Proteins searching for their target on DNA by one-dimensional diffusion: overcoming the “speed-stability” paradox

Shi Yu · Shihu Wang · Ronald G. Larson

Received: 24 January 2013 / Accepted: 27 February 2013 / Published online: 5 April 2013
© Springer Science+Business Media Dordrecht 2013

Abstract The sequence dependence of DNA-protein interactions that allows proteins to find the correct reaction site also slows down the 1D diffusion of the protein along the DNA molecule, leading to the so-called “speed-stability paradox,” wherein fast diffusion along the DNA molecule is seemingly incompatible with stable targeting of the reaction site. Here, we develop diffusion-reaction models that use discrete and continuous Gaussian random 1D diffusion landscapes with or without a high-energy cut-off, and two-state models with a transition to and from a “searching” mode in which the protein diffuses rapidly without recognizing the target. We show the conditions under which such considerations lead to a predicted speed-up of the targeting process, and under which the presence of a “searching” mode in a two-state model is nearly equivalent to the existence of a high-energy cut-off in a one-state model. We also determine the conditions under which the search is either diffusion-limited or reaction-limited, and develop quantitative expressions for the rate of successful targeting as a function of the site-specific reaction rate, the roughness of the DNA-protein interaction potential, and the presence of a “searching” mode. In general, we find that a rough landscape is compatible with a fast search if the highest energy barriers can be avoided by “hopping” or by the protein transitioning to a lower-energy “searching” mode. We validate these predictions with the results of Brownian dynamics, kinetic Metropolis, and kinetic Monte Carlo simulations of the diffusion and targeting process, and apply these concepts to the case of T7 RNA polymerase searching for its target site on T7 DNA.

Keywords Protein–DNA interaction · Sliding · Two-state model · Protein targeting

---

Electronic supplementary material The online version of this article (doi:10.1007/s10867-013-9310-3) contains supplementary material, which is available to authorized users.

S. Yu · S. Wang · R. G. Larson
Department of Chemical Engineering, University of Michigan, Ann Arbor, MI 48109-2136, USA

R. G. Larson (✉)
A150 NCRC, Building 10, Ann Arbor, MI 48105, USA
e-mail: rlarson@umich.edu

1 Introduction

The long-standing problem of proteins (repressor, polymerase, transcription factor, etc.) seeking their target sites among millions of nonspecific sites along DNA is attracting increasing interest. To help explain how proteins might find their targets faster than allowed by ordinary three-dimensional diffusion, in 1981 Berg et al. [1, 2] proposed a “facilitated diffusion” model for this process. Their model, as further elaborated by others, includes four mechanisms: (a) one-dimensional diffusion or “sliding” of a protein along the DNA contour, (b) “hopping” of a protein from one site on DNA to a nearby site a few base pairs away, (c) “jumping” of the protein to a distant site, mediated by 3D diffusion through solvent, and (d) “intersegment transfer” and “intersegmental jumps” [3] of the protein to a site on a DNA sequence close by in 3D space, but possibly many base pairs distant along the DNA contour (since the DNA typically has a coiled configuration). Since then, a large number of experiments [4–12] have been carried out to test this model, both *in vitro* and *in vivo*. Many models have been developed to describe the protein targeting, through a combination of 1D and 3D diffusion [13–15]. However, since the 1D diffusion coefficient of a protein moving along DNA is two or three orders smaller than the 3D diffusion coefficient, long-range sliding of a protein along DNA may slow down rather than speed up the targeting process relative to the predictions of the Smoluchowski equation for three-dimensional diffusion-limited targeting [16, 17], a result that has been confirmed by Brownian dynamics (BD) simulations [18, 19]. Recently, Halford [20] reexamined the early experimental results [21] and claimed that the targeting speed of DNA-binding proteins (DBPs) does not exceed the true 3D diffusion limit, if one includes in the 3D diffusion the effect of electrostatic attraction between the cationic proteins and the anionic DNA in the Smoluchowski equation. In other words, the so-called “facilitated diffusion” mechanism does not really facilitate the targeting, but rather impedes it. Halford’s conclusion remains controversial, however, and in any event single-molecule imaging experiments leave little doubt that proteins do actually slide along DNA [6, 7, 22]. Hence, the 1D diffusion of the protein along DNA remains a very interesting and important problem to understand in greater detail, regardless of whether it actually speeds up the target search or not. Most DNA-binding proteins (DBPs) interact with 5 to 20 base pairs at a time when forming nonspecific complexes with DNA. Both statistical mechanical analysis [23] and experiments [24] demonstrate that the interaction energies between a DBP and each base pair are almost independent of each other and additive. Because of the irregularity of the DNA base pair sequence, it is assumed that the sequence-dependent binding energy can be approximated by an uncorrelated (or random) Gaussian distribution [25]. The problem of 1D diffusion of particles in a random potential is well studied problem [26, 27] and it has also been used to model the motion of dislocations in disordered crystals [28], the unzipping of DNA [29], protein folding [30], and other problems. Zwangzig [31] determined theoretically the simple exponential relationship between the 1D diffusion coefficient D in a random Gaussian one-dimensional potential and the standard deviation, σ , or roughness, of the potential (1), which was subsequently rediscovered by Slutsky and Mirny [25, 32]:

$$D = D_0 \exp \left[- \left(\frac{\sigma}{k_B T} \right)^2 \right] \quad (1)$$

Based on (1), Slutsky and Mirny noted that a fast search process requires that the roughness of the one-dimensional potential be small ($\sigma = 1 \sim 2 k_B T$), which seemingly conflicts

with the requirement that the DNA-protein specific complex be thermodynamically stable once it forms ($\sigma \sim 5 k_B T$) [25]. To resolve this so-called “search-speed/stability paradox”, Slutsky and Mirny proposed a “two-state model”, which allows the protein to change its conformation to perform fast diffusion in a “searching” state and bind stably to the target site in a “recognition” state. A recent single-molecule experiment [8] has confirmed that p53, and fragments of p53 that contain either the “recognition domain” or the “sliding domain,” diffuse differently along DNA in a manner consistent with this “two-state” mechanism.

However, since the DBP finds its target by a combination of 1D and 3D diffusion and the transition from 1D to 3D diffusion should be sequence-dependent, especially for a rough potential, a Gaussian random potential might be an unrealistic model of the actual 1D potential landscape that the protein experiences when sliding along the DNA major groove. One important source of non-randomness is the likely relatively high-energy barrier to motion from one site to the next along the DNA, since the protein–DNA interaction is dominated by highly localized hydrogen bonds between DNA and protein that must be broken for the protein to slide a distance of one base pair along the DNA. Since these barriers must be overcome whether the roughness of the landscape is small or large, (1) may overestimate the effects of potential roughness on the 1D diffusion coefficient. Another consideration is that the sequence-independent electrostatic interaction might decay more slowly in the direction perpendicular to the DNA than do sequence-dependent forces, and hence this relatively weak electrostatic force might by itself be sufficient to reduce the dimensionality of protein Brownian motion from 3D to 1D [33]. A protein can diffuse a very short distance away from the DNA surface to get over a high-energy barrier resulting from sequence-dependent forces, while easily staying close enough to the DNA that electrostatic forces can maintain the 1D diffusion. Thus, the real 1D potential landscape along DNA that the protein experiences during sliding might be better approximated by imposing a maximum-energy threshold onto an otherwise Gaussian-distributed positionally random potential. So, in this paper, we discuss how introducing a threshold into the 1D random potential can reduce the dependence of 1D diffusivity of an over-damped particle on the strength of a Gaussian disordered potential. We consider both a discrete 1D potential landscape and also propose a continuous random 1D potential based on Barbi’s hydrogen-bonding model [34, 35] for the sequence-dependent interaction potential, which we will show reconciles fast sliding with stability of targeting. We also compare the diffusivities along random potentials with wells of various shapes to determine the sensitivity of 1D diffusivity to the shape, as well as the depths, of random potential wells. We obtain diffusivities for these landscapes in three different ways and show that they agree with each other. Using these diffusivities, we determine using both theory and simulations the rate at which the protein finds its target in reaction- and diffusion-limited regimes.

