How a protein searches for its site on DNA: the mechanism of facilitated diffusion

2009 J. Phys. A: Math. Theor. 42 434013

(http://iopscience.iop.org/1751-8121/42/43/434013)

IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **42** (2009) 434013 (23pp) doi:10.1088/1751-8113/42/43/434013

How a protein searches for its site on DNA: the mechanism of facilitated diffusion

Leonid Mirny, Michael Slutsky, Zeba Wunderlich, Anahita Tafvizi, Jason Leith and Andrej Kosmrlj

Harvard-MIT Division of Health Sciences and Technology, Department of Physics, Massachusetts Institute of Technology, 77 Mass ave, Cambridge, MA 02139, USA

E-mail: leonid@mit.edu

Received 19 June 2009, in final form 21 September 2009
Published 13 October 2009
Online at stacks.iop.org/JPhysA/42/434013

Abstract

A number of vital biological processes rely on fast and precise recognition of a specific DNA sequence (site) by a protein. How can a protein find its site on a long DNA molecule among 10^6–10^9 decoy sites? Here, we present our recent studies of the protein–DNA search problem. Seminal biophysical works suggested that the protein–DNA search is facilitated by 1D diffusion of the protein along DNA (sliding). We present a simple framework to calculate the mean search time and focus on several new aspects of the process such as the roles of DNA sequence and protein conformational flexibility. We demonstrate that coupling of DNA recognition with conformational transition within the protein–DNA complex is essential for fast search. To approach the complexity of the *in vivo* environment, we examine how the search can proceed at realistic DNA concentrations and binding constants. We propose a new mechanism for local distance-dependent search that is likely essential in bacteria. Simulations of the search on tightly packed DNA and crowded DNA demonstrate that our theoretical framework can be extended to correctly predicts search time in such complicated environments. We relate our findings to a broad range of experiments and summarize the results of our recent single-molecule studies of a eukaryotic protein (p53) sliding along DNA.

PACS numbers: 87.15.kj, 87.10.−e, 87.15.rp

(Some figures in this article are in colour only in the electronic version)

1. Introduction

1.1. The protein–DNA search problem

Many biological processes are initiated by a single protein binding its specific target sequence (target site) on a long DNA molecule. In the search for its target sequence, such a protein

is faced with two difficulties, one thermodynamic and the other kinetic. The thermodynamic challenge lies in recognizing and tightly binding the target site among the billions of other non-specific DNA sequences. The kinetic difficulty is finding the target site in mere seconds amidst the crowded cellular environment filled with other DNA sequences and proteins. A wide variety of biophysical studies have contributed to a deeper understanding of the thermodynamics of protein–DNA interactions. However, little is known about the kinetics of the search process and the roles of DNA sequence, DNA spatial organization and other DNA-bound proteins.

1.2. History of the problem: experiment

The problem of how a protein finds its target site on DNA has a long history. In 1970, Riggs *et al* [1] measured the association rate of the Lac repressor and its target site on DNA as

$$k_{\text{exp}} = 10^{10}\ \text{M}^{-1}\ \text{s}^{-1}.$$

This astonishingly high rate, which is 100–1000 times higher than any known protein–protein association rate, was also shown to be much higher than the maximal rate of protein–DNA association achievable by diffusion in solution.

Indeed, binding of a protein (P) to a target site (S) on DNA is a bimolecular association reaction

$$\text{P} + \text{S} \rightleftarrows \text{PS}$$

characterized by its rate k_a which is defined macroscopically (and measured experimentally) as a coefficient in $\frac{\text{d[PS]}}{\text{d}t} = k_a[\text{P}][\text{S}]$. The expression for the diffusion-limited rate of bimolecular reactions was obtained by Smoluchowski and in the case of protein–DNA association has the form

$$k_{\text{Smol}} = 4\pi D_{3\text{D}} b a, \tag{1}$$

where $D_{3\text{D}}$ is the diffusion coefficient of the protein (assuming the site on DNA diffuses much more slowly than the protein), b is the cross-section of the binding reaction and a is the fraction of the molecular surface (of the protein) that contains the reactive binding interface. With this, let us estimate the value of the diffusion-limited rate k_{Smol}. We should set $b = 0.34$ nm, the spacing between the base-pairs of DNA, since displacement by a single base-pair (bp) leads to a different DNA sequence recognized by a protein (figure 1). Measured diffusion coefficients of proteins in aqueous environments have a range of $D_{3\text{D}} \approx (1\text{–}5) \times 10^{-6}\ \text{cm}^2\ \text{s}^{-1}$. We can assume a relatively large $a \approx 0.2\text{–}0.5$, since a protein has a relatively large reactive interface and is likely to become oriented correctly as it approaches the DNA by the electrostatic interactions of basic amino acids with negatively charged DNA. Using these numbers and converting to 'moles per liter' by multiplying by the Avogadro constant we obtain

$$k_{\text{Smol}} \approx 2 \times 10^{-19}\ \text{m}^3\ \text{s}^{-1} \approx 10^8\ \text{M}^{-1}\ \text{s}^{-1}.$$

For comparison, the rates for most protein–protein associations (under physiological salt concentrations, i.e., under well-screened electrostatic interactions) are in the range of 10^6–$10^7\ \text{M}^{-1}\ \text{s}^{-1}$ [2] and are indeed well below the Smoluchowski diffusion limit. This comparison demonstrates that the protein–DNA search is about 100 times faster than the diffusion limit and about 1000 times faster than bimolecular associations of protein molecules, suggesting that some special mechanism, referred to as *facilitated diffusion*, provides this speed-up.

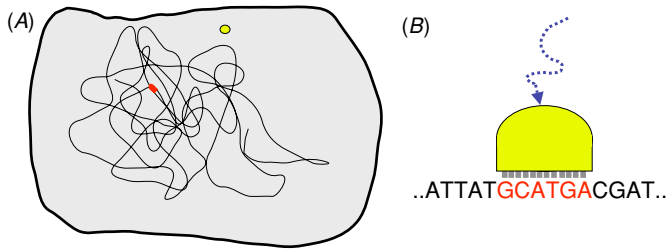

Figure 1. (*A*) Schematic representation of the protein–DNA search problem. The protein (yellow) must find its target site (red) on a long DNA molecule confined within the cell nucleoid (in bacteria) or cell nucleus (in eukaryotes). Compare with figure 9(*A*) which shows confined DNA. (*B*) The target site must be recognized with 1 base-pair (0.34 nm) precision, as displacement by 1 bp results in a different sequence and consequently a different site.

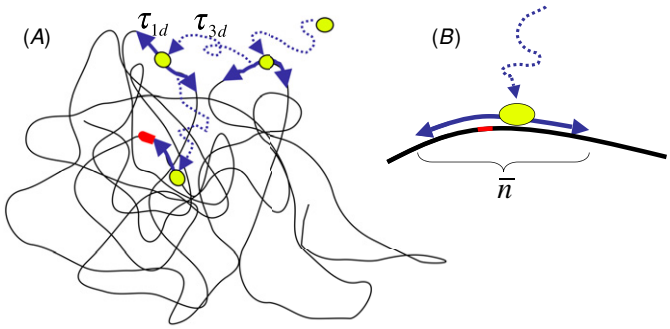

Figure 2. (*A*) The mechanism of facilitated diffusion. The search process consists of alternating rounds of 3D and 1D diffusion, each with average duration τ_{3D} and τ_{1D}, respectively. (*B*) The antenna effect [9]. During 1D diffusion (sliding) along DNA, a protein visits on average $\bar{n}$ sites. This allows the protein to associate some distance $\sim\bar{n}$ away from the target site and reach it by sliding, effectively increasing the reaction cross-section from 1bp to $\sim\bar{n}$. The antenna effect is responsible for the speed-up by facilitated diffusion.

1.3. History of the problem: theory

To resolve this discrepancy, one possible mechanism of facilitated diffusion that includes both 3D diffusion and effectively 1D diffusion of protein along DNA (*the 1D/3D mechanism*) was suggested. This mechanism was first proposed and dismissed by Riggs *et al* [1] but was soon revived and rigorously studied by Richter and Eigen [3], then further expanded and corrected by Berg and Blomberg [4] and finally developed by Berg *et al* [5]. The basic idea of the 1D/3D mechanism is that while searching for its target site, the protein repeatedly binds and unbinds DNA and, while bound non-specifically, slides along the DNA, undergoing one-dimensional (1D) Brownian motion or a random walk. Upon dissociation from the DNA, the protein diffuses three dimensionally in solution and binds to the DNA in a different place for the next round of one-dimensional searching (figure 2(*A*)).

