Downstream Exploration of DNA-Bound Searching Proteins: a Diffusion-Reaction Model

Michael L Mayo, Edward J Perkins  
Environmental Laboratory, Engineer Research and Development Center  
US Army Corps of Engineers  
Vicksburg, MS 39180.  
Email: Michael.L.Mayo@usace.army.mil, Edward.J.Perkins@usace.army.mil

Preetam Ghosh  
Department of Computer Science  
Virginia Commonwealth University  
Richmond, VA 23284.  
Email: pghosh@vcu.edu

Abstract—Proteins search along DNA for targets (e.g. transcription initiation sequences) through a combination of sliding, jumping, and intersegment transfers, wherein the sliding process proceeds until the protein-DNA complex dissociates. As such, we propose a diffusion-reaction model of the sliding phase of a total search process, and study the effect of varying reaction rate on the detailed search kinetics. With increasing dissociation rate, only the “fastest” proteins survive to explore downstream sequences, but at the price of increasing rarity, allowing for fluctuations to dominate the behavior at late times and long distances. Predictions of this model for the total search time agree with experimental estimates across an order of magnitude, providing bounds on dissociation rates that suggest a balance is established between maximizing downstream exploration of the DNA while minimizing the number of rare trajectories to guarantee the greatest chance of a successful search.

Keywords—Diffusion-Reaction; Random Environment; Facilitated Target-Location; Protein-DNA Interaction; Monte Carlo method

I. INTRODUCTION

To initiate transcription, proteins bind nonspecifically to DNA, wherein they undergo short periods of sliding interrupted by jumping events that collectively—and very rapidly—transport searching proteins toward target sites to the promoter region of an appropriate gene. Together, these rounds of sliding and jumping are termed “facilitated target location” [1]. While several seminal theoretical works [2], [3], [4] and many experimental observations [5], [6] firmly establish the need for its inclusion into any model explaining the rapid response of transcriptional systems, the reactive kinetic properties of the sliding phase have yet to be completely studied.

Here, we propose a diffusion-reaction model of protein transport along bacterial DNA under dissociative conditions of the protein-DNA complex, and investigate its implications for protein target-location times by Monte Carlo simulations. As justified below, we consider the sliding phase along nonspecific DNA sequences to occur mainly within an uncorrelated random environment, wherein the probabilities associated with jumping to adjacent sites are chosen at random according a Gaussian-type distribution. As such, the current model falls within the field of probability theory, wherein the seminal works of Sinai [7] showed that for late times, the root-mean-square displacement scales as $\sim \ln^2 t$ (under special conditions for the transition probability distribution), rather than $\sim t^{1/2}$, as expected from “ordinary” diffusive systems. Other notable results include the study of random walks in the presence of random forces [8], and correlated disorder [9].

We use our model to express the average time required to arrive at a given downstream target site for the first time, termed the mean first-passage time—a property commonly used to estimate transit times within a medium, and reaction rates with a boundary, among many other interpretations. In the biological application that we emphasize here, the mean first-passage time places an estimate on the time required for a transcription factor to recruit RNA polymerase to the promoter region of the target gene, under reaction-limited conditions (RNA polymerase is assumed to be present in excess, while also fully accessible to the promoter region). The approximate total search time includes many consecutive rounds of protein sliding and jumping, and can be written as [10]

$$t_s^{avg} = \frac{L}{\Lambda} (t_{\Lambda} + t_{3D}),$$  

(1)

providing an estimate of the rate for transcriptional activation, $r = 1/t_s^{avg}$, under conditions of irreversible elongation initiated by the RNA polymerase binding. In Eqn. 1, $L$ is the total length of an average bacterial genome; $\Lambda$ is the average length along the DNA that a molecule explores before it “reacts” (i.e. dissociates) with the DNA, termed the exploration length [11]; $t_{\Lambda}$ is the average time needed to explore a distance along the DNA equal to the exploration length; $t_{3D}$ is the average time the protein spends free of the DNA. Under “optimal” conditions the search time is minimized, so that $t_{\Lambda} = t_{3D}$, and the approximate total search time can then be found from Eqn. 1:

$$t_s^{avg} \approx 4 \times 10^4 t_{\Lambda},$$  

(2)
Figure 1. One dimensional model of protein sliding on DNA. Adjacent site-to-site transitions are defined by energetic barriers derived from the relative difference in the protein binding free energies.

given that the typical bacterial genome is approximately \( L \approx 10^8 \) bps; experiments on the BbvCI restriction enzyme [12] suggest a value for the exploration length of \( \Lambda \approx 50 \) bps. Equation 2 relies only upon estimates of the time, \( t_A \), for a protein to cover about 50 bases under biologically reactive conditions. The determination of this quantity, and its implications for the speed of transcription initiation, is the main subject of this article.

