Dynamic Strategies for Target-Site Search by DNA-Binding Proteins

Mario A. Díaz de la Rosa,† Elena F. Koslover,‡ Peter J. Mulligan,† and Andrew J. Spakowitz†‡*
†Department of Chemical Engineering and ‡Biophysics Program, Stanford University, Stanford, California

ABSTRACT Gene regulatory proteins find their target sites on DNA remarkably quickly; the experimental binding rate for lac repressor is orders-of-magnitude higher than predicted by free diffusion alone. It has been proposed that nonspecific binding aids the search by allowing proteins to slide and hop along DNA. We develop a reaction-diffusion theory of protein translocation that accounts for transport both on and off the strand and incorporates the physical conformation of DNA. For linear DNA modeled as a wormlike chain, the distribution of hops available to a protein exhibits long, power-law tails that make the long-time displacement along the strand superdiffusive. Our analysis predicts effective superdiffusion coefficients for given nonspecific binding and unbinding rate parameters. Translocation rate exhibits a maximum at intermediate values of the binding rate constant, while search efficiency is optimized at larger binding rate constant values. Thus, our theory predicts a region of values of the nonspecific binding and unbinding rate parameters that balance the protein translocation rate and the efficiency of the search. Published data for several proteins falls within this predicted region of parameter values.

INTRODUCTION

Nature controls biochemical processes at the cellular level with remarkable speed and fidelity. A critical aspect of biochemical control is transcriptional regulation, as virtually all cellular activity is dependent on the timely expression of the information encoded in the cell’s DNA. Gene expression is mediated through the direct interaction between transcription factors and their respective target sites. Surprisingly, the cell achieves this regulation with a relatively small number of transcription factors searching massive lengths of DNA. Escherichia coli, for instance, has ~20 copies of lac repressor that must search its 4,600,000 basepair-long genome (1).

To account for the fast localization rate of lac repressor to its operator, Berg et al. (2) postulated a search model termed facilitated diffusion, in which a protein translocates along DNA through a multistep process of free diffusion and nonspecific binding. Four proposed facilitated-diffusion mechanisms are shown in Fig. 1. Mechanism 1 involves the protein sliding along DNA. A sufficient thermal fluctuation detaches the protein, leading to free diffusion and eventual reattachment, either to a nearby segment (mechanism 2, microhop) or to a distal segment (机制ism 3, macrohop). If a loop exists in the DNA, the protein may be able to traverse along the strand without attaching (机制ism 4, intersegmental transfer). Although other physical effects (such as long-range electrostatic attraction) are proposed to contribute to the rapid binding of lac repressor to its operator (3), it is widely accepted that facilitated diffusion plays an important role in the search process.

Experiments provide evidence for the mechanisms described above and substantiate the validity of the facilitated diffusion model. The sliding mechanism is observed using single-molecule manipulation and fluorescence (4) and total internal reflection fluorescence microscopy (5–12). Hopping events are reported for stretched DNA strands, using single-molecule imaging (11). A range of experiments (4,13–17) provide indirect evidence for the existence of the hopping and intersegmental transfer mechanisms. Notably, DNA conformation is very clearly demonstrated to play a critical role in target-site localization rates for DNA strands that are either supercoiled (15) or stretched using optical tweezers (17).

Modeling efforts of Berg and co-workers (2,18,19) represent the groundwork for a kinetic theory of facilitated diffusion. Microscopic models that capture individual sliding trajectories have addressed effects such as multiple sliding proteins (20,21), irreversible detachment (22), the presence of obstacles along DNA (23), and sequence-dependent translocation (24,25). Various theoretical approaches address the role of DNA conformation. Brownian dynamics simulations are adapted to study protein translocation along linear DNA (26) and association to circular DNA (27). Hop distributions are found from analyses of frozen lattice configurations to assess local and global translocation behaviors (28). Extensions to the kinetic model of Berg et al. (2) incorporate conformation-dependent hops (29). The distribution of hops along stretched DNA is predicted using a reaction-diffusion analysis (30). Scaling arguments for site localization identify regimes for straight, coiled, and globular configurations (31). Theoretical models of translocation including long-range hops approximate the density of foreign segments to be uniform in space, predicting the role of local and distal translocation events (17,32). Effective one-dimensional transport is modeled using fractional Brownian dynamics with anomalous diffusivity to capture the impact of long-range transfer events along the DNA (33).

In this article, we present a theoretical model of protein translocation on DNA that explores the influence of nonspecific binding and DNA conformation on the search. We find...
that the spatial layout of DNA segments plays an integral role in protein translocation rate and efficiency. Our model is sufficiently detailed to address a wide range of biologically relevant issues while still maintaining computational tractability for studying processes at genome length-scales and biological timescales.

THEORETICAL MODEL

We develop a theoretical model of the translocation of a protein along DNA via facilitated diffusion, incorporating the physical behavior of the polymer. In Fig. 2, we show a schematic of our model. In this section, we proceed to translate this schematic representation into a mathematical formalism, which we analyze in Asymptotic Analysis of Theoretical Model. These results also form the basis of a novel dynamic Monte Carlo (MC) formalism used to track individual protein trajectories, as explained in Appendix C in the Supporting Material.

Transport and kinetics in the on-state

In this work, we assume the protein slides along the DNA in a one-dimensional diffusive process that is not biased by the underlying DNA sequence or varying binding states of the protein. Furthermore, we assume the DNA conformation does not affect the on-off kinetics—thus, we ignore transient events of DNA bending or looping that could potentially dislodge the protein from the DNA surface. The chain is assumed to be sufficiently long such that the anomalous transport near the ends of the chain can be ignored and the chain is effectively infinite. We also ignore the presence of other proteins on the chain. Our current goal in this article is to lay down the overarching physical phenomena before turning to these important biological effects in future work.