During 1D sliding the protein is kept on DNA by the binding energy to non-specific DNA. This energy has been measured for several DNA-binding proteins and has a range of 10–15 k_BT (at physiological salt concentration), was shown to be driven primarily by screened electrostatic interactions between charged DNA and protein molecules [6], and

is thus highly sensitive to the concentration of salt ions. The microscopic mechanism of protein–DNA translocation and events that trigger protein dissociations are yet to be understood.

1.3.1. Open questions. Despite significant progress made by early theoretical studies in explaining faster-than-diffusion search and a broad range of experimental observations, aspects of the protein–DNA search from the molecular to the cellular levels remained poorly understood:

- *Molecular level—DNA sequence and protein conformation.* The role of DNA sequence in the search process remains largely unexplored. The sequence-specific energy of protein–DNA binding leads to a rugged energy landscape for protein translocation along DNA. How fast can a protein slide and search despite of this ruggedness?

  DNA-binding proteins exhibit complex intramolecular dynamics that can influence the efficiency of the search process, e.g. commonly observed coupling of folding and binding. How does conformational transition affect search?
- *Spatial effects*. What is the molecular mechanism of protein translocation along DNA? Does a protein intermittently dissociate from and then re-associate to DNA while undergoing Brownian motion (hops) or does it stay on DNA for long periods of time (sliding)? And after dissociating from DNA, the protein undergoes a round of 3D diffusion before binding DNA again (a jump). How far along DNA does the protein re-associate? Do the statistics of the jump lengths $P(x)$ resemble the Levy flight distribution $P(x) \sim x^{-\alpha}$ $(1 < \alpha < 3)$?
- *Cellular level.* How does the search proceed in the cell? This question has become the central focus of many studies as single-molecule methods have made it possible to observe binding of single proteins in the cell. What is the role of the spatial organization of DNA? How fast is search in the cell where molecular crowding and highly complex DNA conformation can slow protein movement? What is the role of other DNA-bound proteins and DNA packing in chromatin?

Ultimately, we would like to understand the extent to which the protein–DNA search is facilitated *in vivo*, whether the remarkable physical mechanism of facilitated diffusion is essential for cell physiology, and whether the mechanism is universal for all DNA-binding proteins, both in prokaryotic and eukaryotic cells.

Recent theoretical studies of protein–DNA search include work by Halford and Marko [7], Coppey *et al* [8], Hu *et al* [9–11], Lomholt *et al* [12, 13] and other groups [14–16]. All these studies used the 1D/3D mechanism of facilitated diffusion as a basic framework.

Here we review several studies from our group and discuss them in the context of other experimental and theoretical results. We start by presenting the single-molecule theoretical approach [17] that our group has been developing (sections 2.1–2.1.2). Next we turn to a study concerned with the role of DNA sequence [17] (section 3.1) and arrive at the important speed-stability paradox (section 3.2), which suggests the essential role of conformational transitions in DNA-binding proteins [17]. We discuss our studies of the conformational transition in section 4.1, the mechanism of kinetic pre-selection (section 4.2) and the landscape model (section 4.3). The role of spatial effects examined in our recent works [18, 19] is discussed in section 6, where new and previously unpublished study of the search on the DNA molecule confined into a small volume is presented in section 6.3. We also present previously unpublished analysis of the search on crowded DNA (section 6.4).

2. Search on DNA

2.1. Theory: a single-molecule approach

In spite of this rich history, most of the analytical studies of facilitated diffusion remained rather complicated. Here we present our original approach which is rather transparent and intuitive. Consider a *single* protein searching for a *single* target site on a long DNA molecule of M bps by the 1D/3D mechanism. The search consists of multiple rounds, each consisting of one round of 1D diffusion followed by one round of 3D diffusion. Then the total search time is given by

$$t_s = \sum_{i=1}^{k} (\tau_{1D,i} + \tau_{3D,i}),$$

where $\tau_{1D,i}$ and $\tau_{3D,i}$ are the durations of 1D and 3D diffusion in the ith round of searching, and k is the number of rounds until the target site is found. The mean search time is then

$$\bar{t}_s = \bar{K}(\tau_{1D} + \tau_{3D}),$$

where τ_{1D} and τ_{3D} are the mean durations of 1D and 3D diffusion rounds, and $\bar{K}$ is the mean number of rounds until the target is found.

It is easy to estimate the average number of rounds until the target site is found among M alternatives. If during each round of sliding the protein scans $\bar{n} \ll M$ sites, the probability that site is found in a single round is $p = \bar{n}/M$. If each time the protein re-associates with the DNA, it does so uniformly along the DNA, i.e., scanning independent sets of n sites, the probability of finding the target (for the first time) on the kth round is given by the geometric distribution $(1-p)^{K-1}p$. The mean number of 1D/3D rounds is then $\bar{K} = 1/p = M/\bar{n}$, yielding the mean search time

$$\bar{t}_s = \frac{M}{\bar{n}}(\tau_{1D} + \tau_{3D}). \tag{2}$$

If 1D sliding proceeds by normal (non-anomalous) diffusion, then $\bar{n} \sim \sqrt{D_{1D}\tau_{1D}}$, where D_{1D} is the diffusion coefficient of sliding. For an exponentially distributed 1D time with a mean of τ_{1D} we obtained the mean number of visited sites (the distance between the left- and right-most visited sites) [19]

$$\bar{n} = 2\sqrt{D_{1D}\tau_{1D}}. \tag{3}$$

Equations (2) and (3) together provide a simple expression for the search time as a function of macroscopic measurable parameters τ_{3D}, D_{1D} and $\bar{n}$ (or τ_{1D}). Note that τ_{3D} in turn depends on the spatial diffusion coefficient D_{3D} and DNA density.

2.1.1. Immediate results. First, one can calculate the optimal partitioning of time between 1D and 3D diffusion modes during the search process. Plugging $\bar{n} \sim \sqrt{D_{1D}\tau_{1D}}$ into (2) and setting $d\bar{t}_s/d\tau_{1D} = 0$, we found that the fastest search is achieved if $\tau_{1D} = \tau_{3D}$ yielding the *optimal* search time

$$\bar{t}_{opt} = \frac{2M}{\bar{n}}\tau_{3D} = M\sqrt{\frac{\tau_{3D}}{D_{1D}}}. \tag{4}$$

Second, one can easily calculate the magnitude of the speed-up due to 1D/3D facilitated diffusion, as compared to a 3D-only or 1D-only mechanism. To obtain the mean search time for 3D diffusion alone, we set $\tau_{1D} = 0$ and $\bar{n} = 1$, yielding $\bar{t}_{3D} = M\tau_{3D}$. Comparison with equation (4) shows that facilitated diffusion is $\bar{n}/2$ times faster. The search time by 1D

diffusion alone is $\bar{t}_{1D} \sim \frac{M^2}{D_{1D}}$, which is $\sim\frac{M}{\bar{n}}$ slower than the optimal search time by facilitated diffusion (equation (4)).

Let us briefly consider an example. For a bacterial genome of $M = 5 \times 10^6$ bps and a sliding length of $\bar{n} = 200 - 500$ bps (e.g. [20]), facilitated diffusion provides the necessary speed-up by a factor of $\sim 10^2$ as compared to diffusion-limited search. Search by 1D sliding alone is extremely inefficient, being $\sim 10^4$ times slower than optimal facilitated diffusion.

Note that the latter results are obtained assuming optimal 1D/3D partitioning $\tau_{1D} = \tau_{3D}$ that, as we demonstrated [18], is not achieved in the cell due to high DNA concentration and high protein affinity for non-specific DNA (see below). Note also that if the search for the target site is performed simultaneously by m protein molecules, the mean search time is approximately m times smaller [17, 19].

