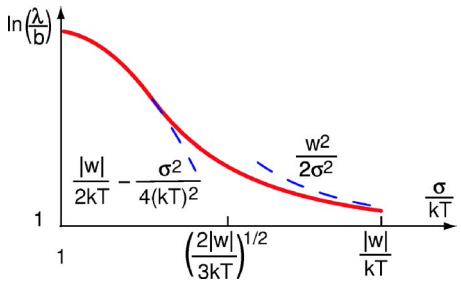

FIG. 3. (Color online) Dependence of the antenna length λ on the disorder strength σ. Dashed lines represent the asymptotic limits.

Mesoscopic antenna. In this case, the upper limit of the integral in Eq. (6) should be replaced by $\epsilon_\lambda \ll \epsilon_{opt}$ which is the largest energy typically available within the antenna. It can be estimated from $n(\epsilon_\lambda) \sim 1$; it is $\epsilon_\lambda \sim w + \sqrt{2}\sigma\sqrt{\ln(\lambda/b)}$. Using w and ϵ_λ, we can estimate the sum in Eq. (6) and get typical 1D current for the case of a mesoscopic antenna:

$$J_1(\lambda) \sim D_3 c_3 b \exp\left[\frac{|w|}{kT} - \sqrt{2\ln(\lambda/b)}\frac{\sigma}{kT}\right]. \tag{15}$$

Equation (15) is apparently different from Eq. (8), valid for the macroscopic antenna. This difference is partially related to the rate enhancement of 1D diffusion at a small time scale noticed for Gaussian disorder in computer simulations [12]. Equating $J_1(\lambda)$ to $J_3 \sim D_3 c_3 \lambda$, we obtain the antenna length

$$\lambda \sim b \exp\left[\left(\sqrt{\frac{|w|}{kT} + \frac{\sigma^2}{2(kT)^2}} - \frac{\sigma}{\sqrt{2}kT}\right)^2\right]. \tag{16}$$

We can check, with this λ, that the condition $\epsilon_\lambda < 0$ still holds. When $|w| < \sigma^2/2kT$, the antenna length $\lambda \sim b\exp(w^2/2\sigma^2)$. For a given adsorption energy w, the dependence $\lambda(\sigma)$ is plotted in Fig. 3. It shows that the decrease of the antenna length with growing disorder strength slows down when the antenna becomes mesoscopic.

The crossover from a relatively weak adsorption to a strong one described by Eqs. (12) again leads to the two scaling regimes for the case of a mesoscopic antenna. They are labeled C and D in Fig. 2. For relatively weak adsorption, when $|w| < \sigma^2/kT$, we obtain regime C, where

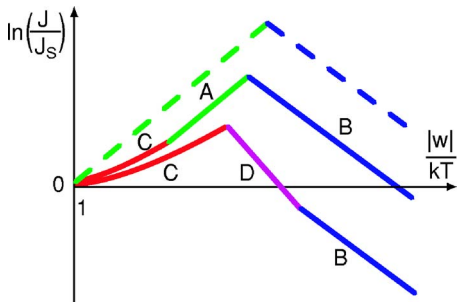

FIG. 4. (Color online) Schematic plot of the dependences of the rate enhancement J/J_s on $|w|$ at $\sigma=\sigma_1$ (upper solid curve) and $\sigma=\sigma_2$ (lower solid curve). Letters A, B, C, and D represent the domains of Fig. 2 they go through. The dashed line shows the limit case of the flat energy profile with $\sigma=0$.

$$\frac{J}{J_s} \sim \exp\left(\frac{w^2}{2\sigma^2}\right) \quad \text{(regime } C\text{)}, \tag{17}$$

while for strong adsorption we have regime D where

$$\frac{J}{J_s} \sim \frac{v}{Lb^2} \exp\left[-\frac{|w|}{kT} - \frac{\sigma^2}{2(kT)^2}\right] \quad \text{(regime } D\text{)}. \tag{18}$$

In experiment, the adsorption energy w can be controlled by the salt concentration changing the Coulomb part of protein-DNA interaction [17]. The dependences of $\ln(J/J_s)$ on $|w|$ at the two specified values of disorder strength σ_1 and σ_2 marked in Fig. 2 are schematically plotted in Fig. 4. For comparison, we also plotted the case of the flat energy profile ($\sigma=0$). In both cases with $\sigma>0$, $\ln(J/J_s)$ first grows proportional to w^2 (regime C), because the antenna is mesoscopic and thus 1D diffusion is faster, when compared to the normal diffusion at macroscopic antenna. For a relatively small disorder $\sigma=\sigma_1$, this rate enhancement continues to regime A but with a rate proportional to $|w|$ because the antenna grows to be macroscopic. For a larger disorder $\sigma=\sigma_2$, strong nonspecific adsorption of proteins on distant pieces of DNA slows down the search rate, when the antenna is still mesoscopic, and $\ln(J/J_s)$ decreases in regime D faster than it does in regime B. The antenna in regime B is macroscopic and $\ln(J/J_s)$ decreases proportional to $|w|$ for both $\sigma=\sigma_1$ and $\sigma=\sigma_2$.