Given our assumptions, the transport of the protein along the DNA is governed by the distribution $G_{\text{on}}(s|s_0; t)$, which gives the joint probability that a protein beginning at $s_0$ at time zero will end at $s$ at time $t$ through one-dimensional diffusive transport. The left image in Fig. 2 represents the diffusive transport for $|s - s_0| = L$ over the observed time $t$. This mathematical function is written as

$$G_{\text{on}}(s|s_0; t) = \frac{1}{\sqrt{4\pi D_{\text{on}} t}} \exp \left[ -\frac{(s - s_0)^2}{4D_{\text{on}} t} \right], \quad (1)$$

where $D_{\text{on}}$ is the effective one-dimensional diffusion coefficient for protein sliding along the chain. The units of $G_{\text{on}}$ are $1/\text{length}$ and the function is normalized.

The transition kinetics for going from the on- to the off-state are modeled as a first-order reaction process. We define $P_{\text{on}}(t)$ as the probability that a protein that transitions from the off- to the on-state at time zero will remain in the on-state at time $t$. We note that $P_{\text{on}}(t)$ is not the probability that the protein is bound to the DNA at any given time, but rather it represents the duration of the on-leg of a single cycle. This on-probability satisfies the differential equation

$$\frac{dP_{\text{on}}}{dt} = -k_u P_{\text{on}}, \quad (2)$$

with $P_{\text{on}}(t=0) = 1$ and unbinding rate $k_u$, resulting in

$$P_{\text{on}}(t) = \exp(-k_u t). \quad (3)$$

Transport and kinetics in the off-state

The off-state transport encompasses both micro- and macrohops and must incorporate both the dynamics of the protein diffusion and the configurational properties of the DNA strand. We assume the three-dimensional transport of the protein proceeds with an effective diffusion coefficient $D_{\text{off}}$. This approximates the motion of the protein through the crowded environment within the heart of the DNA as one within an effective medium. We describe the physical behavior of DNA by an equilibrium distribution function, which neglects conformational memory in the DNA as the protein traverses along the strand. Preliminary simulations of the target-site search process show no qualitative difference between equilibrated and frozen DNA limits. This
approximation of rapid DNA equilibration will be further explored in a future publication.

We define three-dimensional spatial distribution functions for the DNA ($G_D$) and the protein ($G_p$). The function $G_D(\vec{R}; |s-s_0|)$ gives the joint probability of having the DNA segment $s_0$ located at the origin and segment $s$ at position $\vec{R}$. The distribution $G_D$ is specific to the DNA properties; we explore several physical models for DNA in this work. The function $G_p(\vec{R}; t)$ describes the joint probability for a protein beginning at the origin at time zero and ending at spatial position $\vec{R}$ at time $t$.

Ultimately, the protein’s transport and kinetics are closely tied to each other. The transport gives the probability of locating the protein within a distance $a$ of the DNA stand, which in turn affects the rate of protein binding. We begin our framework by defining the spatially dependent rate of binding to any segment of DNA. The binding rate scales with the local probability of simultaneously having both a DNA segment and a protein within a spatial reaction volume $\nu_0 = 4\pi a^3/3$, where $a$ is the reaction radius shown in Fig. 2. Incorporating a reaction zone into the theory is a necessary step to avoid an unphysical divergence in the reaction-diffusion formulation. This is an established approach in addressing diffusion-controlled reactions (34).

The local binding rate at position $\vec{R}$ is written as

$$k_b S(\vec{R}) G_p(\vec{R}; t) = k_b \int d\vec{R}_1 \int_0^\infty ds \ H(a - |\vec{R}_1 - \vec{R}|) \times G_D(\vec{R}_1; |s|) G_p(\vec{R}_1; t),$$  

(4)

where $H(x)$ is the Heaviside step function [$H(x)$ is unity for $x \geq 0$, and zero otherwise]. In Eq. 4, we define $S(\vec{R})$ as the length of DNA as a function of position $\vec{R}$. The first-order rate constant $k_b$ gives the rate per unit length of a protein binding to a DNA strand if the protein is within the reaction radius $a$. Generally, the relationship between $k_a$ and $k_b$ is established by an equilibrium experiment, resulting in an equilibrium constant $K_{eq} = k_{on}/k_{off} = 2.17,35$, where $k_{off} = k_0/v_0$. A separate experiment is necessary to determine either $k_a$ or $k_{on}$ to uniquely identify these parameters (see Appendix E in the Supporting Material for more details).

The protein distribution $G_p$ satisfies the reaction-diffusion equation

$$\left( \frac{\partial}{\partial t} - D_{off} \nabla^2 \right) G_p(\vec{R}; t) = -k_b S(\vec{R}) G_p(\vec{R}; t)$$  

(5)

with the initial condition

$$G_p(\vec{R}; t=0) = \delta(\vec{R}).$$

Without loss of generality, we write the function

$$G_p(\vec{R}; t) = P_{off}(t) \Gamma_p(\vec{R}; t),$$

with $P_{off}(t)$ giving the probability the protein remains in the off-state at time $t$ (and given that it unbinds at time zero) and $\Gamma_p(\vec{R}; t)$ describing the normalized spatial distribution for a protein remaining in the off-state at time $t$. Thus,

$$\int d\vec{R} \Gamma_p(\vec{R}; t) = 1$$

for all time $t$. The off-probability $P_{off}$ satisfies the differential equation

$$\frac{dP_{off}}{dt} = -k_b M(t) P_{off},$$  

(6)

with $P_{off}(t=0) = 1$, where

$$M(t) = \int d\vec{R} S(\vec{R}) \Gamma_p(\vec{R}; t)$$

is the total length of DNA within $a$ of the protein at time $t$ if the protein remains in the off-state at $t$. The solution of the governing differential equation is given by

$$P_{off}(t) = \exp \left[ -k_b \int_0^t M(t') dt' \right],$$  

(7)

a formal result that requires solution of Eq. 5 to evaluate.