2.1.2. Connection to the Smoluchowski equation: speed-up and slow-down. It is easy to see how the formalism presented above is connected to the Smoluchowski rate for bimolecular reactions. The rate and the mean time of the search process are connected by $\bar{t}_s = \frac{1}{k_s[T]}$, where $[T]$ is the concentration of the target sequence, which is related to the total DNA concentration $[T] = [\mathrm{DNA}]/M$. Note that τ_{3D} is the mean diffusion-limited time experienced by the protein before it interacts with *any* region of DNA, and thus, $\tau_{3D} = \frac{1}{k_{\mathrm{Smol}}[\mathrm{DNA}]}$. Using these expressions and equation (2) for the mean search time we arrive at the rate of the search reaction

$$k_s \approx k_{\mathrm{Smol}} \left(\frac{\tau_{3D}}{\tau_{1D} + \tau_{3D}} \right) \bar{n} = 4\pi D_{3D} \left(\frac{\tau_{3D}}{\tau_{1D} + \tau_{3D}} \right) \bar{n} a. \tag{5}$$

Two aspects of the search process become transparent from this equation. First, *the acceleration* of search by sliding effectively increases the cross-section from $b = 1$ bp to $\bar{n}$ base-pairs of DNA, allowing the protein to reach the target site by associating with $\bar{n}$ base-pairs around it. Hu *et al* [9] called this *the antenna effect* (figure 2(*B*)). The second effect is the *slow-down* due to non-specific binding of the protein to DNA. While searching for its target, the protein spends a certain fraction of its time bound to DNA far from the target and thus, not diffusing in 3D. This effect is manifested by the factor $\tau_{3D}/\tau_{1D} + \tau_{3D}$, which is the fraction of time the protein spends diffusing in 3D. Thus, binding non-specifically to DNA leads to a reduction of spatial mobility, which can be taken into account by an effective diffusion coefficient $D_{3D,\mathrm{eff}} = D_{3D}\tau_{3D}/\tau_{1D} + \tau_{3D}$.

Importantly, the slow-down term depends upon a protein's affinity for non-specific DNA and the DNA concentration, but not upon the rate at which it slides along DNA. The speed-up term $\bar{n} \sim \sqrt{D_{1D}\tau_{1D}}$, in contrast, depends on the absolute time spent in each round of sliding and the diffusion coefficient of sliding. Taken together the two effects can lead to speed-up (up to $\sim\bar{n}$ times) or slow-down as compared to the search by 3D diffusion alone. A similar observation that 1D/3D mechanism can lead to a slow search was made by Hu *et al* [9].

2.2. Single-molecule experiments

Several lines of experimental evidence support the 1D/3D search mechanism. These include experiments which demonstrated that the rate of site-specific binding is significantly increased by lengthening non-specific DNA surrounding the site (*the antenna effect*) and numerous elegant biochemical experiments supporting the role of 1D sliding in the search process [20–25]. The most direct experimental approach is the real-time observation of individual proteins sliding along DNA. By mechanically stretching DNA molecules and monitoring the movement of individual, fluorescently labeled proteins along DNA, single-molecule approaches have successfully characterized facilitated diffusion of repair proteins Msh2–Msh6, Ogg1 and Rad51, [26–28] as well as transcription factors LacI [29] and p53 [30].

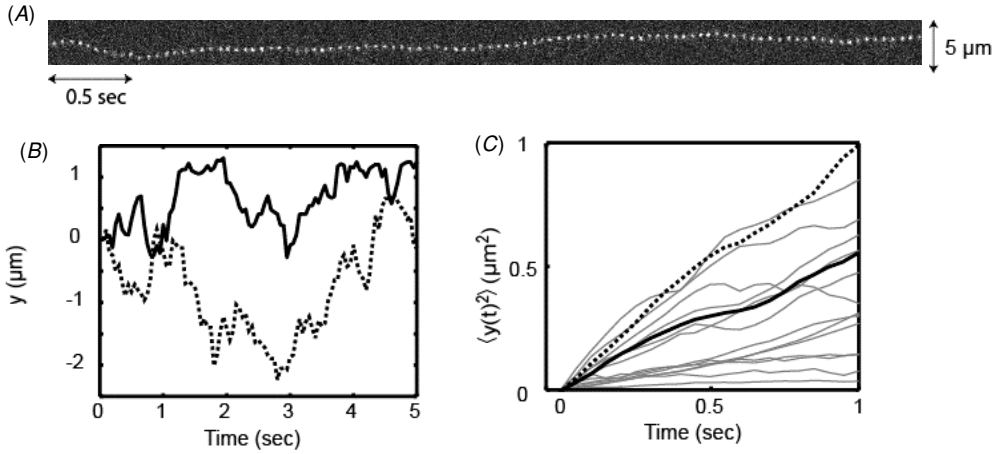

Figure 3. (*A*) The kymogram (a series of microscopy images) of protein sliding. A long DNA molecule is stretched (not visible) along the *y*-axis. The figure presents a series of images taken at high speed of a single labeled p53 molecule moving along DNA. (*B*) Trajectories of sliding shown for two proteins, as extracted from a series of microscopy images. (*C*) Mean-squared displacement of the proteins as a function of time, shown for two trajectories from (*B*) and 13 other trajectories. Note that the mean-squared displacement is linear, demonstrating non-anomalous diffusion. The diffusion coefficient is extracted from the slope of these lines.

In vivo imaging allowed the monitoring of 1D/3D diffusion of LacI in single bacterial cells [21]. While experimental evidence of the 1D/3D mechanism is overwhelming, detailed quantitative characterization of this process so far has been limited to very few proteins.

Single-molecule characterizations of DNA-binding proteins have given three interesting results. First, they demonstrated that protein sliding along DNA proceeds by normal (non-anomalous) diffusion. Figure 3 presents results of single-molecule tracking on a stretched DNA (van Oijen lab [30]). Obtained diffusion coefficients for 1D sliding of transcription factors, such as LacI and p53 [29, 30] have fallen in the range of

$$D_{1D} \approx 10^5\text{–}10^6 \text{ bp}^2 \text{ s}^{-1} = 10^{-10}\text{–}10^{-9} \text{ cm}^2 \text{ s}^{-1},$$

which is within an order of magnitude or two of the limit of $\approx 10^7$ bp^2 s^{-1} calculated from the Stokes–Einstein equations with a correction for the helical rotation of the protein as it tracks the pitch of the DNA. The factor of $\sim$10 difference suggests that the protein has to overcome relatively small free energy barriers $\sim$1–2 k_BT as it slides along DNA. Experimentally obtained diffusion coefficients suggest that a protein can slide over a range as long as $\bar{n} \approx$ 100–5000 bps if it stays bound to non-specific DNA for $\tau_{1D} \sim$ 0.01–5 s, as was measured by *in vivo* FRAP experiments for a range of eukaryotic DNA-binding proteins [31].

The second result concerns the mechanism of sliding. It was suggested that proteins can hop along DNA by dissociating and re-associating close by, movement that, at the resolution of the relevant single-molecule studies, resembles sliding of a protein along DNA. To test this mechanism we studied sliding of p53 along DNA at different salt concentrations [30]. Since monovalent salt ions bind DNA electrostatically, they effectively destabilize protein–DNA complexes. If hopping were the mechanism of translocation, faster 1D diffusion would be expected at high salt concentrations. We observed, however, that KCl had no effect on the diffusion coefficient, while leading to a decrease in the residence time on DNA with high salt

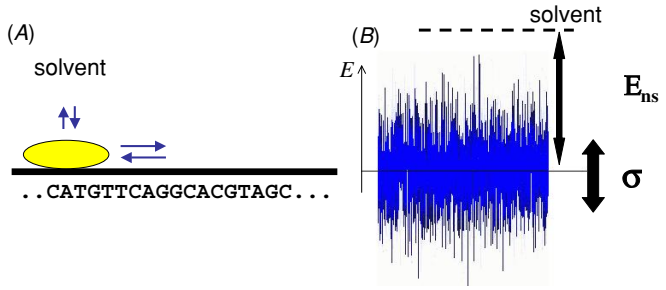

Figure 4. (*A*) Model of the 1D sliding: a protein can slide along DNA or dissociate. (*B*) The sequence-dependent energy landscape of sliding. Protein–DNA binding energy depends on sequence, thus forming the landscape. Two parameters determine protein dynamics: the roughness of the energy landscape σ and the (free) energy of non-specific binding to DNA, $E_{\rm ns}$.

concentration, as expected. These experiments demonstrated that p53 slides along DNA while maintaining constant contact with the DNA molecule [30].