The crossover from the weak disorder to the strong one happens at $\sigma \sim \sigma_0 = kT[(1/2)\ln(v/Lb^2)]^{1/2}$ (see Fig. 2). If one plugs in the achievable experimental conditions with $L/b \sim 150$ and $v \sim L^3$, the estimate of σ_0 is of the order of $2kT$, which falls in the range of estimates of σ from $1kT$ to $6kT$ used in the Refs. [12,14,15]. Apparently σ grows for proteins with a larger number of contacts with DNA and σ_0 decreases with DNA concentration. In order to identify the role of strong disorder, we look forward to more experiments dealing with relatively large concentrations of short straight DNA to guarantee that disorder strength satisfies $\sigma > \sigma_0$.

We know only one observation [17] of the peak in the coordinates of Fig. 4 but for a long and definitely coiled DNA where our theory is not directly applicable. In our recent paper [9], we presented a general theory including Gaussian coiled and globular DNA in the absence of disorder. In the current paper, we concentrated on the simplest regimes labeled A and D in Fig. 4(a) of Ref. [9] and still got the rather complicated diagram of Fig. 2 [18]. That is why we did not try to present our theory for more complicated regimes here.

Recently, the 1D sliding distance along DNA (which coincides with λ for straight DNA) was measured for the BbvCI restriction enzyme [19]. The authors claim that even in low-ionic-strength buffers, the 1D sliding is limited to distances <75 base pairs which do not exceed the DNA persistence length. In this case, our theory, designed for relatively short and, therefore, straight DNA, should be applicable.

We are grateful to A.Yu. Grosberg, S.D. Baranovskii, and J. Zhang for useful discussions.

[1] M. Ptashne and A. Gann, *Genes and Signals* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001).
[2] A. D. Riggs, S. Bourgeois, and M. Cohn, J. Mol. Biol. **53**, 401 (1970).
[3] P. H. Richter and M. Eigen, Biochemistry **2**, 255 (1974).
[4] O. G. Berg, R. B. Winter, and P. H. von Hippel, Biochemistry **20**, 6929 (1981).
[5] R. F. Bruinsma, Physica A **313**, 211 (2002).
[6] S. E. Halford and J. F. Marko, Nucleic Acids Res. **32**, 3040 (2004).
[7] S. E. Halford and M. D. Szczelkun, Eur. Biophys. J. **31**, 257 (2002).
[8] K. V. Klenin, H. Merlitz, J. Langowski, and C. X. Wu, Phys. Rev. Lett. **96**, 018104 (2006).
[9] Tao Hu, A. Yu. Grosberg, and B. I. Shklovskii, Biophys. J. **90**, 2731 (2006).
[10] O. G. Berg and P. H. von Hippel, J. Mol. Biol. **193**, 723 (1987).
[11] U. Gerland, J. D. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **99**, 12015 (2002).
[12] M. Barbi, C. Place, V. Popkov, and M. Salerno, Phys. Rev. E **70**, 041901 (2004); J. Biol. Phys. **30**, 203 (2004).
[13] K. Nadassy, S. J. Wodak, and J. Janin, Biochemistry **38**, 1999 (1999).
[14] M. Slutsky and L. A. Mirny, Biochemistry **87**, 4021 (2004).
[15] M. Slutsky, M. Kardar, and L. A. Mirny, Phys. Rev. E **69**, 061903 (2004).
[16] B. I. Shklovskii and A. L. Efros, *Electronic Properties of Doped Semiconductors* (Springer-Verlag, Berlin, 1984).
[17] R. B. Winter, O. G. Berg, and P. H. von Hippel, Biochemistry **20**, 6961 (1981).
[18] We assume $D_3=D_1$ in the absence of disorder. Thus, with disorder, $d=D_1/D_3<1$ which corresponds to the case represented by Fig. 4(a) of Ref. [9].
[19] D. M. Gowers, G. G. Wilson, and S. E. Halford, Proc. Natl. Acad. Sci. U.S.A. **102**, 15883 (2005).