We define the distribution $G_{off}(s|s_0; t)$ that gives the probability that, for a binding event at time $t$, the protein reenters the DNA at segment $s$, given that it left segment $s_0$ at time zero. The off-state probability distribution $G_{off}(s|s_0; t)$ scales with the reactive flux onto position $s$; thus, we write

$$G_{off}(s|s_0; t) = \frac{1}{M(t)} \int d\vec{R}_1 \int d\vec{R}_2 \ H(a - |\vec{R}_1 - \vec{R}_2|) \times G_D(\vec{R}_1; |s-s_0|) \Gamma_p(\vec{R}_2; t),$$  

(8)

which is normalized. We leave our theoretical development with a general $G_D$ until we explore several explicit polymer models for DNA later in this article.

Equations 7 and 8 fully characterize the transport and kinetics in the off-state of our theoretical model. However, evaluation of these functions requires a solution for $\Gamma_p$ using the reaction-diffusion equation for $G_p = P_{off} \Gamma_p$ (Eq. 5). A suitable approximation for $\Gamma_p$ is to assume the distribution remains relatively unperturbed from the free solution

$$\Gamma_p^{(free)}(\vec{R}; t) = \left( \frac{1}{4\pi D_{off} t} \right)^{3/2} \exp \left( -\frac{\vec{R}^2}{4D_{off} t} \right),$$  

(9)

which satisfies Eq. 5 in the limit $k_b \rightarrow 0$. In this work, we will make use of this approximation along with numerical simulations that effectively solve for $\Gamma_p$.

### ASYMPTOTIC ANALYSIS OF THEORETICAL MODEL

The transport and kinetic processes outlined in the previous section act as inputs into a two-state model that captures the composite behavior. In Appendix A in the Supporting Material, we formally define and derive the composite Green’s
function $G(s|s_0,t)$ that governs the two-state transport. The solution for $G(s|s_0,t)$ requires a sum over all possible transport and kinetic events that bring the particle from $s_0$ to $s$ in time $t$, accounting for the statistical distributions previously discussed.

The exact result for the composite Green’s function (Eq. A5 in the Supporting Material) is the fundamental tool we use to study transport dynamics along a specific model of the DNA configuration, specified by $G_D$. We derive the long-time asymptotic behavior of the individual transport functions appearing in Eq. A5 for the wormlike chain model in Appendix B in the Supporting Material. For ease of notation, we adopt a nondimensionalization of our parameters such that all lengths are made dimensionless by the Kuhn length $b$ and all times are made dimensionless by the on-state diffusion time $t_{\text{diff}} = b^2/D_{\text{on}}$. This redefines our model in terms of the dimensionless parameters $\gamma = D_{\text{off}}/D_{\text{on}}$, $\alpha = a/b$, $\kappa_a = t_{\text{diff}} k_a$, and $\kappa_b = t_{\text{diff}} b k_b$.

Inserting the results from Appendix B into Eq. A5 in the Supporting Material, we arrive at the asymptotic form for the composite Green’s function,

$$\hat{G}(\omega;\nu) = \frac{1}{\nu + \sigma |\omega|^{1/2} + D_{\text{eff}} |\omega|^2},$$

where $\omega$ is the Fourier conjugate of $s$ and $\nu$ is the Laplace conjugate of $t$. The effective transport coefficients $\sigma$ and $D_{\text{eff}}$ are given by

$$\sigma = \frac{t_{\text{off}}^{1/2}}{\tau_{\text{on}} + \tau_{\text{off}}},$$

$$D_{\text{eff}} = \frac{t_{\text{on}}^2}{\tau_{\text{on}} + \tau_{\text{off}}},$$

In the coefficients above, we define the average times (made dimensionless by $t_{\text{diff}}$) spent in the on- and off-states per cycle as

$$\tau_{\text{on}} = \frac{1}{\kappa_a},$$

$$\tau_{\text{off}} = \int_0^\infty d\tau \exp \left[-\frac{\kappa_b}{\gamma} \int_0^{\gamma \tau} dr M(r')\right],$$

and the average translocation lengths (made dimensionless by $b$) in the on- and off-states per cycle as

$$l_{\text{on}} = \sqrt{\frac{1}{\kappa_a}},$$

$$l_{\text{off}} = \left(4\sqrt{3} \alpha \lambda \kappa_b \tau_{\text{off}}\right)^2.$$

Although the leading $\omega$ scaling is independent of the specific DNA model, the precise value of $\sigma$ and $D_{\text{eff}}$ are affected by the model details through the value of $\tau_{\text{off}}$.