The third surprising result is that sliding is observed not only for proteins from bacteria or mitochondria (e.g. bacterial LacI and mitochondrial Ogg1 [29, 26]) where DNA is mostly naked, but also for a eukaryotic protein p53 [23, 30]. In eukaryotes, DNA is tightly packed and bound by myriad other proteins, leaving little space for proteins to slide. Observed sliding by a eukaryotic proteins suggests that the mechanism of facilitated diffusion can occur in eukaryotes despite other DNA-bound proteins and DNA packing. In section 6.4, we present our previously unpublished analysis of sliding on crowded DNA, i.e. in the presence of other DNA-bound proteins.

3. Role of DNA sequence and protein conformation

3.1. Diffusion on a rugged landscape

Our group [17, 32] and others [10] have considered the role of DNA sequence in the process of facilitated diffusion. Since the energy of protein binding depends on the DNA sequence of the bound fragment, different sequences will have different energies. Thus, while sliding along DNA, the protein has different energies at different positions along the DNA, turning sliding into 1D diffusion in an external coordinate-dependent field (figure 4(*B*)). Let us consider the sequence-dependent field as a random field with energies independently and normally distributed. We choose the normal distribution as it closely resembles the distribution of the protein–DNA binding energies computed using a popular position-weight matrix approximation [33]. The approximation assumes that bound DNA base-pairs contribute independently and additively to the total binding energy, making the distribution of energies for random sequences normal due to the central limit theorem. By averaging the mean-first-passage time for a 1D random walk over the normally distributed energies, we obtained

$$D_{\rm 1D} \sim \exp\left[-\gamma\left(\frac{\sigma}{k_B T}\right)^2\right], \tag{6}$$

where $\gamma \sim 1$, and σ^2 is the variance of the protein–DNA binding energy. Such rapid decay of the diffusion coefficient with the ruggedness of the energy landscape clearly demonstrates that sliding fast enough to account for experimental evidence is possible only for $\sigma \lesssim 1$–$2\ k_B T$. Subsequent single-molecule experiments have confirmed this result (see section 2.2).

Hu and Shklovskii [10] have recently re-examined the effect of disorder on sliding and suggested two possible regimes of dynamics: macroscopic and mesoscopic ones. The macroscopic regime is characterized by sliding over sufficiently long distances, leading to averaging over specific realizations of the disorder (i.e. specific DNA sequence) and is identical to the regime discussed above with diffusion coefficient given by equation (6). The mesoscopic regime, in contrast, requires short sliding distances. If the sliding distance is sufficiently short, transition over a single highest barrier is the rate limiting step that leads to the diffusion coefficient $D_{1D} \sim \exp(-\sqrt{n}\sigma/k_BT)$ with less dramatic dependence on σ/k_BT. The suggested acceleration of search due to rare barriers in the mesoscopic regime, however, requires sufficiently small energy of non-specific binding: $|E_{ns}| \lesssim \sigma^2$ (in k_BT units). This regime however is unlikely to produce significant acceleration of sliding for the majority of DNA-binding proteins due to their high affinity for non-specific DNA: $E_{ns} \approx 10$–$15k_BT$ [6]. Thus the mesoscopic regime becomes relevant only for $\sigma \gtrsim 4k_BT$, when sliding is prohibitively slow. Our conclusion that fast search on DNA requires smooth sliding landscape ($\sigma \sim 1$–$2k_BT$) should hold for most DNA-binding proteins. Consistent with this result most well-characterized proteins have been shown to slide sufficiently fast with $\sigma \approx 1$–$2k_BT$ [21, 26, 29, 30, 34]. It is possible that some DNA-binding proteins that exhibit short sliding distance and low affinity for nonspecific DNA can be described by the mesoscopic regime and can afford to have more rugged landscape.

3.2. The search-stability paradox

We demonstrated that a relatively small variance of the sequence-dependent energy landscape $\sigma \lesssim 1$–$2k_BT$ can lead to fast sliding and overall fast searching, assuming 1D/3D partitioning is at its optimal value $\tau_{3D} = \tau_{1D}$.

The protein must find its target site sufficiently fast and then stay bound to it. We demonstrated that the requirements of fast search and stability of the protein–DNA complex impose different and mutually exclusive constraints on σ (see figure 5). Indeed, the variance of the sequence-dependent binding energy σ determines not only the rate of sliding, but also the energy of the target site E_0, and hence the equilibrium occupancy of the target site P_{eq}:

$$P_{eq} = \frac{\exp(-E_0/k_BT)}{\sum_{i=1}^{M}\exp(-E_i/k_BT)},$$

where energies E_i of individual sites are drawn from a normal distribution with the variance σ^2 and the target site has the lowest energy in the genome $E_0 = \min_{i=1,\ldots,M}\{E_i\} \approx -\sigma\sqrt{2\log M}$ ($M \approx 10^6$ bp for bacterial genomes). It is easy to see numerically that $P_b \gtrsim 0.25$ requires $\sigma \gtrsim 5k_BT$.

In summary, fast searching requires $\sigma \lesssim 1$–$2k_BT$, while stability requires $\sigma \gtrsim 5k_BT$. The two conditions are mutually exclusive and lead to the *speed-stability paradox*.

The paradox is analogous to that in protein folding [35, 36]. Analysis and folding simulations for a random protein demonstrated that at high T the protein is able to fold by overcoming energy barriers and escaping local minima, but the folded (ground) state is unstable. At low T, the folded state is stable, but folding becomes prohibitively slow. The speed-stability paradox in protein folding is resolved by designing sequences that have a pronounced energy gap between the native folded conformation and the bulk of unfolded ones [35, 36].

We demonstrated that a similar approach to the speed-stability paradox cannot work for protein–DNA interactions. The energy gap between the target site and all other (random) sites is $\approx\sigma\sqrt{4L}$, where L is the length of the site (i.e. number of strongly interacting base-pairs),

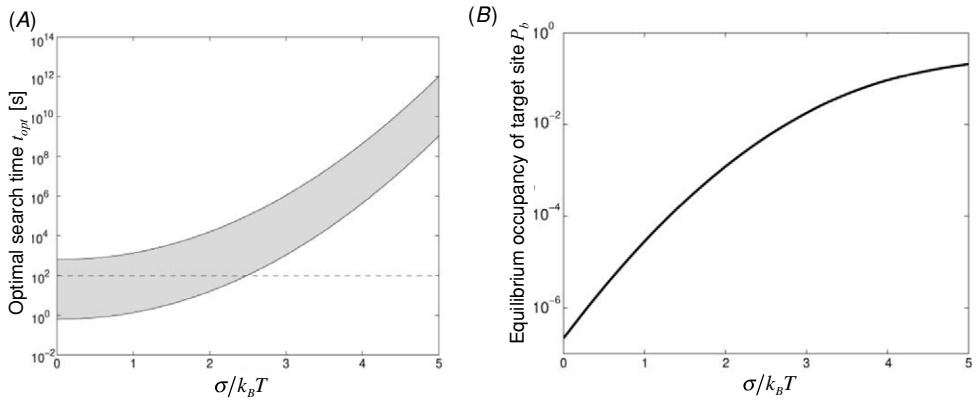

Figure 5. The speed-stability paradox. (*A*) The optimal search time for a single protein and a single target site on the entire bacterial DNA. The lane corresponds to possible values for the search time depending on parameters of the model and assuming optimal 1D/3D partitioning. Fast searching is possible only if $\sigma < 2k_BT$. (*B*) The equilibrium occupancy of the target site that has the lowest possible energy among $M = 5 \times 10^6$ sites. High equilibrium occupancy (i.e. stability of the protein–DNA complex) requires $\sigma \gtrsim 5k_BT$. It is impossible to achieve both fast searching and stability if the classical model of sequence-dependent protein–DNA interactions applies.

and 4 comes from the total number of possible base-pairs. Due to relatively small $L \approx 5$–10, large energy gaps are unattainable.

We proposed that the presence of (at least) two distinct conformational states in the protein (or protein–DNA complex) can resolve the paradox, allowing the protein both to slide rapidly and to form a stable complex with the target site.