Unlike synthetic nanoparticles, a protein can change its conformation frequently. We therefore also propose a simple two-state model, similar to that of Mirny and coworkers, to explore whether the sliding motion of DBPs can be accelerated substantially by switching their conformations. In our simple model, the transitions between the two states on different sites are controlled by a constant activation barrier; in this respect our model differs from most previous two-state models [32, 36]. In addition, we choose transition rates between states that satisfy detailed balance, guaranteeing that our model does not violate thermodynamic requirements [37]. We then carry out kinetic Monte Carlo simulations to see if this simple two-state model improves the targeting process.

Before discussing our diffusion models, we address some general conditions that govern the rate of target search and the conditions under which this search is controlled by the

overall one-dimensional diffusivity along the DNA, or by the rate of reaction at the target site, or by both. The “search-speed/stability paradox” mentioned above arises only if both the one-dimensional diffusion coefficient and the rate of reaction at the target site *both* influence the overall targeting rate. If the rate of reaction at the target site is slow compared to the rate of diffusion, then most proteins will not react the first time they reach the target site, but will occupy sites along the DNA with probability given by simple Boltzmann statistics, governed by the site-specific free energy. The rate of reaction will then be controlled by this occupancy probability and the reaction rate constant, and not by the diffusion speed. On the other hand, if the reaction rate at the target site is very fast, then any protein reaching the target will react immediately, irrespective of whether the protein remains on the site for very long, and the “stability” of the binding site will be unimportant, while the speed of diffusion dominates the targeting rate. To demonstrate this quantitatively, in the next section we derive the targeting rate for a simple 1D reaction-diffusion model, and then show how the effective diffusion coefficient and effective reaction rate are influenced by properties of the one-dimensional potential.

2 Theory and models

2.1 Diffusion-limited vs. reaction-limited targeting

We assume that away from the binding site, the targeting proteins can associate with DNA with overall rate constant k_{on} , and dissociate from DNA with rate constant k_{off} . The bulk concentration of protein can be considered to be a constant c_{bulk} . Then, the concentration profile of the protein on DNA can be approximated simply by a steady-state analysis balancing flux of proteins to and from the DNA and 1D diffusion along the DNA molecule. At steady state, we have:

$$D \frac{\partial^2 c}{\partial x^2} + k_{on} c_{bulk} - k_{off} c = 0 \quad (x \leq 0) \quad (2)$$

where D is the effective 1D diffusion coefficient of protein along DNA, in units of inverse time, c is the protein concentration (or probability of occupancy) along the DNA, and x is a dimensionless distance along the DNA molecule in units of base pairs. To simplify the analysis, we divide the one-dimensional domain into a “far field,” where we can take the diffusivity D , and k_{on} , and k_{off} to be constants that are pre-averaged over the free-energy landscape, and the “near field”, which is the target site, where we will take account of the strength of binding of the DNA, which will affect its probability of reacting there. Within this approximation (which we will later test using simulations), the concentration of protein along the DNA can be obtained by applying the boundary conditions:

$$\begin{aligned} c &= c_{\infty} = (k_{on}/k_{off}) c_{bulk} \quad (x \rightarrow -\infty) \\ -2D \frac{\partial c}{\partial x} - \Omega^{eff} c &= 0 \quad (x = 0) \end{aligned} \quad (3)$$

where Ω^{eff} is the effective reaction rate constant at the target site ($x = 0$). We only consider proteins to the left (negative side) of the target site ($x \in (-\infty, 0]$) and account for diffusion from the positive side to the target site at $x = 0$ by the factor of 2 multiplying the diffusivity in (3). The effective reaction rate $\Omega^{eff} = \Omega \exp \left[\frac{\mu - E_T}{k_B T} - \frac{1}{2} \left(\frac{\sigma}{k_B T} \right)^2 \right]$

(see [Supporting Material](#)) accounts for the residence time of the protein at the target site as well as rates of transition to and from one “state” to the other in the case of a two-state model, as will be discussed below. We also assume that $k_{off} \ll D$, so that 1D diffusion is an important part of the target-searching. (If $k_{off} > D$, the proteins will not remain on the DNA long enough to slide one dimensionally and so will reach the target site primarily by three-dimensional diffusion.) Since we take $k_{off} \ll D$, we can neglect the terms involving k_{off} and k_{on} in the second boundary condition above, so that we disallow the protein to hop on or off the DNA at the target site. (The small contribution of this could be easily included, if desired.) Solving (2) and (3), the protein concentration along the DNA is:

$$c = \frac{k_{on}}{k_{off}} c_{bulk} - \frac{k_{on}}{k_{off}} c_{bulk} \left(\frac{\Omega^{eff}}{2D\sqrt{\frac{k_{off}}{D}} + \Omega^{eff}} \right) \exp \left(x\sqrt{\frac{k_{off}}{D}} \right) \quad (4)$$

where we only consider $x < 0$. Note in the above, that the typical *diffusion distance* traveled by the protein between its adsorption from bulk solution and its subsequent desorption, in units of base pairs, is given by $\sqrt{D/k_{off}}$. The reaction rate is given by:

$$r = \Omega^{eff} c_{x=0} = k_{on} c_{bulk} \left(\frac{2\sqrt{D/k_{off}}}{1 + \frac{2\sqrt{Dk_{off}}}{\Omega^{eff}}} \right) \quad (5)$$

When $\Omega^{eff} \gg \sqrt{Dk_{off}}$, the targeting process is *diffusion limited*, the concentration of protein at the target site is very small, and the reaction rate is given by:

$$r = 2k_{on} c_{bulk} \sqrt{\frac{D}{k_{off}}} = 2 \frac{(k_{on}/k_{off}) c_{bulk} D}{\sqrt{D/k_{off}}} = \frac{2c_{\infty} D}{\sqrt{D/k_{off}}} \quad (6)$$

This rate, r , can also be written as the product of the diffusivity D times the 1D concentration gradient of protein, which is the concentration of protein far from the target, namely $c_{\infty} \equiv (k_{on}/k_{off}) c_{bulk}$, divided by the diffusion distance $\sqrt{D/k_{off}}$. The factor of two again arises because the protein can diffuse to the target from either the right or the left.

We performed a series of Metropolis Monte-Carlo simulations over a periodic landscape of 2,000 base pairs with different roughness using the Metropolis scheme to verify (6). The simulations were performed by randomly introducing a protein along the periodic landscape, and simulating its diffusion, until it either reached the target site and reacted, or dissociated from the DNA. The sites were treated as discrete with energies E_i distributed with Gaussian distribution, $E_i = N(0, \sigma)$, and hopping from one site to the next was accomplished by the usual Metropolis method. Thus, randomly selected diffusion moves between neighboring sites were accepted with probability given by the Boltzmann factor for the energy difference between the sites, if the move is unfavorable, or unity otherwise. The probability of dissociation $k_{off,i}$ was also site-specific with probability given by a Boltzmann of the energy difference between the site and the energy in free solution, i.e., $k_{off,i} = \exp[-(E_{bulk} - E_i)/k_B T]$. We adjusted the energy in free solution E_{bulk} according to the roughness of the potential along DNA by setting $E_{bulk} = -\ln(D_0/250^2) - (\sigma/k_B T)^2/2$ to fix the value of the diffusion distance at $\sqrt{D/k_{off}} \approx 250$, since $D \approx D_0 \exp\left[-\left(\frac{\sigma}{k_B T}\right)^2\right]$,