The asymptotic form of $\hat{G}$ found in Eq. 10 suggests that the composite transport behavior obeys a Lévy distribution (36) that is superdiffusive in its trajectory. The general functional form of a Lévy distribution is governed by the Fourier-transformed effective Green’s function (36) $\hat{G}(\omega,\tau) = \exp(-\sigma |\omega|^{\beta})$, where $\mu < 2$. In our case, Eq. 10 reduces to this distribution with $\mu = 1/2$, where the diffusive term (order $\omega^2$) is removed due to the long-time behavior being dominated by the superdiffusive transport. The general form leads to transport dynamics with an average displacement that scales as

$$\left< |s-s_0|^\beta \right>^{1/\beta} \sim \tau^{1/\beta},$$

where $\beta$ must be within the range $0 < \beta < \mu$ due to a divergence in averages with $\beta \geq \mu$ (36). For our problem, the protein transport is predicted to have an average (nondimensionalized) displacement given by

$$\left< |s-s_0|^\beta \right>^{1/\beta} = \left[4 \frac{\Gamma(1+\beta)\Gamma(-2\beta)}{\Gamma(1+\beta/2)\Gamma(-\beta/2)} \bar{\sigma}^\beta \right]^{1/\beta},$$

where the parameter $\beta$ must be within the range $0 < \beta < 1$. Previous theoretical treatments (33) also suggest that DNA conformation leads directly to superdiffusive motion, where the superdiffusive scaling is found by the polymer looping probability and the superdiffusion coefficient $\sigma$ is a phenomenological parameter. Our work provides a method to determine this coefficient based on the specific polymer and protein properties.

**DISCUSSION**

Our reaction-diffusion model of protein translocation permits a detailed examination of the effect of DNA conformation on the target-site search of gene regulatory proteins. The goal in this section is to present an analysis of our theory and discuss the biological impact of these results. Throughout this section, we use dimensionless variables to illustrate fundamental features of our model; however, parameter values that are relevant to specific experiments are identified.

The transport in the off-state (corresponds to hops that the protein takes while disengaged from the DNA. The superdiffusive transport predicted by our model emerges as a direct result of the long-range hops associated with off-state transport. Thus, the nature and likelihood of these hops are integral to the overall search dynamics.

The hop distribution is contained in $G_{\text{off}}(s|s_0, t)$, the probability of a protein leaving the strand at location $s_0$ and subsequently traveling to location $s$ at time $t$. This reentry time hinges on the binding rate $\kappa_a$: large values favor short reentry times while small values yield long reentry times. The top plot within Fig. 3 shows a surface plot of the hop probability $M(r)G_{\text{off}}(N_0; \gamma t \tau)$ (nonnormalized for clarity) versus the hop length $N = L/b$ and time $\tau \gamma t = D_{\text{off}}/b^2$. We consider the DNA Kuhn length $b = 106$ nm (37,38), $a = 10.6$ nm, and $D_{\text{eff}} \approx 4 \times 10^7$ nm$^2$/s for LacI (7),

Biophysical Journal 98(12) 2943–2953
at short times (\( \gamma \tau = D_{\text{off}}/b^2 \)) to long time (\( \gamma \tau \sim 10^3 \)) for a wormlike chain as indicated within the inset of Fig. 3. As the large-\( N \) limit is identical to the behavior of a Gaussian chain, we can exploit the results from Appendix A in the Supporting Material to address this asymptotic behavior. Analysis of \( M(\tau)G_{\text{off}}(N;\tau) \) for the Gaussian-chain model reveals a limiting form of

\[
M^{(GC)}(\tau)G^{(GC)}_{\text{off}}(N;\tau) \sim \frac{1}{6\sqrt{\pi} (N/6 + \tau)^{3/2}} \sim N^{-3/2} \quad (19)
\]

in the limit of large \( N \). This behavior is exhibited in all five time curves in Fig. 3 B in the large-\( N \) limit.

The inset in Fig. 3B shows the \( \gamma \tau = 10^{-3} (t = 0.27 \mu s) \) curve plotted with the hop distribution for transport along a rigid-rod polymer (dashed curve) and a Gaussian-chain polymer (dotted curve). Analytical results for the rigid-rod model and the Gaussian-chain model are provided in Appendix D in the Supporting Material. The wormlike chain solution exhibits close agreement with the rigid-rod behavior for \( N < 0.1 \) and with the Gaussian-chain model for \( N > 10 \). These limiting forms are expected from the physical behavior of a semiflexible polymer. At short times, the hops available to a protein translocating along a semiflexible polymer like DNA are thus a combination of those available on a rigid rod (microhops) and on a Gaussian chain (macrohops).

At long times, the distinction between microhops and macrohops disappears. The more time a protein is allowed to diffuse, the more it will have access to distal segments of the DNA strand, eliminating the ability to resolve small-scale structural details. Thus, if the protein is to distinguish between microhops and macrohops, three-dimensional excursions must be relatively short, approximately a microsecond.

In the large-\( N \) limit, the hop distribution decays as \( N^{-3/2} \), as indicated within the inset of Fig. 3 B. As the large-\( N \) limit is identical to the behavior of a Gaussian chain, we can exploit the results from Appendix A in the Supporting Material to address this asymptotic behavior. Analysis of \( M(\tau)G_{\text{off}}(N;\tau) \) for the Gaussian-chain model reveals a limiting form of

\[
M^{(GC)}(\tau)G^{(GC)}_{\text{off}}(N;\tau) \sim \frac{1}{6\sqrt{\pi} (N/6 + \tau)^{3/2}} \sim N^{-3/2} \quad (19)
\]

giving a range of times in Fig. 3 of \( t = 0.27 \mu s (\gamma \tau = 10^{-3}) \) to \( t = 2.7 \) ms (\( \gamma \tau = 10^3 \)). The data found in Fig. 3 exploit the free-diffusion approximation \( \Gamma_{\mu} = \Gamma_{\mu}^{(\text{free})} \) for transport in the off-state. The applicability of this approximation is established below.