4. Coupling of search and the conformational transition

4.1. The two-state model: effective energy gap

Here we consider a search by a protein (generally, a protein–DNA complex) that has two conformations: *the recognition (R)* state with $\sigma > 5k_BT$ and *the search (S)* state that has $\sigma \approx 1k_BT$. In the *S* state, the protein can slide fast along DNA, while in the *R* state it cannot slide far along DNA, but binds the target site tightly. Two additional parameters fully determine the dynamics of the system: the difference between the mean energy in the two states ΔG_{RS} and the rate of transitions between the two states (or the barrier that separates them). We introduced the model in [17] and further studied in [37].

While simulations were used to study the full model, analytical treatment of a simplified model described regimes of dynamics with sufficient accuracy and helps to intuitively understand the system. For this purpose, we assumed constant rates of the conformational transitions between the states: $k_{S\to R}$ and $k_{R\to S}$, and the equilibrium constant $K_{\mathrm{eq}} = k_{S\to R}/k_{R\to S} = \exp(-\Delta G_{RS})$ (see figure 6). We demonstrated that in this case the average search time can be calculated explicitly. A similar result was independently obtained by Hu *et al* [11].

Similar to [11], we found that the presence of the fast search state *S* has two opposite effects on search: it helps to accelerate search by fast 1D sliding in the *S* state, but it can make search prohibitively slow if the target site is visited only in the *S* state and not recognized in the

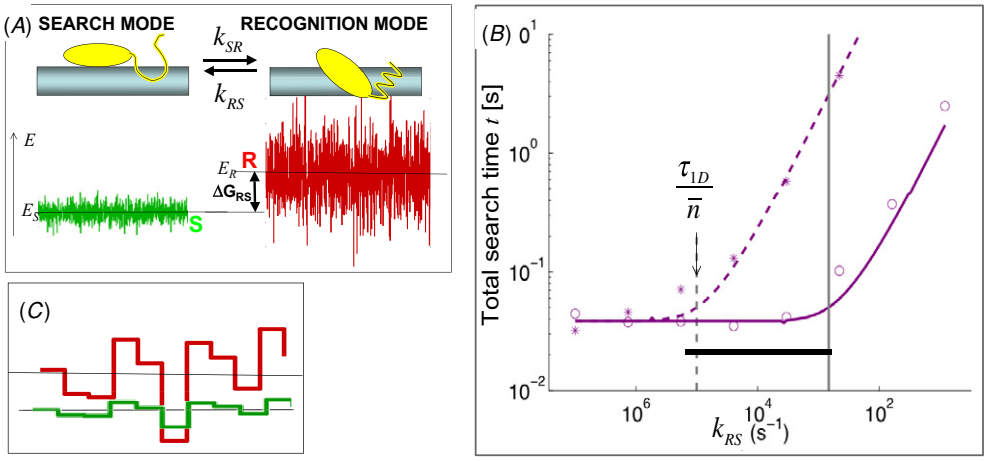

Figure 6. (*A*) The two-state model. The search state has a small ruggedness $\sigma < 2k_BT$ allowing the protein to slide rapidly. The recognition state has $\sigma > 5k_BT$ required for tight binding. Higher mean energy of the recognition state $E_R > E_S$ is required for the protein to spend most of the time sliding and occasionally sample rare sites that have $E_R(i) < E_S(i)$. (*B*) Results of the search simulations for the uncorrelated two-state model (dashed line) and correlated model of *kinetic pre-selection* (solid line). The search time is shown as a function of the rate of $R \to S$ transition (the *X*-axis is reversed: from high to low). A sufficiently high rate of the transition is required for fast searching. Kinetic pre-selection allows fast searching even for sufficiently slow transition rates ($\sim 10^3$ s^{-1}). (*C*) A diagram illustrating the correlated two-state model required for kinetic pre-selection.

R state. The more time the protein spends in the *S* state, the more significant the acceleration is, but the more likely the protein misses the target site while sliding pass it in the *S* state.

Assuming that the protein cannot slide in the *R* state, only the fraction of time spent in the *S* state $\left(1/\left(1+K_{\rm eq}^{-1}\right)\right)$ contributes to sliding. The second effect can be characterized by the probability P_f of recognizing (not missing) the site upon sliding in its vicinity. The total search time needs to be multiplied by the average number of times the global search need to be repeated until recognition, i.e. $1/P_f$. Taking both effects into account we obtain

$$\bar{t}_s = \frac{M}{\bar{n}}\frac{1}{P_f}\left(\tau_{3D}+\tau_{1D}\right), \qquad \bar{n} = \sqrt{\frac{4D_{1D}\tau_{1D}}{1+K_{\rm eq}^{-1}}} \tag{7}$$

This probability P_f determines the efficiency of search: for $P_f \approx 1$ the target is recognized (i.e. visited in the *R* state) in the first search round that visits the site, while for $P_f \ll 1$ the protein is unlikely to recognize the site upon the first arrival and many more rounds of search are required.

The probability P_f, in turn, depends on the total time the protein spends on the target site while sliding in its vicinity $\sim\tau_{1D}/\bar{n}$ and the transition rate $k_{S\to R}$. A slow transition ($k_{S\to R}\tau_{1D}/\bar{n} \lesssim 1$) leads to low probability of recognition P_f and significant slow-down of the search process. This effect is evident from figure 6 which presents results of simulations as compared to theoretical predictions of equation (7). In agreement with the theory, simulations demonstrate a transition from fast to slow search at $k_{S\to R}\tau_{1D}/\bar{n} \approx 1$.

The two-state model allows for the reconciliation of stability with fast search, but requires the protein to undergo a sufficiently fast conformational transition (faster than $\simeq \bar{n}/\tau_{1D} \simeq 3\cdot(10^3\text{–}10^4)$ s^{-1}). Several DNA-binding proteins have been shown to undergo

conformational transitions upon binding to the target site [38–41] with rapid transition rates comparable to the estimate given above [42].

In addition to a sufficiently fast conformational transition, the two-state mechanism requires that the specific binding conformation of the protein–DNA complex (state *R*), which is favorable when the protein is on the target site, be generally unfavorable elsewhere—that is, the mean energy in the *R* state must be much higher than that of the *S* state (figure 6). This entails that most of the time the protein spends in the *S* state, which has little sequence-dependent variance $\sigma_S \approx 1k_BT$, and is thus sliding fast. This effectively creates a landscape with small ruggedness and a few low-energy sites with $E_R(i) < E_S(i)$. Such landscape was postulated by Gerland *et al* [43] who demonstrated that it provides fast search and satisfied thermodynamic requirements of tight and 'programmable' binding.

Such a landscape picture is also consistent with recent microfluidic measurements of the binding energy for the yeast Cbf1 protein to all possible short DNA sequences [44]. The experiment demonstrated that the binding energy has little sequence dependence for sites differing significantly from the target site. The requirement of the model that the *R* state has an average energy above that of the *S* state, is consistent with significant deformation of DNA by proteins bound to their target sites. In agreement with our model, the structure of Lac repressor bound to non-specific DNA (i.e. in the *S* state) [39] shows no DNA deformation and very few contacts between the protein and the nucleobases, suggesting small sequence-specific contribution to the energy, and hence small σ in the *S* state.

4.2. Kinetic pre-selection

We demonstrated [37] that the process of search by a two-state protein can be much more efficient if the protein can undergo the transition most readily at the target site, entailing that P_f be close to 1 while keeping $\bar{n}$ sufficiently large in equation (7). We proposed the *kinetic pre-selection* mechanism which is capable of achieving this search-expediting behavior.

The central idea of kinetic preselection is that the *S* and *R* landscapes are correlated (see figure 6(*C*)). Then for each site *i*, $E_S(i) = E_R(i)\sigma_S/\sigma_R - \Delta G_{RS}$ ($\Delta G_{RS} > 0$). If the landscapes are correlated, a deep minimum in the *R* state corresponds to a local minimum in the *S* state, entailing that the protein stay longer at the site while sliding, and hence increasing the probability of the conformational transition occurring at such site. This leads to an effective kinetic pre-selection of low-energy sites, making *S*-to-*R* transitions on such sites more likely. Simulations demonstrate that kinetic pre-selection leads to fast searching even when the conformational transition is as slow as experimentally observed [42] (figure 6(*C*)).