and $k_{off} \approx \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n k_{off,i} = \exp \left[-E_{bulk} + (\sigma/k_B T)^2 / 2 \right]$, as discussed further below. The reaction rate is just the inverse of the average time between reaction events times the protein “concentration,” which is $1/2,000$ in this case, since proteins are tracked one at time over a landscape with 2,000 sites. The results, plotted in Fig. 1, show that (6) remains correct even for a rough potential landscape with standard deviation up to $\sigma = 2 k_B T$. Note that in the upper insert, the occupancy of sites along the DNA is very “noisy” even for a relatively small roughness of $\sigma = 0.5 k_B T$, not due to insufficient averaging over multiple proteins, but due to this roughness, which strongly favors some sites over neighboring ones. As the potential landscape becomes rougher, the noise of both the site occupancy and the local diffusion coefficient increases greatly, but (6) holds up to at least $\sigma = 2 k_B T$, showing that our use of a far-field average diffusivity to compute the rate of diffusion to the reaction site remains valid at least to this level of roughness. (The occupancy of sites for $\sigma = 2 k_B T$ is even “noisier” than for $\sigma = 0.5 k_B T$ as shown in the bottom inset of Fig. 1.) For σ higher than this, the diffusion becomes too slow to obtain adequate averaging; however, the use of a far-field average diffusivity will eventually break down for large enough σ . We can estimate the point at which this will occur by considering the landscape-dependent variance of D_n , the diffusion coefficient averaged over the diffusion distance $n = \sqrt{D/k_{off}}$. As shown in the [Supporting Material](#), the variance of the D_n/D_0 can be estimated by $2\sigma^4/n$. So, when the diffusion distance is about 100 ($n \approx 100$, typical value for DBP’s), the variance of D_n over this diffusion distance will be significant for σ larger than $3 k_B T$. Thus, for $\sigma > 3 k_B T$, instead of using the far-field 1D diffusion coefficient for D in (6), we would need to compute D_n from the local base pair sequence neighboring the target to determine the reaction rate in the *diffusion-limited* regime.

Fig. 1 Steady-state reaction rate of a protein at the target site (site number = 1,000), assuming instantaneous reaction at the reaction site, for a discrete Gaussian-distributed landscape, obtained by Metropolis Monte-Carlo simulations (squares). Also shown is the reaction rate from (6), using MC simulations to determine the diffusion coefficient (dashed line); and using $D = \exp[-(\sigma/k_B T)^2]$ to approximate the diffusion coefficient (continuous line). The insets show the rough concentration profiles over the 2,000 base pairs for $\sigma = 0.5 k_B T$ and $\sigma = 2.0 k_B T$

On the other hand, when $\Omega^{eff} < \sqrt{Dk_{off}}$, the targeting process becomes *reaction limited* and the reaction rate is given by:

$$r = (k_{on}/k_{off}) c_{bulk} \Omega^{eff} = c_{\infty} \Omega^{eff} \quad (7)$$

Thus, in the reaction-limited regime, the protein targeting speed is only determined by the effective reaction rate at the target, and by the overall concentration of protein on the DNA, which is given by $(k_{on}/k_{off})c_{bulk}$, and not by the diffusivity. Note that the effective reaction rate must include the effect of the residence time of the protein at the target site, and thus by the stability of binding to the target. In the reaction-limited regime, diffusion is fast enough compared to reaction that the protein concentration, or probability distribution, along the DNA reaches thermodynamic equilibrium. At the target site, the protein finds a local free-energy minimum, and so its concentration there will be higher than elsewhere according to the ratio of Boltzmann factors for the target site relative to the average over all sites.

Recent experiments [11] have shown that *lac* repressor frequently (>90%) diffuses over its natural operator several times before binding, which indicates the targeting of *lac* repressor, is not diffusion-limited. Therefore, determination of conditions for diffusion-limited vs. reaction-limited targeting, as described above, is an important element of the analysis of the target search on DNA. One aspect of targeting *in vivo* not considered here is the role of proteins, such as histones, which can block 1D diffusion along DNA in the cell. Our current theory is applicable only to *in vitro* single-molecule experiments that lack such blockages to diffusion, but in the future, the effect of blocking proteins will need to be considered.

2.2 Diffusion models

In this section, we limit ourselves to the diffusion-limited case, where the protein targeting speed is determined by the 1D diffusion coefficient, and hence the targeting speed is given in (6). We will determine the “far-field” protein diffusivity for models with three different 1D potentials, namely 1) a Gaussian-distributed square-well potential with a threshold, 2) a Gaussian-distributed square-well potential with a threshold and regular barriers, and 3) a continuous disordered potential. These potentials are illustrated in Fig. 2a below. We will also consider 4) a two-state model (Fig. 2b), where the protein can switch between a fast “searching state” with a weak or flat potential and a “recognition state” where it feels a rough Gaussian-distributed random potential.

We can calculate the diffusion coefficient along a random energy landscape using two different approaches. One is to compute the mean first passage time (MFPT) for the protein to diffuse from a starting point $y = 0$ (where $y = 0$ is also a reflecting point to prevent particles from diffusing towards negative y) to a finishing point $y = L$ along a randomly generated landscape, and then average this MFPT over multiple randomly generated landscapes. This gives an average time t_{ave} to diffuse a given distance L , and the diffusion coefficient can then be estimated from the standard formula $D \approx L^2/2t_{ave}$. This MFPT can be obtained from the more general formula for a particle starting at position y [38]:

$$t(y) \propto \frac{1}{D_0} \int_y^L dy_1 e^{U(y_1)/k_B T} \int_0^{y_1} e^{-U(x)/k_B T} dx \quad (8)$$

and taking $y = 0$ in our calculation.

Fig. 2 **a** 1D disordered potential landscapes (from *top to bottom*): I. Gaussian random potential with threshold; II. Threshold Gaussian random potential with activation barriers between each site; III. Sinusoidal disordered potential with threshold.

b two-state model with transition rates between different sites or states (given by the ω 's) determined by (19). “R” is the recognition state and “S” the search state. We will show later that in some cases the two-state model gives a diffusivity nearly identical to a one-state set model with a low threshold set by the searching state energy as shown in the bottom diagram

The second approach to obtain an analytical solution for the long-time D is to use the formula for a periodic potential derived by Festa and d’Agliano [39], namely:

$$D = \frac{D_0}{\langle e^{U/k_B T} \rangle \langle e^{-U/k_B T} \rangle} \quad (9)$$

where the brackets indicate an average over the period of the potential:

$$\langle e^{U/k_B T} \rangle = \frac{1}{L} \int_0^L e^{U(x)/k_B T} dx \quad (10)$$

This formula can be used for a non-periodic random potential by using quasi-random potentials that are random over a finite distance, beyond which they are periodic. The result for a random potential is then obtained by taking the limit of a long repeat distance L , i.e., taking $L \rightarrow \infty$.

2.2.1 Gaussian-distributed square-well potential with threshold

As discussed above, the hopping/jumping of a protein can be highly sequence-dependent. Once the protein encounters a high free-energy barrier, it can retreat back, diffuse over the barrier, or be expelled into bulk solution. Because of this third possibility, the real energy landscape that the protein actually experiences cannot strictly be Gaussian disordered, since the high-energy tail of the distribution will have energy higher than the protein in free solution, and the protein will likely escape to free solution if a sequence with high enough energy in this tail is encountered. To simplify our model, we therefore assume that the protein will never feel a net energy that is higher than energy $E_{threshold}$. Thus, after generating an energy landscape from a Gaussian distribution, we set any energy higher than $E_{threshold}$ to $E_{threshold}$. This assumes that when the threshold energy is encountered, the protein briefly loses attachment to the DNA and then quickly binds to a neighboring site, rather than disappearing into the bulk. Hence, we are only including very short “hopping” motions of the protein, but not longer-range “jumping” or “intersegment transfer.” We will also show below that in some cases, the diffusivity for the “two-state” model introduced below can be approximated by a one-state model with a low threshold energy, which provides additional motivation for studying this simple model. Thus, for these cases, our potential landscape can be approximated by random square-well potential with threshold, given below:

$$E_i = \min [N(\mu, \sigma), \quad E_{threshold} + \mu], \quad i = 1, 2, 3, \dots \quad (11)$$