The details of the time sensitivity of the hop distribution are illustrated in the five time slices shown in the bottom plot in Fig. 3. At short times (\( \gamma \tau = 10^{-3} \) or \( t = 0.27 \mu s \), blue curve), a bimodal hop distribution emerges that is reminiscent of the microhop and macrohop mechanisms proposed by Berg et al. (2) (see Introduction and Fig. 1). The short hop-length peak (microhop) is a result of the large probability of the protein seeing segments that are very close to the exit point due to the rigidity of the DNA strand at short lengths. The long hop-length peak (macrohop) occurs as a result of the DNA forming looped conformations that place these distal segments within close spatial proximity of the exit point.

DNA looping plays an important role in gene regulation and has been a critical determinant of DNA elasticity through cyclization-rate experiments (39,40). Such experiments give the \( J \)-factor, the effective concentration of one end of a DNA chain near the other end. The \( J \)-factor has a maximum value at \(-670 \) bp or \( N = 2 \) (39–41), which is very close to the location of the macrohop peak in Fig. 3. At very short times (\( \gamma \tau << 1 \)), the protein has not diffused far enough to leave the spatial location of the exit point, and the hop probability scales with the local concentration of chain segments near the exit point. Within our current model, long-range hops at these short times embody an intersegmental hop (32) mechanism, requiring complete dissociation of the protein from the DNA, unlike the intersegmental transfer illustrated in Fig. 1.

**FIGURE 3** (A) Surface plot (logarithmic color scale) from low in blue to high in red of the hop distribution \( M(\tau)G_{\text{off}}(s,N;\tau) \) for a wormlike chain as a function of hop length \( N = L/b \) and time \( \gamma \tau = D_{\text{off}}/b^2 \). (B) Five time slices from the surface plot, showing the evolution of the hop distribution from short time \( (\gamma \tau = 10^{-3} \) or \( t = 0.27 \) \( \mu s \) for DNA (37,38) and LacI (7) parameters, blue) to long time \( (\gamma \tau = 10^3 \) or \( t = 2.7 \) ms, red). (Inset) The blue curve \( (\gamma \tau = 10^{-3}) \), plotted with the hop distributions for a rigid rod (dashed curve) and a Gaussian chain (dotted curve).
Our analysis in the previous section reveals the overall transport to be superdiffusive with an average displacement that scales as
\[
\langle |s - s_0|^\beta \rangle^{1/\beta} \sim \tau^2,
\] (20)
in the long-time limit. The fundamental cause of the superdiffusive behavior lies in the power-law tails within the hop distribution. Analysis of a general Lévy flight \((36)\) reveals that having a hop distribution with the power-law tail \(G_{\text{off}} \sim N^{-1-\mu}\) results in the transport scaling given by Eq. 17. Our case corresponds to \(\mu = 1/2\), and the superdiffusive behavior is consistent with the hop distribution presented in Fig. 3.

The average reentry time \(\tau_{\text{off}}\) (Eq. 14) is determined by a combination of both the length of DNA that is accessible for binding \([M(\tau)]\) and the binding rate \(k_b\). A plot of \(\tau_{\text{off}}\) for a wormlike chain as a function of the dimensionless rate constant \(k_b\) is given in Fig. 4. As in Fig. 3, we use the free-diffusion approximation \(\Gamma_p \approx \Gamma_p^{(\text{free})}\) for transport in the off-state. In Fig. 4, we fix \(\gamma = 4286\) to model the behavior of representative proteins \(\text{LacI}\) and \(\text{EcoRV}\) with \(D_{\text{off}} \approx 4 \times 10^7 \text{nm}^2/\text{s}\) and \(D_{\text{on}} \approx 9 \times 10^3 \text{nm}^2/\text{s}\) \((7, 11)\). The other lines represent \(\tau_{\text{off}}\) for a Gaussian chain in the long-time limit (red dashed) and rigid rod in the short-time limit (blue dash-dotted), which are discussed below.

For high values of \(k_b\), the likelihood of rebinding is so high that the protein is effectively limited to very fast hops and short-range capture, as reflected in the hop distribution at short times (blue dashed curve in Fig. 3). For short hops, the DNA behavior can be approximated as a rigid rod. Taking the short-time limit of the rigid-rod off-state accessible length (see Appendix D in the Supporting Material) gives \(M^{(\text{RR})} \approx 2\alpha\). Assuming that the dominant contribution to \(\tau_{\text{off}}\) comes from the short-time limit of \(M^{(\text{RR})}\), we have
\[
\tau_{\text{off}}^{(\text{RR})} \approx \frac{1}{2\alpha k_b}.\] (21)

This approximate form is slightly modified for the wormlike-chain model by accounting for the additional accessible length that arises from looping distal segments into the exit point. This slight correction is included in Fig. 4, although the quantitative difference is not substantial (7-6% additional length from looping). This large-\(k_b\) behavior is plotted as the blue dashed-dotted curve in Fig. 4.

For small values of \(k_b\), the rebinding rate is sufficiently slow that the protein diffuses far enough from the exit point for the polymer to appear as an effective Gaussian chain. Taking the long-time limit of the Gaussian-chain off-state accessible length (see Appendix D in the Supporting Material) gives \(M^{(\text{GC})} \approx 4\alpha^3/\sqrt{\pi}\). Assuming the main contribution to \(\tau_{\text{off}}\) comes from the long-time limit of \(M^{(\text{GC})}\), we find
\[
\tau_{\text{off}}^{(\text{GC})} \approx \frac{\pi \gamma}{32 \alpha k_b}.\] (22)

This gives the small-\(k_b\) behavior plotted as the red dashed curve in Fig. 4. Under such conditions, the transport rate depends on the reaction radius through the combination of parameters \(\alpha^3 k_b\), and our results are insensitive to our choice of \(\alpha = 10.6\text{ nm}\) in this limit.