4.3. The landscape model

We also examined [37] an interesting generalization of the two-state model is a model system with a continuum of conformational states, i.e., a reaction coordinate z for the conformational transition (see figure 7). We introduced this simple system based on the following assumptions: (i) there is a minimum in the conformational energy at some value z_0: $E(z) \sim \alpha(z - z_0)^2$ and (ii) the sequence-specific energy decays exponentially with z: $E(x, z) \sim E(x)\exp(-z)$. These lead to the energy function along DNA (x coordinate) and the conformational coordinate z:

$$E(x, z) = E(x)\exp(-z) + \frac{\alpha}{2}(z - z_0)^2,$$

where $E(x)$ is the sequence-specific binding energy (that is equivalent to the energy in the *R* state of the two-state model), and α and z_0 are parameters which determine the shape of

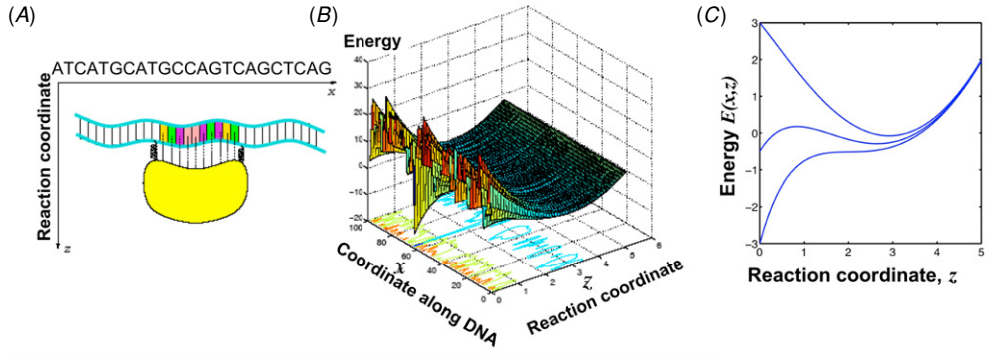

Figure 7. (*A*) The landscape generalization of the two-state model with z as a reaction coordinate of the conformational transition and the sequence-specific energy decaying exponentially with z. (*B*) The resulting 2D landscape of sliding and conformational transition. With appropriately chosen parameters, the ruggedness in the sliding valley ($z = z_0$) is small, allowing fast sliding and occasional transitions into the *R* state ($z = 0$). The search time as a function of the conformational mobility (along the z-axis) is similar to figure 6(*B*). (*C*) The cross-section of the landscape $E(x, z)$ versus reaction coordinate z of the conformational transition. The profiles of energy for three different sites are shown: the target site (the lower curve), a low-energy site (the medium curve) and a high-energy site (the top curve). Note that the barrier for transition along the z-coordinate depends on the energy of the site: the target site has no barrier leading to very fast transitions and likely recognition from the first arrival $P_f \approx 1$ in equation (7).

the landscape. The landscape defined this way, illustrated in figure 7(*B*), has a 'groove' at z_0 which corresponds to the *S* state of the two-state model. Importantly, the two profiles along the x coordinate in the two states are correlated naturally leading to the kinetic pre-selection discussed above.

For a proper choice of parameters, the search resembles that of the correlated two-state model with a noticeable difference: the barrier for the conformational transition is lower at the low-energy (target) sites, making the transition faster and more likely to happen at such sites. Figure 7(*C*) illustrates this observation by showing $E(x_0, z)$ versus z profiles for three sites with different energy at $z = 0$ ($E(x) = E_0$, $E(x) > 0$, $E(x) \approx 0$). Note that the target size (lower line) can be reached from the z_0 state (where sliding is fast) without a barrier along the z reaction coordinate. In this model, the search time depends on the effective diffusion coefficient along the reaction coordinate, requiring sufficiently high conformational mobility to achieve fast searching (similar to the two-state model figure 6(*C*)).

Both the kinetic pre-selection and the landscape models lead to effective coupling of the conformational transition and binding, i.e., a conformational transition in the protein–DNA complex when the protein binds to the target site. Such coupling of folding and binding has indeed been detected for a broad range of DNA-binding and other ligand-binding proteins, but the role of this phenomenon in recognition remained unknown. Our models suggest that fast searching for a target site among $\sim 10^6$–10^9 decoys requires the protein to have at least two distinct conformations, rapidly exchange between them and exhibit coupling between the conformational transition and binding to the target. It is possible that these conclusions can be generalized to other search processes where the target is encoded as a short string to be detected among a combinatorially large number of possible decoy strings.

5. Non-specific binding: 1D/3D partitioning

5.1. Non-optimal 1D/3D partitioning in bacteria

Our analysis demonstrated that the fastest search is achieved if $\tau_{1D} = \tau_{3D}$. Recently, we studied whether this condition is satisfied in a bacterial cell. We showed that the partitioning of time between 1D and 3D depends on proteins' affinity for non-specific DNA, measured by the binding constant $K_d^{ns} = [P][DNA]/[P \cdot DNA]$, and upon the intracellular DNA concentration. High affinity for non-specific DNA measured for a range of bacterial DNA-binding proteins $K_d^{ns} \approx 10^{-3}$–10^{-6} M [6] and the significant DNA concentration inside a bacterial cell [DNA] $\approx 10^{-2}$ M (assuming 5×10^6 bps confined in micron-sized nucleoid) entail that the protein spends most of the time on DNA:

$$\frac{\tau_{1D}}{\tau_{3D}} = \frac{[DNA]}{K_d^{ns}} = 10^1\text{–}10^4. \tag{8}$$

This theoretical estimate for the fraction of time on DNA: $\tau_{1D}/(\tau_{1D} + \tau_{3D}) \gtrsim 90\%$ is close to recent *in vivo* measurements which estimate this fraction as 87% [21].

5.2. Speed-up and slow-down

Such strong deviation from optimal partitioning has significant implications for the search process. Substituting this high value of τ_{1D}/τ_{3D} into equation (5) demonstrates that binding to non-specific DNA leads to significant slow-down of the search process by a factor of $\tau_{3D}/(\tau_{3D} + \tau_{1D}) = 10^{-1}$–$10^{-4}$. This slow-down can be barely compensated by the speed-up of 1D sliding $\bar{n} = 10^2$–10^3 term in equation (5), leading to the search time $t_s \approx 15$–100 min for ten proteins per cell.

This analysis brings us to the surprising conclusion that the 1D/3D mechanism does not allow for significant facilitation of the protein–DNA search in bacteria; the speed-up due to sliding is eliminated by the slow-down due to non-specific binding to DNA. Can proteins rapidly find their sites in spite of this slow-down? What is the role of this whole mechanism if it does not provide effective facilitation? A possible answer to the first question is provided below. A possible answer to the second is that high affinity for non-specific DNA could be essential for other mechanisms such as equilibrium sequestration of the protein from the cognate site to non-specific DNA (see [43] and appendix of [18] for more details). Thus sliding provides a way of facilitating search under conditions of high DNA concentration and high affinity for non-specific DNA, that otherwise would lead to a very slow search. It is possible that certain proteins can have low affinity for non-specific DNA and thus would not benefit from sliding, but can instead accelerate search by large number of copies of this protein in cell.

6. Spatial effects, DNA conformation and crowding

The process of protein–DNA search is influenced by a broad range of spatial effects. This includes, among other phenomena, the role of DNA conformation [9] and density [12], hopping and intersegmental transfer [45], and the dependence of the search time on the initial distance [18, 19].