We now compute the diffusivity over this potential using both the MFPT and the Festa and d’Agliano expression (9) and (10) above, with $L \rightarrow \infty$, and taking the mean site free energy $\mu = 0$ $k_B T$. For a Gaussian random potential with threshold (11), this yields:

$$D = \frac{D_0}{\left\{ e^{\tilde{\sigma}^2/2} \left[1 - \Phi \left(\frac{-\tilde{v} + \tilde{\sigma}^2}{\tilde{\sigma}} \right) \right] + e^{\tilde{v}} \left[1 - \Phi \left(\frac{\tilde{v}}{\tilde{\sigma}} \right) \right] \right\} \left\{ e^{\tilde{\sigma}^2/2} \Phi \left(\frac{\tilde{v} + \tilde{\sigma}^2}{\tilde{\sigma}} \right) + e^{-\tilde{v}} \left[1 - \Phi \left(\frac{\tilde{v}}{\tilde{\sigma}} \right) \right] \right\}} \quad (12)$$

where we have made the equation more compact by defining $v = E_{threshold}$, $\tilde{v} = v/k_B T$; $\tilde{\sigma} = \sigma/k_B T$. In the following discussion, all parameters with a tilde will represent these parameters normalized by $k_B T$. In the above, Φ is the cumulative distribution function of the normal distribution:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x}{\sqrt{2}} \right) \right] \quad (x \in \mathbb{R}) \quad (13)$$

Equation (12), which is based on the equation of Festa and d’Agliano, agrees very well with the numerical results we obtain from the MFPT, which are plotted in Fig. 3a. When $\tilde{\sigma}$ is large ($\tilde{\sigma} \gg |\tilde{v}|$), (12) can be reduced to:

$$D = 2D_0 \exp \left(-\tilde{v} - \frac{1}{2}\tilde{\sigma}^2 \right) \quad (14)$$

As shown in Fig. 3b, (12) and (14) agree with each other for $\tilde{\sigma} \gg |\tilde{v}|$, whether $\tilde{v}$ is positive or negative. We will later show that a very similar formula also results from analysis of the diffusivity in the two-state model, which we will discuss later.

Fig. 3 **a** Normalized diffusivity D/D_0 vs. Gaussian standard deviation σ for Model I, namely a Gaussian-distributed potential with a threshold (11). Symbols are obtained numerically from the MFPT, while lines are calculated from (12). Note that σ describes the standard deviation of the Gaussian potential, while the true standard deviation is lower than this because of the cut-off applied to the tail of the Gaussian. **b** Normalized diffusivity D/D_0 vs. Gaussian standard deviation σ for different energy threshold, *lines* are obtained by (12), *symbols* are obtained by (14)

For a Gaussian disordered potential with a threshold (11), the diffusion coefficient depends strongly on both $E_{\text{threshold}}$ (or $\bar{\nu}$) and σ . We note here that σ is the standard deviation of the Gaussian energy distribution before applying the cut-off, and the standard deviation of the actual potential is therefore less than σ . When $E_{\text{threshold}}/\sigma$ is large, only a very small fraction of the Gaussian-distributed energies are cut off by $E_{\text{threshold}}$. Therefore, for small σ , (12) almost converges to (1) ($E_{\text{threshold}} > 0 k_B T$). The difference between (12) and (1) increases with σ . According to a model of Slutsky and Mirny [25], σ should be at least $5 k_B T$ to maintain the stability of the DNA-protein specific complex. In this case, the value of D given by (12) can be as much as 10^4 times higher than that given by (1), if $E_{\text{threshold}}$ is small enough ($\sim 3 k_B T$); see Fig. 3a. By calculating the partition function numerically, we find that the fraction of time that the protein spends on the target site when system is in equilibrium for small $E_{\text{threshold}}$ ($E_{\text{threshold}} = 3 k_B T$) and large roughness ($\sigma = 5 k_B T$) is about 0.25, which is the same value as obtained for the Gaussian-distributed potential without threshold [25], because the threshold only affects high-energy states that contribute little to the partition function. Therefore, if the highest-energy states along the DNA can be eluded by hopping/jumping of the protein short distances along an otherwise Gaussian-distributed random potential (or by transitioning to a searching mode as discussed below), then the diffusion coefficient can be increased greatly without compromising the fidelity

of targeting. Thus, the “speed-stability paradox” might be overcome by brief excursions of the protein from the DNA surface when the high-energy states are encountered, thus reducing the effective roughness of the landscape. We also note that these few high-energy states might be selected against by mutation of the few base pairs responsible for them, and this would be favored by natural selection if they indeed have a large negative effect on protein mobility. In any case, we might expect that since the high-energy states have a large effect on mobility but a small effect on binding stability, we should consider the possibility that they are somehow avoided, and we have modeled this avoidance by a simple Gaussian landscape with high energy-states truncated by a threshold.

2.2.2 Gaussian-distributed square-well potential with threshold and regular barriers

During translocation from one site to the next site, the protein may have to overcome activation barriers (E_a) because of charge mismatch or hydrogen bond deformation and breakage. Thus, we consider a random potential augmented by a periodic potential to add activation barriers into our model. All activation barriers E_a and site potential wells are modeled by square barriers with width α and square wells of the width β , respectively ($\alpha + \beta = 1$).

$$E_{2j} = \min [N(\mu, \sigma), E_{\text{threshold}} + \mu], \quad E_{2j+1} = E_a \quad (j = 0, 1, 2, 3, \dots) \quad (15)$$

where $N(\mu, \sigma)$ is again a normal (or Gaussian) distribution with mean μ and standard deviation σ . For simplicity, let $\mu = 0$, $E_{\text{threshold}} = E_a = v$ (shown in Fig. 2a II). By applying (9) to this random potential, we have:

$$D = \frac{D_0}{\left\{ \alpha + \beta \left[e^{-\tilde{v} + \tilde{\sigma}^2/2} \left[1 - \Phi \left(\frac{-\tilde{v} + \tilde{\sigma}^2}{\tilde{\sigma}} \right) \right] + 1 - \Phi \left(\frac{\tilde{v}}{\tilde{\sigma}} \right) \right] \right\} \left\{ \alpha + \beta \left[e^{\tilde{v} + \tilde{\sigma}^2/2} \Phi \left(\frac{\tilde{v} + \tilde{\sigma}^2}{\tilde{\sigma}} \right) + 1 - \Phi \left(\frac{\tilde{v}}{\tilde{\sigma}} \right) \right] \right\}} \quad (16)$$

The relations between 1D diffusion coefficients and the roughness of different random potentials are plotted in Fig. 4. Numerical results obtained using the MFPT agree well with analytical results based on the equation of Festa and d’Agliano (16).

Fig. 4 Normalized diffusivity D/D_0 vs. roughness σ for Model II, which is a Gaussian-distributed potential with a threshold and an energy barrier (15), where the threshold and barrier are the same height. Symbols are obtained from the MFPT while lines are obtained from (16), with $v = E_{\text{threshold}}$, $\alpha = \beta$

When high activation barriers are introduced into this random potential in Model II, the sliding of the protein is dominated by these activation barriers. Thus, the random potential has even less effect on the 1D diffusion. Such periodic barriers are probably important for DNA-protein systems, because many DNA-protein nonspecific complexes are stable enough for the structure of those complexes to be resolved [40]. While such barriers make the diffusivity less sensitive to landscape roughness, they will slow down the overall one dimensional diffusion, relative to that for bulk 3D diffusion. (Note that the diffusivity plotted in Fig. 4 is normalized by the diffusion coefficient for no roughness, but with barriers.)

2.2.3 Sinusoidal disordered potential

In reality, of course, the DNA-protein interaction energy landscape is not discrete. So, here we also consider a continuous potential landscape with continuous first derivative, which is given by:

$$E(x + i \cdot l) = E_i \left[\sin^{2n} \left(\pi \frac{x}{l} \right) \right], \quad (0 \leq x < l, \quad i = 1, 2, 3, \dots) \quad (17)$$

where l is the width of each “well” of the potential, where the potential returns to zero at positions $0, l, 2l$, etc., and each well has a depth of E_i (a negative quantity) halfway between these points. The exponent $2n$ sets the shape of this periodic potential, with larger n leading to steeper walls of the potential wells. A periodic potential is obtained by setting all E_i to the same value E .