Fig. 4 exploits the approximation \(\Gamma_p \approx \Gamma_p^{(\text{free})}\), which essentially states that the binding reaction does not substantially perturb the off-state diffusion away from the free-diffusion Gaussian distribution. We performed Brownian dynamics simulations to numerically capture the reaction-diffusion behavior governed by Eq. 5. Comparing these results with those found in Fig. 4 demonstrates that the approximation \(\Gamma_p \approx \Gamma_p^{(\text{free})}\) has a maximum error for \(\tau_{\text{off}}\) of \(-35\%\) in the small-\(k_b\) limit; the error asymptotes to zero as \(k_b\) transitions from long-range to short-range capture. We do not include this data in Fig. 4 as the curves are indistinguishable over the range of values within the figure. We make use of the approximation \(\Gamma_p \approx \Gamma_p^{(\text{free})}\) throughout this work.

The off-state dynamic behavior outlined in Figs. 3 and 4 leads to dramatic consequences for the translocation of proteins on DNA. To explore protein translocation, we use dynamic MC simulations (see Appendix C in the Supporting Material) to generate individual protein trajectories. Fig. 5 shows the average displacement \(\langle |s - s_0|^\beta \rangle^{1/\beta} \) (with \(\beta = 0.2\)) for 10,000 total trajectories. All three simulations have \(\gamma = 4286\), \(\alpha = 0.1\), and \(\kappa_s = 39.2\). The binding rate \(k_b\) is given by \(k_b = 10^3\) (blue curve marked \(A\)), \(k_b = 10^3\) (purple curve marked \(B\)), and \(k_b = 10^6\) (red curve marked \(C\)). Curve \(B\) corresponds to rate parameters for \(\text{EcoRV}\) in a salt concentration of approximately 0 mM \((17)\). The parameters for
curves A and C span the range of values of \( \kappa_b \) for several DNA binding proteins (see Appendix E in the Supporting Material).

The mean protein displacement has a diffusive regime at short times, characteristic of the sliding and microhop mechanisms that dominate at these times. This short-time diffusive behavior is represented by the black dotted curve with scaling \( t^{1/2} \). At longer times, the protein is able to occasionally take very long steps, accelerating its translocation along the DNA. This ultimately leads to a long-time superdiffusive scaling of \( t^2 \), consistent with our predictions. The superdiffusive timescaling exists for all values of \( \kappa_b \); however, the superdiffusion coefficient \( \sigma \) exhibits a nonmonotonic behavior with a maximum at some intermediate \( \kappa_b \) value. Assuming a diffusion-controlled enzymatic reaction, the appearance of a localization rate dependence on salt concentration, including a maximum effect, is consistent with in vitro experiments (17,42).

We mark in Fig. 5 the displacement that approximately matches the length of the human genome. This plot shows that the superdiffusive motion accelerates the search such that the protein translocates over a genome length of DNA between 1 s and 100 s for these parameters. We note that in vivo measurements of lac repressor on the E. coli genome result in a reported time for initial binding to the target site of 59 s (8), which is comparable to our range of predicted translocation times. In comparison, a purely diffusive sliding search would require \( \sim 2 \times 10^{14} \) s or 6,000,000 years to traverse the entire genome.

The dashed curves in Fig. 5 provide our long-time asymptotic predictions for protein translocation. The agreement between protein-trajectory simulations and the asymptotic transport solution suggests that the composite transport function \( G \) is the appropriate tool to address the target-site search strategy in the long-time limit. We now proceed to explore the influence of the binding kinetics on the rate of translocation and the overall efficiency of the search. The metrics of interest include the superdiffusion coefficient \( \sigma \) and the average times and lengths the protein spends in each transport state per cycle, all defined in the previous section.

The power-law tails of the hop probabilities lead to long-time superdiffusive displacement with a transport coefficient \( \sigma = \lambda_{\text{off}}^{1/2}/(\tau_{\text{on}} + \tau_{\text{off}}) \). The resulting transport scaling (Eq. 18) is universal for the class of polymers we are discussing, while \( \gamma \) varies with the kinetic and transport parameters of the protein (\( \kappa_{\text{on}}, \kappa_{\text{off}}, \) and \( \gamma \)). This dependence is seen in the surface shown in Fig. 6, which plots \( \sigma \) versus unbinding rate \( \kappa_{\text{on}} \) and binding rate \( \kappa_{\text{off}} \) with \( \gamma = 4286 \). The dotted line shows the transition from short-range to long-range capture illustrated in Fig. 4. The dots A–C indicate rate parameters that correspond to the simulation curves in Fig. 5.

The bottom plot of Fig. 6 shows four slices in \( \kappa_{\text{off}} \) that exhibit a maximum search rate with nonspecific binding affinity. From our discussion of the hop distribution (Fig. 3), the binding affinity controls the distance traveled and time spent in the off-state, which in turn controls translocation acceleration through \( \sigma \). At high \( \kappa_{\text{off}} \), restriction to short-range capture yields a constant \( \sigma \). The superdiffusion coefficient does not begin to increase until the binding rate is lowered enough that the time the protein spends in three-dimensional excursions allows it to see distal DNA segments (in the long-range capture domain of Fig. 4). The subsequent peak in \( \sigma \) indicates that, past a certain point, rebinding is so infrequent that facilitated diffusion no longer aids the search.