Hu *et al* [9] have conducted the first systematic study that examined the role of DNA conformation in the search process. Using a continuous diffusion approximation, an approach complementary to the single-protein view discussed above, the authors examined the search process and established an elegant analogy to a problem in electrostatics. They identified a

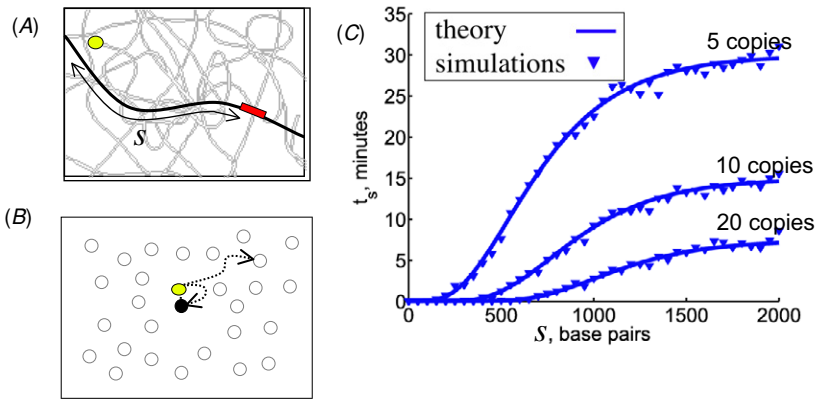

Figure 8. (*A*) A mechanism of fast searching by a protein that starts a distance s away from the target site but close to DNA. (*B*) A simple model that takes into account such spatial effects as (i) re-associations to the same strand of DNA ('hops') and (ii) associations with other remote strands of DNA ('jumps'). For lengths shorter than persistence length of DNA, DNA can be considered as a straight rod, and the problem of jumps versus hops can be considered as a 2D problem in the orthogonal cross-section, as shown in the diagram. (*C*) The search time as a function of the distance between the initial position of the protein along DNA and the target site: theoretical lines and simulation datapoints.

wealth of different scaling regimes, found the maximal possible acceleration of search due to sliding and demonstrated that high DNA concentration in combination with excessive affinity of proteins for non-specific DNA can lead to a slow-down rather than an acceleration of the search process. This study also showed that various DNA conformations (e.g. a Gaussian coil) could promote correlated re-associations (reabsorption) that extends effective antenna lengths beyond those of individual sliding events. While providing a thorough examination of scaling behaviors in possible search regimes, this study left several questions open, such as the possibility of distance-dependent search and the microscopic mechanism of re-associations (hops) during a single round of sliding.

6.1. Distance-dependent search time

As Polya purportedly told the drunkard wandering the streets looking for his home, 'You can't miss; just keep walking, and stay out of 3D!' In 3D, diffusion is non-redundant, i.e. the probability of revisiting a particular site is much less than one [46]. As a consequence of this property, the average time required to find a particular site does not depend on the initial position, as long as this distance is greater than the size of the target. In contrast, the time of search in two dimensions (2D) (e.g. on a membrane) or in 1D (e.g. along DNA or along a filament) is distance dependent [46]. Therefore, we ask: can the 1D component of facilitated diffusion make search much faster for proteins starting a small distance from the target site?

Since sliding increases the effective target size by a factor of $\bar{n}$, we expect fast and distance-dependent search if the protein starts close to DNA and $\sim\bar{n}$ bps away from the target (see figure 8(*A*)). In other words, if a protein can find its site by sliding along DNA and not dissociate (i.e. 'staying out of 3D'), in what we call a local search, the search time will be dependent on its initial position [19, 18]. The length scale of this effect can be further increased by re-associations of the protein to DNA, replacing $\bar{n}$ with some $\bar{n}_{\text{eff}} > \bar{n}$.

6.2. Re-associations

Models discussed above assume that after a protein dissociates from the DNA, it re-associates with *uniform* probability along DNA, i.e. all sites on DNA are equally likely to be the association site. Scaling arguments of Hu *et al* [9] support this uniform distribution of landing points for DNA organized into a random globule. Other DNA conformations, however, can increase the probability of associating near the dissociation point [9].

As a first approximation, we assume that spatial excursions can be of two extreme types: *hops*, short-range dissociations from the DNA in which the protein re-associates in the same region of DNA at a distance smaller than or equal to its persistence length (150 bp), and *jumps*, long-range excursions in which each site of DNA is equally likely to be the re-association point (see figure 8(*B*)). Note that associating far (along the contour length of DNA) from the dissociation point does not necessarily require excursions far in 3D space: jumping onto a strand that is nearby spatially is likely to achieve this, especially if the DNA is packed into a random globule (see [9] and simulations below). Thus jumps may also include 'inter-segmental transfer' of the protein from one DNA region to another—a mechanism that is frequently discussed in the literature but is rather poorly defined microscopically. To calculate the total search time we coarse-grain hops into a longer sliding distance.

We used simulations and analytical approaches to calculate the probability of a hop (versus a jump) upon dissociation. Since at a scale smaller than the persistence length the DNA can be considered as a straight rod, the problem becomes a 2D problem of return to the origin (a hop) versus being adsorbed by the traps that represent other DNA fragments (see figure 8(*B*)). We showed that at intra-bacterial DNA density the probability of a hop $P_{\rm hop} \approx 80$–90% entails that a protein hop for $1/(1 - P_{\rm hop}) \approx 6$–$9$ times before it dissociates to make a jump. This makes the effective sliding distance $\bar{n}_{\rm eff} = \bar{n}/\sqrt{1 - P_{\rm hop}}$ and thus the span of fast local search about 2.5–3 times longer.

Upon a jump, the protein is assumed to re-associate uniformly on DNA, allowing us to use equation (2) with $\tau_{\rm 1D}$ and $\bar{n}$ replaced by their effective counterparts that take into account hops:

$$t_s = \frac{M}{\bar{n}_{\rm eff}}(\tau_{\rm 1D,eff} + \tau_{\rm 3D}). \tag{9}$$

This equation constitutes a simple modification of equation (2) that allows us to account for spatial effects such as re-associations and DNA density, which both determine $P_{\rm hop}$.

Figure 8(*C*) presents simulation and analytical results for the search time as a function of the distance s between the target and the initial position of the protein along DNA. Clearly, the search is much faster at distance $s \lesssim \bar{n}_{\rm eff}$. As s increases the protein is less likely to find the target by sliding and hopping, and is likely to make a jump to a random location on DNA, which entails that it performs a global search. When the search is performed by multiple proteins in parallel (until any one of them binds the site), there is significant gain in search speed, allowing the starting distance to be much greater with $s \approx 1000$ bps.

This result has several important implications. First, it reconciles seemingly conflicting experimental data: a protein can make a small number of hops, as recently observed experimentally [22], but slide along DNA in between the hops, as evident from single-molecule experiments [25, 30]. Second, the mechanism of fast searching by a protein that starts close to its target can be relevant for protein–DNA search in bacteria. Indeed, proteins are produced close to the location of the gene encoding it [47]. Thus, if a gene encoding a DNA-binding protein is located close to the target site of this protein, then the search can be very fast. We demonstrated that in agreement with this conjecture, genes encoding bacterial transcription factors have a tendency to be located close to the targets of these factors [18].

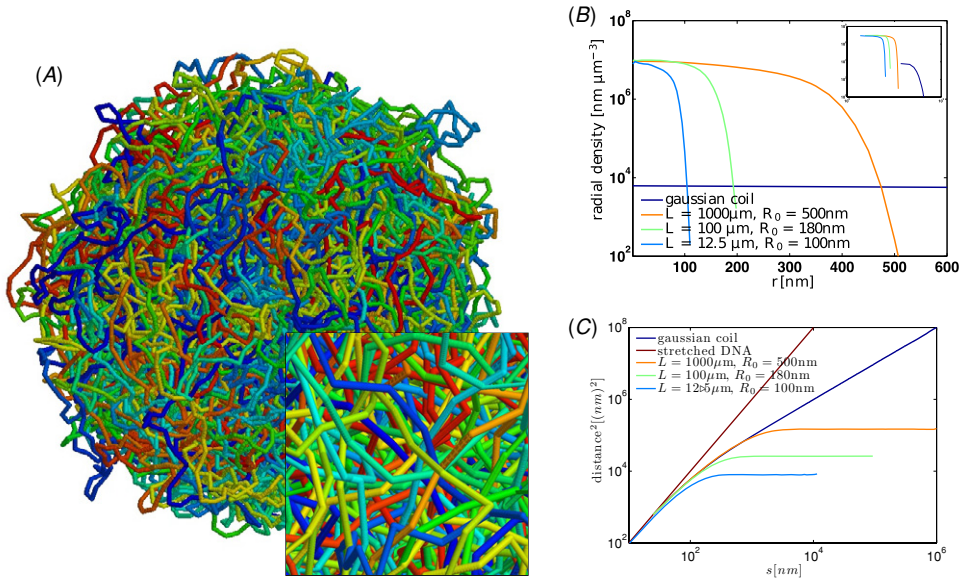

Figure 9. Compact DNA conformations obtained by Monte Carlo methods to simulate search inside dense DNA globule. (*A*) An example of the DNA globule obtained by simulating confinement of 3×10^6 bps (1000 μm) of DNA inside a spherical volume of $R_0 = 500$ nm. The DNA is modeled by 40 000 rods (two per persistence length) with the bending energy set to obtain the correct persistence length. The inset shows the high density of DNA inside the globule. (*B*) The radial DNA density of obtained globules, compared to that of the Gaussian coil (averaged over 1000 conformations). Note the nearly uniform density inside the globules is in agreement with the theory [50]. (*C*) Mean-squared distance between monomers with distance s apart along the chain for obtained globules, the Gaussian coil and straight DNA. As expected for an ideal globule [50], for $s < R_0$ the chain inside the globules behaves as the Gaussian coil, while for $s > R_0$ the mean-squared displacement is constant due to the confinement.