We now obtain the relationship between D and potential well depth E for periodic potentials, including the continuous periodic potential above, with $n = 1$, and 1,000, as well as the periodic potential with alternating square wells and square barriers of the same width using (9). The results are presented in Fig. 5. Note that for shallow potential wells ($< 3 k_B T$), D decreases slowly as the well depth increases and then drops exponentially with increasing

Fig. 5 D vs. periodic potential well depth. The dashed line is for the square-well potential, and the solid line is for the continuous sinusoidal potential, (17) with $n = 1$, while the dotted line is the sinusoidal potential with $n = 1000$

energy well depth at depths greater than around $5 k_B T$. Taking the continuous well wall steepness towards infinite steepness by setting n to infinity, the periodic potential ceases to impede diffusing particles, as it becomes flat with infinitely narrow, delta-function, peaks. This illustrates that the shape of the 1D potential along DNA strongly affects the diffusivity.

2.2.4 Two-state model

To investigate the diffusion of proteins that might use two domains to bind DNA or switch between two conformations, we propose a two-state model similar to that suggested by Slutsky and Mirny, in which a protein can switch between a “searching” (S) state and “recognition” (R) state on any site. However, we do not allow dissociation of the protein from the DNA. For both states, we assume the site energies to be Gaussian distributed:

$$E_i^S = N(\mu^S, \sigma^S), \quad E_i^R = N(\mu^R, \sigma^R) \quad (18)$$

We assume the protein can slide along the DNA in both S and R states (as p53 does) with the S state allowing for much faster diffusion than the R state does ($D^S \gg D^R, \sigma^S \ll \sigma^R$). Both the sliding along DNA and the conformational switching are controlled by activation barriers. The diffusive properties of the two-state model can be characterized by a sliding transition rate in the S state $\omega_{i,i\pm 1}^S$, a sliding transition rate in the R state $\omega_{i,i\pm 1}^R$, a transition rate from the S to the R state ω_i^{SR} , and a transition rate from the R to the S state ω_i^{RS} . These rates are given by:

$$\begin{aligned} \omega_{i,i\pm 1}^S &= \omega_0 \exp[-(\tilde{E}_a^S - \tilde{E}_i^S)] \\ \omega_{i,i\pm 1}^R &= \omega_0 \exp[-(\tilde{E}_a^R - \tilde{E}_i^R)] \\ \omega_i^{SR} &= \omega_0 \exp[-(\tilde{E}_a^C - \tilde{E}_i^S)] \\ \omega_i^{RS} &= \omega_0 \exp[-(\tilde{E}_a^C - \tilde{E}_i^R)] \end{aligned} \quad (19)$$

$\tilde{E}_a^S$, $\tilde{E}_a^R$, and $\tilde{E}_a^C$ are activation barriers (normalized by $k_B T$) for diffusing from one site to the next in the “searching” state, for diffusing in the “recognition” state, and for conformational transitioning from the search to the recognition state, respectively. $\tilde{E}_i^S$ ($\tilde{E}_i^S \equiv E_i^S/k_B T$) and $\tilde{E}_i^R$ ($\tilde{E}_i^R \equiv E_i^R/k_B T$) are Gaussian-distributed site-specific (normalized) energies in the “searching” state and in “recognition” state, respectively, where, without loss of generality, the S state can be taken to have zero average energy ($\tilde{\mu}^S = 0$). This is equivalent to interpreting $\tilde{\mu}^R$ as the energy difference between the recognition state and the searching state. We use the same prefactors for all the rates above, but note that we can absorb differences in prefactors into the apparent activation energies, since we are not considering effects of varying temperature here. Note that the ratios of forward to reverse rates of transition between any two microstates (i to $i \pm 1$ in either S or R configuration and S to R or R to S at any site i) are given by Boltzmann functions of the energy difference between the two states. Thus, detailed balance is satisfied and at a long time the protein will sample an equilibrium distribution of states. Since the site-specific free energies $\tilde{E}_i^S$ and $\tilde{E}_i^R$ are each Gaussian distributed without threshold about a mean energy $\tilde{\mu}^S$ and $\tilde{\mu}^R$, $\tilde{E}_i^S$ and $\tilde{E}_i^R$ can be higher than the activation barriers. In that case, the transition rate of translocation from site i to site $i \pm 1$ or the transition rate from the R state to the S state at site i will be overestimated. However, if the activation barrier is large compared to the standard deviations, such events should be very rare, and have little effect on the diffusion coefficient.

To obtain the relation between D and the roughness of the energy landscape in the R state, a kinetic Monte Carlo simulation was performed using the Bortz-Kalos-Lebowitz (BKL) algorithm [41]. The diffusion coefficient was obtained by taking the slope of mean square displacement vs. time.

Since the 1D diffusion coefficient for the R state is much lower than that for the S state, in other words, the S state energy roughness is much less than the R state energy roughness ($\tilde{\sigma}^S < \tilde{\sigma}^R$), we need the protein to prefer the S state if we wish a fast search. Therefore, since we are taking the mean energy of the search state to be zero, the mean energy of the recognition state (relative to the searching state) $\tilde{\mu}^R$ should be positive, especially for large $\tilde{\sigma}^R$. So, although in one-state Models I and II, we chose a zero mean energy for the Gaussian-distributed energy, for this model we vary $\tilde{\mu}^R$ from 0 to 4. Although the protein thermodynamically prefers the S state away from the target site, the two-state model assumes that at the target site the free energy in the recognition state is at its minimum, and at the target the free energy is lower in the recognition state than in the search state. The conformation transition activation barrier $\tilde{E}_a^C$ determines the frequency of state switching. Obviously, a small barrier will lead to a fast switching between these two states and thereby facilitate the sliding motion of the protein along DNA. However, a small $\tilde{E}_a^C$ will also result in a small average residence time at the target site τ_T , which will allow fast reaction only when the reaction rate constant Ω is much larger than $\omega_{T,T\pm 1}^R + \omega_T^{RS}$, where T is the target site number. Otherwise, the fast switching between R and S states will block any speed up of the targeting process, since the target site, once found, will be abandoned before a reaction can occur. Similar to the one-state analysis, as shown in the [Supporting Material](#), the reaction rate r ($x = 0$) is given by:

$$\frac{c_\infty^S}{r} = \frac{1}{2\sqrt{k_{off}D}} + \frac{1}{\omega_T^{SR} \exp(-\tilde{E}_T^S)} + \frac{1}{\Omega \exp(-\tilde{E}_T^R)} \quad (20)$$

where $\tilde{E}_T^S$, $\tilde{E}_T^R$, c_∞^S are target energy in “searching state”, target energy in “recognition state”, and far field protein concentration in “searching state”, respectively. From (20), when $\Omega < \omega_T^{SR} \exp(\tilde{E}_T^R - \tilde{E}_T^S) = \omega_T^{RS}$ and $\Omega < \sqrt{k_{off}D} \exp(\tilde{E}_T^R)$, the protein targeting is *reaction limited*, and the reaction rate is:

$$r = c_\infty^S \exp(-\tilde{E}_T^R) \Omega \quad (21)$$

Note that at the target site the free energy in the recognition state is low, while the free energy of the search state is relatively insensitive to protein location along the DNA. Thus, although away from the target site the searching-state free energy is generally lower than the recognition state energy, at the target site, we can assume that $\tilde{E}_T^S - \tilde{E}_T^R > 0$, and the reaction rate is *sped up* a factor of $\exp(-\tilde{E}_T^R)$, relative to the reaction rate for a flat free-energy landscape in the recognition site. The reason is that in the reaction-limited regime, all protein states are sampled in proportion to their equilibrium distribution before the protein is able to finally react. Since the protein in the recognition state at the target site has a low free energy, it is thermodynamically favored; that is the protein is *stable* at the target site and this increases the net reaction rate. To account for the concentration of protein at the target site in the recognition state, relative to the concentration away from the target, we need an additional factor, analogous to that given by (7) for the one-state model. Since there

are both recognition and search states, the concentration of protein in the recognition state at the target is:

$$\begin{aligned}
 c_T^R &= \frac{\exp(-\tilde{E}_T^R)}{\frac{1}{n} \sum_i^n \exp(-\tilde{E}_i^S) + \frac{1}{n} \sum_i^n \exp(-\tilde{E}_i^R)} c_\infty \\
 &= \frac{\exp(-\tilde{E}_T^R)}{\exp\left[\frac{1}{2}(\tilde{\sigma}^S)^2\right] + \exp\left[-\tilde{\mu}^R + \frac{1}{2}(\tilde{\sigma}^R)^2\right]} c_\infty \\
 &\approx \frac{\exp(-\tilde{E}_T^R)}{1 + \exp\left[-\tilde{\mu}^R + \frac{1}{2}(\tilde{\sigma}^R)^2\right]} c_\infty
 \end{aligned} \quad (22)$$

where we have again used analytical results for the expected values of the log-normal distributions given by the above averages of exponentials of $\tilde{E}_i^R$ and $\tilde{E}_i^S$, which are Gaussian-distributed quantities. Also, in the above, we remember that we have as a reference energy taken the average energy in the searching state μ^S to be zero, and since we assume that the energy roughness in the searching state is low, we have neglected the factor involving σ^S in the final equation. We have an overall reaction rate of

$$r = c_\infty \frac{\exp(-\tilde{E}_T^R) \Omega}{1 + \exp\left[-\tilde{\mu}^R + \frac{1}{2}(\tilde{\sigma}^R)^2\right]} \quad (23)$$