The dashed curves in Fig. 6 are the Gaussian chain results, which show that DNA bending rigidity ultimately lowers the potential acceleration for high values of \( \kappa_{\text{off}} \). This also demonstrates that the short-time hopping that occurs for large \( \kappa_{\text{off}} \) distinguishes between the fine-scale features of the polymer model but that long-range hopping leads to identical results for different polymer models (e.g., the wormlike-chain and Gaussian-chain results shown here).

The preceding observations suggest an optimal set of parameters that result in a maximum transport rate along the strand. We note that the Gaussian-chain results (dashed curves in Fig. 6 B) accurately capture the wormlike-chain results (solid curves) near the maximum. Using the small-\( \kappa_{\text{off}} \) results for the off-time, \( \tau_{\text{off}} \approx \tau_{\text{off}}^{\text{(GC)}} \) (Eq. 22), we can approximate \( \sigma \) as

\[
\sigma \approx \frac{4\sqrt{5}\alpha^3 k_b \tau_{\text{off}}^{\text{(GC)}}}{\tau_{\text{on}} + \tau_{\text{off}}^{\text{(GC)}}}.
\]

Noting that \( \partial \sigma^{\text{(GC)}} / \partial k_b = -2\tau_{\text{off}}^{\text{(GC)}} / k_b \), the optimal translocation rate (with \( \partial \sigma / \partial k_b = 0 \)) occurs when there are equal average times spent by the protein in the on- and off-states,

\[
\tau_{\text{on}} = \tau_{\text{off}}^{\text{(GC)}}.
\]

Biophysical Journal 98(12) 2943–2953
i.e., $t_{on} = t_{off}^{(GC)}$. This rate-optimization condition holds for the class of polymers that we consider in this article; namely, polymer models that behave as an ideal random walk in the long-length limit.

The optimization condition of equal times on and off the DNA was derived previously by finding the optimal number of sites scanned during a sliding event that would lead to the fastest overall search time (43). Similar treatments optimize sliding lengths (29) and dissociation frequencies (44). We note, however, that the reasoning behind this conclusion is fundamentally different from our approach. Therein, the search time is optimized through the one-dimensional sliding time (governed in our model by the unbinding rate constant $k_u$), whereas our model relies on optimizing the superdiffusive acceleration (exemplified by $\sigma$) by means of the time the protein spends off the DNA (governed in our model by the binding rate constant $k_b$ and the polymer configuration).

An ideal target-site search strategy must not only allow the protein to quickly translocate but also to efficiently cover the strand. Ideally, the search process would not oversample basepairs or altogether skip over vast swaths of the genome. So far, our discussion has only focused on the former aspect of the search. To analyze search efficiency, we define the search-length ratio

$$\Gamma_1 = \lim_{\beta \to 0} \left( \frac{\langle |s - S_0|^\beta \rangle_{on}}{\langle |s - S_0|^\beta \rangle_{off}} \right) = \frac{l_{on}}{l_{off}} \quad (24)$$

as the ratio of the average distance traveled by the protein in the on- and off-states per cycle. A surface plot of this metric as a function of $k_u$ and $k_b$ is shown in Fig. 7. The dotted line once again represents the transition from short- to long-range capture domains. Parameter values that give $\Gamma_1 >> 1$ represent an inefficient oversampling of the DNA strand during the search where the protein revisits sites that have already been traversed. Conversely, values that give $\Gamma_1 << 1$ point to cases where the hops are so large that the protein tends to pass over sites without properly identifying them. Neither scenario is desirable, and we assert that search efficiency is optimized when the search lengths are balanced with $\Gamma_1 = 1$. The dashed line represents the values of $k_u$ and $k_b$ that give this optimized efficiency. The three dots marked A–C correspond to the three-parameter sets used in the simulations within Fig. 5.

From our analysis, we identify the set of kinetic parameters that optimize the translocation rate ($\tau_{on} = \tau_{off}^{(GC)}$) and

$$\Gamma_1 = \lim_{\beta \to 0} \left( \frac{\langle |s - S_0|^\beta \rangle_{on}}{\langle |s - S_0|^\beta \rangle_{off}} \right) = \frac{l_{on}}{l_{off}} \quad (24)$$

as the ratio of the average distance traveled by the protein in the on- and off-states per cycle. A surface plot of this metric as a function of $k_u$ and $k_b$ is shown in Fig. 7. The dotted line once again represents the transition from short- to long-range capture domains. Parameter values that give $\Gamma_1 >> 1$ represent an inefficient oversampling of the DNA strand during the search where the protein revisits sites that have already been traversed. Conversely, values that give $\Gamma_1 << 1$ point to cases where the hops are so large that the protein tends to pass over sites without properly identifying them. Neither scenario is desirable, and we assert that search efficiency is optimized when the search lengths are balanced with $\Gamma_1 = 1$. The dashed line represents the values of $k_u$ and $k_b$ that give this optimized efficiency. The three dots marked A–C correspond to the three-parameter sets used in the simulations within Fig. 5.