II. THE MODEL

Upon binding to the DNA the protein moves along its contour, effectively reducing the dimensionality of the diffusive target-location process from three (in bulk) to one (on DNA). As such, we consider a segment of DNA of length \( N \) (in units of bases) to be a one-dimensional lattice, wherein the lattice sites host base-pairs separated by a lattice constant given by \( \Delta l = 0.34 \text{nm} \) (Fig. 1). Diffusion along the lattice proceeds according to probabilities describing the success of a protein to “hop,” or transition, to or from adjacent lattice sites. Such hopping proteins face energetic barriers, which determine the speed of the diffusion, among other characteristics.

A. Energetic basis of protein diffusion along DNA

The free energy (with respect to the cytoplasm) related to protein-DNA binding can be estimated from sequence data using the position weight matrix method [13], and has been carried out for several bacterial proteins [10]. Such free energy profiles exhibit some dependence on the actual nucleotide sequence. Following tradition [10], [15], [16], we decompose the total binding free energy into sequence-dependent, \( U(\vec{s}) \), and -independent, \( E_{ns} \), parts. The former quantity is termed specific, while the later is termed nonspecific:

\[
U_i = U(\vec{s} = \{s_i, \ldots, s_{i+l-1}\}) + E_{ns},
\]

wherein \( \vec{s} = \{s_i, \ldots, s_{i+l-1}\} \) is the nucleotide binding sequence of length \( l \), and \( s_i \in \{A, C, G, T\} \) labels a base [10].

For an arbitrary DNA segment, the total binding free energy, \( U_i \), is approximately Gaussian-distributed [10]:

\[
\rho(U_i) = \frac{1}{\sqrt{2\pi\mu^2}} e^{-(U_i-(\mu))^2/2\mu^2},
\]

supporting a view that the interaction between protein and DNA can be approximated as a Gaussian random variable. The identity of the actual protein is contained within the free energy terms according to its mass, charge, and so on. Here and below, the mean value of this free energy distribution is taken to be \( (U) = 0k_BT \), with \( k_B \) and \( T \) equal to Boltzmann’s constant and the system temperature, respectively. Biological evidence suggests these binding energies are mostly uncorrelated [10], so we adopt this assumption here; however, nontrivial energetic correlations across bases may potentially arise from several sources, such as AT/GC-rich isochors [14], in which long nucleotide sections exhibit nonrandom nucleotide preference, or they may arise from geometrical considerations, such as DNA curvature/elasticity [14].

For a random walker that hops to site \( i \pm 1 \) from its current site \( i \), the energy barrier it faces is \( U_i - U_{i\pm1} \); the probability to jump per second, \( \omega_{i\pm1, i} \), can be shown to acquire the following Arrhenius form [10], [17]:

\[
\omega_{i\pm1, i} = \frac{\nu}{2} \times \begin{cases} 
1 & \text{for } U_{i\pm1} \geq U_i \\
0 & \text{otherwise}
\end{cases}
\]

In this expression, \( \nu \) is the maximum number of hop attempts per second, which has been previously estimated to be approximately \( 10^8 \) Hz [10]; the factor of two follows from normalization, i.e., \( \max\{\omega_{i+1, i} + \omega_{i-1, i}\} = \nu \). In addition, the probability to not hop may be nonzero, \( \nu - \omega_{i+1, i} - \omega_{i-1, i} \geq 0 \), and is termed the sojourn probability [18].

At any time during the walk, the protein-DNA complex may dissociate from site \( i \) with rate \( r_i \) (Fig. 1), triggering protein macro- and micro-hopping events, depending largely on the nonspecific energy contributions [10] to the total binding free energy. Along with Eqs 3-5, the inclusion of the reaction rate \( r_i \) establishes the protein target-location as a diffusion-reaction process along the DNA in one-dimension. Figure 1 shows the diffusion-reaction model proposed here; diffusion proceeds along the DNA, a one-dimensional lattice of length \( N \) (in units of base-pairs), wherein transitions...
between sites occur according to transition probabilities defined by Eqn. 5.