Note c_∞ is the far-field protein concentration in both “searching” and “recognition” states. This result shows that *there is no enhancement in reaction rate to be gained by having a low value of E_T^S* , i.e., having a local minimum in free energy at the target site in the searching state. Such a local minimum would enhance the concentration of protein there, but would correspondingly diminish the ratio of concentration in the recognition state at the target, relative to that in the searching state, and so produce no gain in reaction rate in the reaction limited regime. The reason is that in the reaction-limited regime, only the equilibrium concentration at the target in the recognition state matters. In their two-state model, Slutsky and Mirny [32] suggested that a low free energy in the searching state at the target might enhance the overall reaction rate, but the argument just given indicates that this will not be effective in the reaction-limited regime. In the diffusion-limited regime, it will of course also be ineffective, since in that case the far-field diffusivity controls the reaction rate, and not the concentration at the target, in either the recognition or searching state.

Returning to (20), when $\omega_T^{RS} \ll \Omega$ and $\omega_T^{SR} \ll \sqrt{k_{off}D} \exp(\tilde{E}_T^S)$, from (20), the targeting process is *conformational-switch limited*, and the reaction rate is:

$$r = c_\infty^S \exp(-\tilde{E}_T^S) \omega_T^{SR} \quad (24)$$

This regime is entered only if $\omega_T^{RS}/\Omega \ll 1$, but the preference of the protein for the recognition state to the searching state at the target requires that $\omega_T^{RS} < \omega_T^{SR}$, although away from the target site, as mentioned above, the protein prefers the search state. Thus, it is possible that this regime could lead to a speed-up of targeting, if $\omega_T^{SR} > \Omega \gg \omega_T^{RS}$. In this case, the protein is trapped at the target site long enough to react there, even if the rate of reaction is not fast. Hence, the overall rate of reaction might be enhanced.

Finally, when $\sqrt{k_{\text{off}}D} \ll \omega_T^{SR} \exp(-\tilde{E}_T^S)$ and $\sqrt{k_{\text{off}}D} \ll \Omega \exp(-\tilde{E}_T^R)$, the targeting process is *diffusion limited*, and the reaction rate is:

$$r = \frac{2c_\infty^S D}{\sqrt{D/k_{\text{off}}}} \quad (25)$$

This formula is the same as for the one-state model in the diffusion-limited regime, namely (6). However, the diffusion coefficient D in the above is the average diffusion coefficient away from the binding site, and this diffusion coefficient is dominated by diffusion in the search state, which is fast, since we have assumed that away from the target site, the free energy of the search state is lower than that of the recognition state; i.e., $\tilde{E}_i^S - \tilde{E}_i^R < 0$ ($i \neq T$).

Therefore, the apparent 1D diffusion coefficient in the two-state model can be approximated by:

$$D \approx D^S f^S \approx D^S (1 - f^R) \quad (26)$$

where D^S is the one-dimensional diffusion coefficient of the protein in “searching state”, and f^S and f^R are the fraction of time or the probability that protein stays in “searching” and “recognition” states, respectively.

At equilibrium, the probability of residing in the search state can be computed by $f^S = \frac{\sum_{i=1}^n \exp(-\tilde{E}_i^S)}{\sum_{i=1}^n \exp(-\tilde{E}_i^S) + \sum_{i=1}^n \exp(-\tilde{E}_i^R)} = \frac{\langle \exp(-\tilde{E}_i^S) \rangle}{\langle \exp(-\tilde{E}_i^S) \rangle + \langle \exp(-\tilde{E}_i^R) \rangle} = \frac{\exp[-\tilde{\mu}^S + \frac{1}{2}(\tilde{\sigma}^S)^2]}{\exp[-\tilde{\mu}^S + \frac{1}{2}(\tilde{\sigma}^S)^2] + \exp[-\tilde{\mu}^R + \frac{1}{2}(\tilde{\sigma}^R)^2]}$, where n is the number of sites, and both $\langle \exp(-\tilde{E}_i^S) \rangle$ and $\langle \exp(-\tilde{E}_i^R) \rangle$ are expected values of log-normal distributions. By taking $\tilde{\mu}^S = 0$, $\tilde{\sigma}^S = 0$ (very small roughness of S state), we have $f^S = \frac{1}{1 + \exp[-\tilde{\mu}^R + \frac{1}{2}(\tilde{\sigma}^R)^2]}$. Assuming the 1D diffusion in the R state contributes little to the overall 1D diffusion, the 1D diffusion coefficient can be approximated by:

$$D \approx D^S \cdot f^S = \frac{D^S}{1 + \exp[-\tilde{\mu}^R + \frac{1}{2}(\tilde{\sigma}^R)^2]} \quad (27)$$

When the activation energy for transitioning from the search to the recognition state is much lower than that for diffusing in the search state, $E_a^C \ll E_a^S$, the protein can switch its conformation frequently before translocating to the next base pair. So, the protein will usually go directly to the recognition state whenever the recognition state energy E_i^R is lower than the searching state energy E_i^S . Then introducing the searching state will have a similar effect as introducing a low threshold to the Gaussian random potential, low enough that the protein usually takes advantage of it to diffuse from one site to another. This can also be seen by examining the free-energy landscape at the bottom of Fig. 2 for the two-state model, where the searching state free energy acts like a low-energy cut-off for the free-energy landscape of a one-state model. Thus, in this limit, the diffusion is along a flat landscape with occasional deep traps, which in the two-state model are the few locations where the recognition state has lower free energy than the searching state. Note that when σ (σ^R) becomes large, (12) reduces to (14), which gives a value twice as high as given by (27), where the constant ν in (12) now is the threshold energy or average search state energy relative to the average energy of the recognition state, or $\nu = \mu^S - \mu^R = -\mu^R$.

The denominator in (27) is the same factor that slows down targeting in the reaction-limited regime discussed above; see (23). The relation between D and the roughness of the free-energy landscape in the recognition state as well as the relation between f^R and the free-energy roughness in the recognition state can also be obtained by the BKL algorithm and are compared with the theoretical result from (27) in Fig. 6a, where the average free energy in the “recognition” state is taken to be $\mu^R = 2 k_B T$.

The dependence of the fraction of time in the recognition state f^R on average energy μ^R and roughness of recognition state is plotted in Fig. 6b. This fraction is higher for larger

Fig. 6 **a** 1D diffusion coefficient and fraction of time that protein is in the “recognition” state vs. energy landscape roughness in the “recognition” state based on (27) (lines) and KMC simulation results (squares and circles). **b** Colors indicate the fraction of time that protein is in the recognition state as a function of both energy roughness and average energy of the recognition state

roughness or lower average recognition state energy. As a result, to speed up the protein sliding, we need the protein to spend a high fraction of time in the searching state and a low fraction of time in the recognition state, and so the average energy of the recognition state should be high relative to the searching state.