From our analysis, we identify the set of kinetic parameters that optimize the translocation rate ($\tau_{on} = \tau_{off}^{(GC)}$) and

$$\Gamma_1 = \lim_{\beta \to 0} \left( \frac{\langle |s - S_0|^\beta \rangle_{on}}{\langle |s - S_0|^\beta \rangle_{off}} \right) = \frac{l_{on}}{l_{off}} \quad (24)$$

as the ratio of the average distance traveled by the protein in the on- and off-states per cycle. A surface plot of this metric as a function of $k_u$ and $k_b$ is shown in Fig. 7. The dotted line once again represents the transition from short- to long-range capture domains. Parameter values that give $\Gamma_1 >> 1$ represent an inefficient oversampling of the DNA strand during the search where the protein revisits sites that have already been traversed. Conversely, values that give $\Gamma_1 << 1$ point to cases where the hops are so large that the protein tends to pass over sites without properly identifying them. Neither scenario is desirable, and we assert that search efficiency is optimized when the search lengths are balanced with $\Gamma_1 = 1$. The dashed line represents the values of $k_u$ and $k_b$ that give this optimized efficiency. The three dots marked A–C correspond to the three-parameter sets used in the simulations within Fig. 5.

From our analysis, we identify the set of kinetic parameters that optimize the translocation rate ($\tau_{on} = \tau_{off}^{(GC)}$) and

$$\Gamma_1 = \lim_{\beta \to 0} \left( \frac{\langle |s - S_0|^\beta \rangle_{on}}{\langle |s - S_0|^\beta \rangle_{off}} \right) = \frac{l_{on}}{l_{off}} \quad (24)$$

as the ratio of the average distance traveled by the protein in the on- and off-states per cycle. A surface plot of this metric as a function of $k_u$ and $k_b$ is shown in Fig. 7. The dotted line once again represents the transition from short- to long-range capture domains. Parameter values that give $\Gamma_1 >> 1$ represent an inefficient oversampling of the DNA strand during the search where the protein revisits sites that have already been traversed. Conversely, values that give $\Gamma_1 << 1$ point to cases where the hops are so large that the protein tends to pass over sites without properly identifying them. Neither scenario is desirable, and we assert that search efficiency is optimized when the search lengths are balanced with $\Gamma_1 = 1$. The dashed line represents the values of $k_u$ and $k_b$ that give this optimized efficiency. The three dots marked A–C correspond to the three-parameter sets used in the simulations within Fig. 5.

From our analysis, we identify the set of kinetic parameters that optimize the translocation rate ($\tau_{on} = \tau_{off}^{(GC)}$) and

$$\Gamma_1 = \lim_{\beta \to 0} \left( \frac{\langle |s - S_0|^\beta \rangle_{on}}{\langle |s - S_0|^\beta \rangle_{off}} \right) = \frac{l_{on}}{l_{off}} \quad (24)$$
the search efficiency ($l_{on} = l_{off}$). These optimal-parameter limits define a search enhancement window wherein both of these imperatives are balanced. Therefore, it is in the best interest of the cell for DNA-binding proteins to operate within this range. We find that this is indeed the case for EcoRV, whose rate parameters at various salt concentrations (11,45) place it in this window, shown in Fig. 8. Kinetic parameters for EcoRI (46,47) fall exactly on the optimized translocation rate line. Other representative enzymes at physiologically relevant salt concentrations also fall within or close to this search enhancement window, including human RAD51 protein (interacting with ssDNA) (48) and lac (7,49) and cro (47,50) repressors, all of which are plotted in Fig. 8. The proteins shown appear to slightly favor speed over efficiency, which might be due to their primary roles of defense (restriction) and repair (recombinases). Our model and analysis show how proteins use their nonspecific binding ability to optimize their search for target-sites without sacrificing site-scanning fidelity.

CONCLUSIONS

This work presents a theoretical model for facilitated diffusion of a protein along DNA. The model formulation provides a flexible platform for studying the roles of DNA conformation, as well as protein transport and kinetics. The relative contributions of the sliding and jumping mechanisms proposed by Berg et al. (2) are given as functions of kinetic binding and unbinding constants, thus allowing for direct investigation of their effect on facilitated diffusion. We generalized this approach by deriving a composite Green’s function for two-state transport with applicability beyond the treatment given in this article. Modeling DNA as an infinite wormlike chain, we find that translocation along DNA becomes superdiffusive due to power-law tails in the hop distribution, which arise from a protein’s ability to bind onto distal segments of DNA.

An important feature of our model is that it analytically predicts the effective translocation-rate coefficients from basic kinetic and transport parameters. Our analytical results allow us to explore the optimization of the search and reach important conclusions regarding the overall search strategy of DNA-binding proteins. Ideally, a target-site search strategy optimizes the translocation rate (achieved by the protein spending equal times on and off the DNA) and the search efficiency (achieved by the protein traversing equal lengths on and off the DNA), defining a search enhancement window. This careful balance of priorities is found in nature, with several DNA-binding proteins falling within this window.

The focus of this work is on the translocation strategies of DNA-binding proteins and is the first step in a complete physical description of target-site search localization dynamics. In a future publication, we will address the presence of a specific target site, the effects of finite-length DNA, and how the resulting predictions compare with experimentally determined localization rates (42) and hop distributions (30). Preliminary simulations of the target-site search process indicate that the localization rate is also influenced by factors other than the translocation rate and can be optimized by varying $\kappa_{on}$, as previously predicted (42). Further work will consider DNA dynamics, supercoiling, and genome packaging, as well as the effects of protein crowding, transient exposure, and confinement. Those studies will exploit the ability of our model to incorporate DNA configurational behavior and to capture dynamic phenomena involving molecular binding and unbinding events. Our future work will build on the fundamental understanding provided by the theoretical approach presented here, revealing further insights into the rich physical phenomena underlying target-site search by DNA-binding proteins.

SUPPORTING MATERIAL

Five Appendix sections and one table are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(10)00349-8.

This work was supported by the following funding sources: Bio-X Graduate Student Fellowship, the Hertz Foundation, National Science Foundation Graduate Research Fellowship, and National Science Foundation Career Award.

REFERENCES