When a sufficient quantity of RNA polymerase is accessible to the DNA, the simplest view of transcription is one in which it is initiated (or inhibited) by the arrival of transcription factors (or repressors). The main quantity of interest, then, is the mean first-passage time, \( \tau_i \), to arrive at the target site \( i \), downstream of the site of first association, wherein the rate of transcription will be given approximately by \( 1/\tau_i \).

### B. Mean first-passage time

The first-passage distribution, \( F_{i,1} \), for a random walk beginning from the origin of the lattice (\( i = 1 \)), provides the probability per unit time, \( \tau_i \), that a protein reaches a site \( i = 1, 2, \ldots, n \) for the first time, and can be written in terms of the occupation probabilities, \( p_{i,1}(t) \), for each site along the lattice at the elapsed time \( t \) [19]:

\[
p_{i,1}(t) = \delta_{i1} \delta(t) + \int_0^t F_{i,1}(t')p_{i,1}(t-t')dt'.
\]  

(6)

The first term on the right hand side of Eqn. 6 ensures that the probability is sharp at \( i = 1 \) and \( t = 0 \). The quantity \( p_{i,1}(t-t') \) in the second term gives the probability to find a protein at site \( i \) during the time period \( t-t' \geq 0 \), after it has reached this site for the first time.

Equation 6 is a convolution, and a Laplace transform (with respect to time) can be applied to Eqn. 6 to decouple the first-passage distribution from the occupation probabilities. After solving for the transformed first-passage distribution in terms of the transformed occupation probabilities, the inverse Laplace transform can be computed, for example, by the Bromwich integral. From this distribution the mean first-passage time can be computed,

\[
\tau_i = \int_0^\infty t F_{i,1}(t)dt.
\]

(7)

However, the first-passage distribution must satisfy \( \int_0^\infty F_{i,1}(t)dt = 1 \), which is not the case when probability is not conserved by reaction with the lattice. Instead, we use exact numerical simulations of the transport process to directly measure the mean first-passage time, Eqn. 7, by implementing the following Monte Carlo algorithm.

### III. NUMERICAL EVALUATION OF THE MODEL BY MONTE CARLO SIMULATION

Exact numerical simulations of the model were conducted to study the diffusion-reaction process in terms of the mean first-passage time for a protein to reach a downstream target site \( i = n \) (\( n = 2, 3, \ldots, N \)), using a variation of the Monte Carlo method. Simulations were run for individual trajectories and the first-passage times recorded for each site along the lattice. An arithmetic average of these times provides an approximation of the mean first-passage time (Eqn. 7).

The simulation begins with a attachment event, in which a protein is found with unit probability at the origin. During each simulation time-step (in as units of \( 1/\nu \)) several checks are made that define the site-to-site transport along the lattice.

Boundary conditions must be provided at the beginning and end of the lattice segment to actually simulate the random walk. At the target site \( i = n \) (defined by the user), the simulation terminates when 1000 “complete trajectories” are satisfied; i.e., when a protein starting at the origin survives to reach the target site exactly 1000 times. If it reaches the target site, we increment a counter for complete trajectories, and another trajectory is started. If, at any time, the check for dissociation returns TRUE, the current trajectory is terminated and another is started again at site \( i = 1 \).

At the origin, reflecting boundary conditions are provided to better quantify the exploration length from the site of association. Even though the walk is initially influenced by this boundary condition, this influence vanishes at biologically-relevant time (~ 1 second) and length scales (~ 100 bases), due to the uncorrelated diffusive nature of the transport process. These boundary conditions therefore provide a suitable model of the biological protein transport process under study.

The data structures for the algorithm Monte Carlo Random Walk (MCRW), are depicted in Fig. 2 and outlined as follows:

1) A 2D matrix, \( \text{total\_simulation\_step} \), records the number of hops traversed by the protein in each trajectory.
The diffusion-reaction along the DNA, given that it began at $x = 75$.) The superdiffusive trajectories (of bases in an equivalent time–survive along the segment to reach the downstream targets during the sliding phase. However, as this probability increases, such trajectories become increasingly rare, indicating that transcription cannot possibly rely upon sliding alone—an idea generally accepted since the early seminal works by Berg, Winter and von Hippel (see introduction). If sliding were the only mechanism, a balance between maximizing downstream exploration and minimizing the overall search times would be required to waste a minimum of sparse biological resources (typically, there are ~ 10-100 transcription factors available to a gene).