Both (27) and KMC simulations demonstrate that this two-state model can facilitate protein targeting process in the *diffusion limited* regime, since it can increase the overall 1D diffusion coefficient dramatically, especially when the average energy of the recognition state is high, which results in a faster reaction rate (25). However, as μ^R increases, the frequency of protein conformational switch from S to R state will decrease. In other words, for high average energy of the recognition state, the conformational switch will become the speed-limiting step.

3 Applications and discussion

For most DNA-binding proteins (DBPs), the sequence-dependent free energy (E_i) and the roughness of the 1D disordered free-energy landscape along DNA are inaccessible. However, it is noteworthy that the one-dimensional diffusion coefficients of proteins with long average sliding length ($<L_{\text{sliding}}>$), such as T7 RNA polymerase (T7 RNAP, $<L_{\text{sliding}}> > 1,000$ bp) and hOgg1 ($<L_{\text{sliding}}> \sim 200$ bp) are of the order 10^{-13} m²/s, which is close to Schurr's [42] estimate $D \approx 5 \times 10^{-13}$ m²/s of the effective 1D diffusivity of a protein following the helical path marked by the major groove of the DNA, accounting only for viscous drag of moving along this path and ignoring the random potential. Thus, for some proteins, there may be little effect of the potential on the diffusivity. Either the roughness of the potential is very small ($\sigma \leq 0.25 k_B T$, based on (1)); or there is a mechanism to reduce the effective roughness of the potential, such as an energy threshold or the searching-state mechanism discussed above. Note that for a flat potential with very small roughness ($\sigma \leq 0.25 k_B T$), the concentration or the occupancy probability of the protein along the DNA is roughly sequence-independent. In other words, the enhancement of the effective reaction rate r resulted from accumulation of DBP's on the target site will be very limited. However, by introducing the threshold, even for a rough landscape the protein can maintain its relatively high probability at the target site without slowing down its diffusion significantly.

To illustrate how our model can be used to determine 1D diffusion coefficients and targeting rates, we consider the free-energy landscape model of Barbi [34, 35]. This model uses the number of hydrogen bonds (HB) formed at each position of the protein along the DNA, to set the depth of the site-specific potential well. Thus, in Barbi's model, the well depth for a given site is simply estimated by:

$$E_i = -N_i \varepsilon \quad (28)$$

where N_i is the number of the hydrogen bonds and ε is the strength of the hydrogen bonds. We can then obtain the 1D potential landscape of the T7 RNAP-DNA system by plugging (28) into (17). Since the actual well shape of the 1D potential along DNA is unknown, we set n in (17) to 1, which guarantees that the protein feels a smooth continuous force along this potential landscape for simplicity. This method assumes that both electrostatic and hydrophobic forces have a negligible effect on the roughness. This is a crude assumption, but allows us to illustrate how to compute long-range 1D diffusivity from any model that gives a free-energy landscape for DNA-binding proteins.

Using the base sequence for T7 DNA, we thus obtain the number of hydrogen bonds (HB) for each site, where a “site” consists of the 5-bp DNA sequence that T7 RNAP binds to at any one instant when moving along the DNA. The number of hydrogen bonds for each site was obtained for a sequence of 3,000 base pairs (T7 genome, bases 17,001 to 20,000 [43]) using Barbi’s method. Barbi’s method provides a 5×4 matrix that allows the total number of hydrogen bonds to be computed for any ordered sequence of five bases. The T7 RNAP was modeled by an over-damped Brownian particle 5 nm in diameter. To obtain the 1D diffusivity of the protein, we perform a Brownian dynamics simulation using 30,000 non-interacting particles that are initially evenly distributed over the 1,000 base pairs in the

Fig. 7 **a** Normalized diffusivity D/D_0 vs. roughness σ of the continuous random potential (12) (Model III). Squares are obtained by Brownian dynamics (BD) simulations, and the line is from the MFPT. **b** Histogram of particle distribution over 50 potential wells, where the *black bars* are obtained by BD simulations, and *white bars* from Boltzmann weights with $\sigma = 2.4 k_B T$. The *inset* is the histogram of particle distribution over the central 1,000 base pairs from a set of 3,000 base pairs for seven obtained by BD simulation over with $\sigma = 2.4 k_B T$

center of the sequence of 3,000 base pairs. In Barbi's model, the number of hydrogen bonds N_i for each 5-bp site can be well fitted by a Gaussian distribution with standard deviation $\sigma_{HB} \approx 2$. (Since this can yield negative values for N_i , in our simulations, any N_i below 0 is set to 0.) Here we vary the HB strength ε up to $1.5 k_B T$, a range that includes the strength typical of a water-mediated hydrogen bond, i.e., $0.8 k_B T$ [44]. The standard deviation of the distribution of energies, E_i , is then found to be $\sigma = \sigma_{HB} \varepsilon \approx 2\varepsilon$, which varies from 0 to $3 k_B T$. The values of D for different potential roughness were obtained by taking the slopes of plots of mean square displacement versus time from the simulation and averaging over all 30,000 particles. For comparison, the diffusion coefficients over this potential were also obtained numerically using the MFPT, computed from (8).

The resulting relation between D and roughness is plotted in Fig. 7a. For roughness as high as $\sigma = 2 k_B T$, the effect of this random potential on one dimensional diffusion coefficient is less than a factor of 10, which is roughly consistent with the measurement of D of the single molecule experiments on T7 RNAP or hOgg1. If the potential well is made narrower by increasing n in (17), the disordered potential will have even less effect on D .

The final locations after 10^7 time steps simulation of the 30,000 non-interacting particles were used to generate the equilibrium number density distribution over the potential well, which is shown in Fig. 7b, and found to match that from the Boltzmann distribution, which indicates that equilibrium was achieved in this run. As a result, our proposed continuous random potential indicates that the sliding of T7 RNAP can be rapid without losing searching fidelity.

To test and improve this model, long DNA with repeated base pair sequences (e.g., T7 RNAP promoter sequences) could be used to compute the 1D diffusion coefficients of T7 RNAP on different periodic potential landscapes, and the results compared to experimental measurements of the 1D diffusion coefficient.

4 Conclusions

Using several models of protein/DNA diffusion, and multiple methods of solving those models, we have explored conditions under which proteins might find and react quickly with their target sites on DNA. A key consideration is whether the targeting is reaction-limited or diffusion-limited. In the former case, thermodynamic considerations dominate, and the stability or residence time of the target binding is paramount. The targeting rate is given by the reaction rate constant at the target site times the concentration or probability of the protein to be at this site. In this case, a deep potential well at the target site is needed so that the protein has a high thermodynamic probability to be at the target site long enough to react. If this is achieved by creating a very rough free-energy landscape, then the 1D diffusivity can be greatly slowed, possibly pushing the process towards the diffusion-limited regime. The targeting in the diffusion-limited regime is set by the concentration or probability of the protein far from the reaction site, times the average diffusivity of the protein, divided by the diffusion distance, which is the square root of the ratio of diffusivity to the dissociation constant. We showed that for diffusion distances of order 100 base pairs, using a diffusion coefficient averaged over the whole landscape gives good accuracy for the targeting rate, unless the landscape roughness exceeds around $3 k_B T$. Fast targeting in the diffusion-limited regime is enhanced by a high diffusivity, which requires an *effectively* shallow free-energy landscape. This can be achieved even if the bare landscape is quite rough, in a number of ways explored here. Firstly, the shape of the landscape matters, and if barriers are narrow

and sharply peaked, then they are effectively less rough. Secondly, if proteins can “hop”—i.e., briefly escape into solution and rapidly re-bind at a nearby site, then diffusion is much faster. Finally, if the protein can change conformation and undergo rapid diffusion in a conformation in which it does not recognize the DNA, then if the switching between searching and recognition states is fast compared to diffusion, the search efficiency is improved. Both “hopping” and introduction of a “searching” mode can under such circumstances be modeled simply by putting a maximum threshold into an otherwise Gaussian random landscape. We have shown that multiple theoretical methods, including first passage time methods, and the Festa-d’Agliano formula, allow us to compute analytically the diffusivity of such models. We have confirmed our analytical expressions using Brownian dynamics and kinetic Monte Carlo methods. Further progress in understanding this complex problem will probably require more detailed understanding of the actual free-energy landscape experienced by real proteins.