Interestingly, the reaction-free trajectory, $\sigma = 0$, scales as $x \sim t^{1/2}$; an increase in the reaction rate filters all but the superdiffusive trajectories ($x \sim t$). This is consistent with intuition: an increasing reaction rate selects for ever faster surviving trajectories, which should approach the theoretical maximum (see linear trajectory within Fig 3). This behavior cannot hold asymptotically (i.e., $\sigma \to 1$), because no molecules survive on the lattice to be transported anywhere downstream of $i = 1$. 

Figure 3 illustrates the mean first-passage times for a searching protein along the DNA for a segment of 100 bases. As mentioned above, experiments on the reaction enzyme BbvCI [12] indicate that the average exploration length is approximately 50 bases, suggesting that an appropriate length scale for the problem at hand is about 100 bases. In general, the mean first-passage time decreases with increasing dissociation probability, demonstrating that only the “faster” molecules—those molecules clearing a maximum of bases in an equivalent time—survive along the segment to reach the downstream targets during the sliding phase. However, as this probability increases, such trajectories become increasingly rare, indicating that transcription cannot possibly rely upon sliding alone—an idea generally accepted since the early seminal works by Berg, Winter and von Hippel (see introduction). If sliding were the only mechanism, a balance between maximizing downstream exploration and minimizing the overall search times would be required to waste a minimum of sparse biological resources (typically, there are ~ 10-100 transcription factors available to a gene).

Interestingly, the reaction-free trajectory, $\sigma = 0$, scales as $x \sim t^{1/2}$; an increase in the reaction rate filters all but the superdiffusive trajectories ($x \sim t$). This is consistent with intuition: an increasing reaction rate selects for ever faster surviving trajectories, which should approach the theoretical maximum (see linear trajectory within Fig 3). This behavior cannot hold asymptotically (i.e., $\sigma \to 1$), because no molecules survive on the lattice to be transported anywhere downstream of $i = 1$. 

Figure 5. Root-mean-square displacement along the DNA, $\sqrt{\langle \Delta x^2 \rangle}$, computed from MCRW for various values of the reaction rate $\sigma$. 

Note that while we have only discussed the measurement of the mean first-passage times using MCRW, it is straightforward to compute other quantities, too, such as the root mean square displacement, and other statistical endpoints as they vary with the dissociation probability, random energy background, target site, and so on.

IV. RESULTS

A. Preferential selection of fast-moving proteins

For bacterial cells, we interpret the mean first-passage time as the minimum time needed for a protein to locate its target along the DNA purely by sliding. As mentioned above, the actual physiological search process involves rounds of sliding interrupted by jumping events that can be either long or very short ($\approx 1$ base [20]). The diffusion-reaction model, as presented herein, naturally describes a sliding phase terminated by jumping events that are initiated by dissociation of the protein-DNA complex. We emphasize, then,
B. Proteins are screened from downstream exploration by the reaction rate

Figure 4 illustrates the probability for a trajectory to ever pass a given lattice site. As expected, the probability for a trajectory that passes the indicated base (i.e. 10, 25, 50, or 75 bps) decays with distance. For trajectories that survive to pass bases nearest to the origin (i.e. 10, 25 bps), a qualitative transition is observed in at approximately 50 bases (panels (a) and (b)), beyond which, the probability is strongly quenched. We conclude, then, that most molecules dissociate before reaching 50 bases under the reactive conditions studied, but of those that survive to this distance, many also reach the 100th base; random walks surviving to 50 bases are “faster” than their shorter-living counterparts, covering the most ground in the same time spent exposed to the reactive media.

Figure 5 illustrates the root-mean-square displacement for several trajectories under varying conditions of protein dissociation. As the random walks approach an average displacement of approximately 50 bases, fluctuations begin to dominate the kinetics for times greater than about $10^4$ simulation steps, indicating that fewer trajectories survive past 50 bases. In other words, the emergence of strong fluctuations arise here due to statistical calculations conducted upon an increasingly smaller sample size as distance from the origin increases (and distance to the far boundary decreases).