Acknowledgements We acknowledge support from NSF under grant EFRI 0938019. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

References

1. Berg, O.G., Winter, R.B., von Hippel, P.H.: Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* **20**, 6929–6948 (1981)
2. von Hippel, P.H., Berg, O.G.: Facilitated target location in biological systems. *J. Biol. Chem.* **264**, 675–678 (1989)
3. van den Broek, B., Lomholt, M.A., Kalisch, S.J., Metzler, R., Wuite, G.J.L.: How DNA coiling enhances target localization. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 15738–15742 (2008)
4. Jeltsch, A., Wenz, C., Stahl, F., Pingoud, A.: Linear diffusion of the restriction endonuclease EcoRV on DNA is essential for the in vivo function of the enzyme. *EMBO J.* **15**, 5104–5111 (1996)
5. Gowers, D.M., Wilson, G.G., Halford, S.E.: Measurement of the contributions of 1D and 3D pathways to the translocation of a protein along DNA. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15883–15888 (2005)
6. Blainey, P.C., van Oijen, A.M., Banerjee, A., Verdine, G.L., Xie, X.S.: A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5752–5757 (2006)
7. Kim, J.H., Larson, R.G.: Single-molecule analysis of 1D diffusion and transcription elongation of T7 RNA polymerase along individual stretched DNA molecules. *Nucleic Acids Res.* **35**, 3848–3858 (2007)
8. Tafvizi, A., Huang, F., Fersht, A.R., Mirny, L.A., van Oijen, A.M.: A single-molecule characterization of p53 search on DNA. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 563–568 (2011)
9. Schonhoft, J.D., Stivers, J.T.: Timing facilitated site transfer of an enzyme on DNA. *Nat. Chem. Biol.* **8**, 205–210 (2012)
10. Wang, Y., Austin, R., Cox, E.: Single molecule measurements of repressor protein 1D diffusion on DNA. *Phys. Rev. Lett.* **97**, 048302 (2006)
11. Hammar, P., Leroy, P., Mahmutovic, A.: The lac repressor displays facilitated diffusion in living cells. *Science* **336**, 1595–1598 (2012)
12. Elf, J., Li, G.-W., Xie, X.S.: Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* **316**, 1191–1194 (2007)
13. Lomholt, M.A., van den Broek, B., Kalisch, S.-M.J., Wuite, G.J.L., Metzler, R.: Facilitated diffusion with DNA coiling. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 8204–8208 (2009)
14. Brackley, C.A., Cates, M.E., Marenduzzo, D.: Facilitated diffusion on mobile DNA: configurational traps and sequence heterogeneity. *Phys. Rev. Lett.* **109**, 168103 (2012)
15. Bauer, M., Metzler, R.: In vivo facilitated diffusion model. *PLoS ONE* **8**, e53956 (2013)
16. Halford, S.E., Marko, J.F.: How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.* **32**, 3040–3052 (2004)
17. Cherstvy, A., Kolomeisky, A., Kornyshev, A.: Protein–DNA interactions: reaching and recognizing the targets. *J. Phys. Chem. B* **112**, 4741–4750 (2008)

18. Florescu, A.-M., Joyeux, M.: Description of nonspecific DNA-protein interaction and facilitated diffusion with a dynamical model. *J. Chem. Phys.* **130**, 015103 (2009)
19. Florescu, A.-M., Joyeux, M.: Comparison of kinetic and dynamical models of DNA-protein interaction and facilitated diffusion. *J. Phys. Chem. A* **114**, 9662–9672 (2010)
20. Halford, S.E.: An end to 40 years of mistakes in DNA-protein association kinetics? *Biochem. Soc. Trans.* **37**, 343–348 (2009)
21. Riggs, A., Bourgeois, S.: The lac repressor-operator interaction* 1,* 2:: III. Kinetic studies. *J. Mol. Biol.* **401**–417 (1970)
22. Bonnet, I., Biebricher, A., Porté, P.-L., Loverdo, C., Bénichou, O., Voituriez, R., Escudé, C., Wende, W., Pingoud, A., Desbiolles, P.: Sliding and jumping of single EcoRV restriction enzymes on non-cognate DNA. *Nucleic Acids Res.* **36**, 4118–4127 (2008)
23. Berg, O.G., von Hippel, P.H.: Selection of DNA binding sites by regulatory proteins 1: statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–743 (1987)
24. Fields, D.S., He, Y., Al-Uzri, A.Y., Stormo, G.D.: Quantitative specificity of the Mnt repressor. *J. Mol. Biol.* **271**, 178–194 (1997)
25. Slutsky, M., Mirny, L.A.: Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophys. J.* **87**, 4021–4035 (2004)
26. Denisov, S., Horsthemke, W.: Mean first-passage time for an overdamped particle in a disordered force field. *Phys. Rev. E* **62**, 3311–3321 (2000)
27. Sheinman, M., Bénichou, O., Voituriez, R., Kafri, Y.: First-passage time distribution for a random walker on a random forcing energy landscape. *J. Stat. Mech. Theory Exp.* **2010**, P09005 (2010)
28. Hirth, J.P., Lothe, J.: *Theory of Dislocations*. McGraw-Hill, New York (1968)
29. Mathé, J., Visram, H., Viasnoff, V., Rabin, Y., Meller, A.: Nanopore unzipping of individual DNA hairpin molecules. *Biophys. J.* **87**, 3205–3212 (2004)
30. Jun, B., Weaver, D.L.: One-dimensional potential barrier model of protein folding with intermediates. *J. Chem. Phys.* **116**, 418–426 (2002)
31. Zwanzig, R.: Diffusion in a rough potential. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2029–2030 (1988)
32. Mirny, L., Slutsky, M., Wunderlich, Z., Tafvizi, A., Leith, J., Kosmrlj, A.: How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J. Phys. A: Math. Theor.* **42**, 434013 (2009)
33. Minoura, I., Katayama, E., Sekimoto, K., Muto, E.: One-dimensional Brownian motion of charged nanoparticles along microtubules: a model system for weak binding interactions. *Biophys. J.* **98**, 1589–1597 (2010)
34. Barbi, M., Place, C., Popkov, V., Salerno, M.: A model of sequence-dependent protein diffusion along DNA. *J. Biol. Phys.* **30**, 203–226 (2004)
35. Barbi, M., Place, C., Popkov, V., Salerno, M.: Base-sequence-dependent sliding of proteins on DNA. *Phys. Rev. E* **70**, 1–6 (2004)
36. Bénichou, O., Kafri, Y., Sheinman, M., Voituriez, R.: Searching fast for a target on DNA without falling to traps. *Phys. Rev. Lett.* **103**, 138102 (2009)
37. Hu, L., Grosberg, A.Y., Bruinsma, R.: Are DNA transcription factor proteins Maxwellian demons? *Biophys. J.* **95**, 1151–1156 (2008)
38. Gorokhov, D.: Diffusion and creep of a particle in a random potential. *Phys. Rev. B* **58**, 213–217 (1998)
39. Festa, R., D'Agliano, E.G.: Diffusion coefficient for a Brownian particle in a periodic field of force: I. Large friction limit. *Physica A: Stat. Mech. Appl.* **90**, 229–244 (1978)
40. Kalodimos, C.G., Biris, N., Bonvin, A.M.J.J., Levandoski, M.M., Guennegues, M., Boelens, R., Kaptein, R.: Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science*. **305**, 386–389 (2004)
41. Bortz, A., Kalos, M.: A new algorithm for Monte Carlo simulation of Ising spin systems. *J. Comp. Phys.* **18**, 10–18 (1975)
42. Schurr, J.: The one-dimensional diffusion coefficient of proteins absorbed on DNA hydrodynamic considerations. *Biophys. Chemist.* **9**, 413–414 (1975)
43. Dunn, J.J., Studier, F.W.: Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol.* **166**, 477–535 (1983)
44. Deng, Y., Glimm, J., Wang, Y.: Prediction of protein binding to DNA in the presence of water-mediated hydrogen bonds. *J. Mol. Model.* **5**, 125–133 (1999)