Collectively, these analyses describe two qualitative transitions along the segment of 100 bases: one in space at approximately 50 bases, and one in time at approximately $10^4$ steps. By this time, however, reaction events with the medium and the far boundary dominate the behavior of the surviving proteins. For intermediate sites on longer segments, boundary effects are less visible, leaving only reaction with the medium to affect the diffusion characteristics.

In any case, it is clear that an increasing rarity of surviving trajectories is problematic for biological processes, wherein such rare events are the most uncertain. We conclude from these analyses that reaction with the medium must be balanced by downstream exploration to provide the most likely chance of timely, and reliable, transcriptional activation.

C. Total search times under reactive conditions

As described in the Introduction, Eqn. 2 provides an expression for the “optimal” overall search time for a single protein search process incorporating both sliding and jumping from one site to another adjacent one. Strictly speaking, we would need to invert the expression $\Lambda(\tau_s) = 50$ bps, to find the time required to explore a distance of approximately 50 bps before the protein-DNA complex experiences a reaction event and dissociates. The first-passage time, however, bounds this value from below, $\tau_{50} \leq \tau_s$, so that the fastest possible search time is approximately

$$ t_{50}^{avg} \geq t_{min} = 4 \times 10^4 \tau_{50}. \quad (8) $$

As the first-passage time depends on the particular value of the dissociation rate of the protein-DNA complex, the minimum search time varies parametrically with it, too.

Figure 6 illustrates the mean first-passage time to arrive at the 50th base (left axis) as a function of the dissociation probability for one particular realization of the binding free energy landscape (see inset of Fig 3). Numerical simulations were found to be computationally prohibitive for reactive regions $\sigma > 10^{-3}$. However, as the dissociation probability approaches one, $\sigma \rightarrow 1$, no molecules survive along the lattice, so the mean first-passage time in this limit is zero for all downstream sites. Exactly how this limit is reached, and the particular value of the critical reaction rate, $\sigma_c$, will be discussed elsewhere.

Assuming that the maximum hopping frequency is $\nu = 10^8$ Hz and a bacterial genome is approximately $L = 10^6$ bps, Fig. 6 provides an estimate of the minimum search time needed for a protein to locate its target along a typical bacterial genome (right axis), provided $\Lambda = 50$ bps.

Estimates derived from experimental data place this value in the range of 1-10 seconds [10], in good agreement with the values predicted by the diffusion-reaction model presented here. In particular, the biological reaction rate implied here may be bounded from above by $\sigma \leq 10^{-3}$, and bounded from below by $\sigma \geq 0$. Further experimental and theoretical work is needed to place more restrictive bounds on this value in the computationally-prohibitive regions ($\sigma > 10^{-3}$).

V. Conclusion

A one-dimensional diffusion-reaction model was proposed to analyze the nonequilibrium protein sliding kinetics along a segment of bacterial DNA, and evaluated numerically by Monte Carlo simulation. Downstream targets are screened
from access by upstream diffusing proteins according to the value of the dissociation rate of the protein-DNA complex; however, anomalous characteristics of the surviving associated proteins allow for faster search times under an increasing dissociation rate.

Estimates of the minimum “optimal” search time derived from recent experimental work places bounds on the rate of protein dissociation consistent with predictions of the proposed model, for a wide range of reaction/dissociation rates. The proposed model supports a view in which the entirety of a typical bacterial genome generates an uncorrelated binding free energy profile with a typical protein, suggesting that long-range correlations can be safely neglected; however, more work is required to arrive at a similar result for short-range correlations. Moreover, access limitations induced by reaction/dissociation with the DNA suggests that nature balances downstream exploration with the competing need for faster search times, possibly allowing for the most “efficient” use of resources, i.e., the few transcription factors typically present for each gene.

Reliability in the face of fluctuations is an overall theme in biological regulation; in the case of transcriptional activation by searching proteins, fluctuations may be minimized by reducing the reaction rate (or equivalently, increasing the exploration length), reducing the temperature thereby reducing the energetic barriers to base-to-base transport (though, at the cost of more weakly-bound proteins), or by simply increasing the number of searching proteins accessible to the DNA.

ACKNOWLEDGMENT

Funding was provided by the US Army Environmental Quality and Installations 6.1 Basic Research Program. The Chief of Engineers approved this material for publication.

REFERENCES