6.3. The role of DNA conformation

In the models presented above, we attempt to describe the search process more realistically by including hops. However, we still assume that jumps are completely randomizing. The compact conformation of DNA can make jumps non-uniform as well, e.g. making it more likely for a protein to re-associate with DNA a certain distance away from a dissociation point (but much further than a hop). To better test how DNA conformation influences the search process, we performed large-scale simulations of long DNA molecules confined in a small volume. Using these simulations we tested whether possible correlated re-associations can lead to significant deviations from the theory [48].

We modeled the DNA conformation inside a bacterial nucleoid as a semiflexible chain of 5×10^6 bps confined to a spherical volume 1 μm in diameter (see figure 9). The parameters and Monte Carlo dynamics of DNA were modeled following Vologodsky *et al* [49] with 2–5 rod segments per persistence length l_p. Figure 9 demonstrates that the statistical properties of the chain resemble those of an ideal polymer globule [50].

We used the subsequently obtained DNA conformations to simulate the search process. The off-lattice simulations of diffusion were performed by using an efficient method widely applied in simulations of diffusion-limited aggregation [46]. Results of these simulations are presented in figure 10.

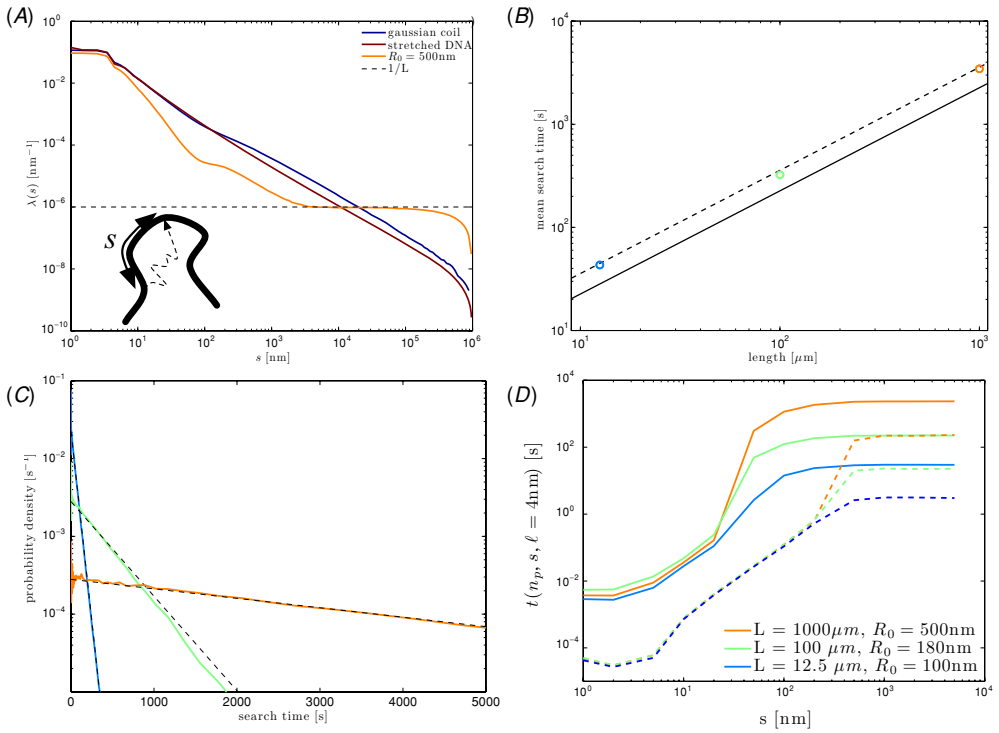

Figure 10. Search inside a dense DNA globule. (*A*) The probability $\lambda(s)$ of a jump between two points that are distance s apart along the contour length of DNA for the globule (orange), the Gaussian coil (blue) and stretched DNA (red). Note that for small s, the jumps in the globule are less likely than for straight DNA since associations to other nearby DNA strands are possible in the globule. For $s > 10^3$ nm, the jumps in the globule become uniformly distributed (close to $1/L$, dashed line), as assumed by the theory. The two regimes are well captured by the modified theory (equation (9)). (*B*) The search time as a function of DNA length (at a constant DNA density). Simulations (circles) are in good agreement with the original theory (solid line, equation (2)) which assumes uniformly random jumps, but are in a much better agreement with the modified theory (dashed line) which considers re-associations (equation (9); for this density $P_{\text{hop}} = 0.6$ as obtained from simulations). No fitting parameters were used. (*C*) Exponential distributions of the search time (solid lines) are in perfect agreement with the theory (dashed lines, same cases and colors as in (*B*)). The mean search time was calculated using equation (9). (*D*) The search time as a function of the distance s along DNA between the target site and the initial position of the protein that was also displaced 4 nm away from the DNA. The solid line is the search by one protein, the dashed line is for 10 proteins searching simultaneously. Note that for 10 proteins the search is about 10 times faster and the region of fast searching extends up to $\sim$ 500 nm (1500 bps).

First, we examined the statistics of the jumps, i.e., the distribution of the DNA contour length s between the dissociation and association points. Two regimes are observed. For small $s \lesssim l_p < R_0$, the probability of re-association (hop) distance s away decays more rapidly than that for straight DNA or a Gaussian coil, likely due to capture of the protein by a remote DNA strand that is spatially close in a dense DNA globule. For $s > R_0$, the probability distribution approaches that of the uniform one. A broad cross-over region with correlated, non-uniform re-associations is observed for the intermediate s.

Second, we tested whether our modified theory that takes into account hops can predict the search time in the DNA globule. Results presented in figure 10(*B*) clearly demonstrate

that the mean search time can be accurately predicted by the theory (equation (9)). Moreover, the complicated shape of the jump statistics (figure 10(*A*)) does not lead to any deviations from the linear dependence of the mean search time on DNA length: $t_s \sim M$, or deviations of the first-passage time distribution from the exponential distribution (figure 10(*C*)). Thus, correlated re-associations in the globule do not contribute significantly to the search dynamics.

Third, we studied distance dependence of the search time by initiating the search some distance s along the DNA from the target site, and 4 nm away from the DNA. We found (figure 10(*D*)) that at distances $\sim\bar{n}$ the search is extremely fast, which is in agreement with our observations using a simpler model that assumed short-range hops and long-range uniform jumps (see above). The region of fast searching is extended considerably by re-associations if several proteins are searching simultaneously.

Little is known about DNA organization inside the cell. Recent experimental studies provide clues about statistical properties of such organizations [51], allowing studies of protein–DNA search in complicated DNA arrangements to be undertaken.

6.4. Search on crowded DNA

Most theoretical studies assumed DNA to be fully accessible for protein binding. In a cell, however, DNA is bound by numerous proteins (e.g. histones) that makes some regions partially inaccessible to other proteins. Experimental efforts aimed at mapping nucleosomes [52, 53] and detection of accessible DNA [54], suggest that only about 5–10% of eukaryotic DNA is accessible. Here we summarize our recent study aimed to understand how the presence of other DNA-bound proteins can affect a DNA-binding proteins search process. The details of this analysis will be published elsewhere.

We considered other DNA-binding proteins and protein complexes such as nucleosomes as immobile obstacles. Specifically we assumed that (i) bound proteins occlude binding to the same region of DNA by the searching protein; (ii) obstacles stay at the same location for times longer than the search time (see figure 11); (iii) a searching protein cannot slide by other DNA-bound proteins; (iv) the rate of sliding and the energy of non-specific binding to regions not bound by other proteins remain unaffected (see figure 11). The first two assumptions are consistent with several experimental measurements of nucleosome stability and turnover rates *in vitro* and *in vivo* [53, 52, 55]. It remains to be examined experimentally whether sliding of one protein is obstructed by another DNA-bound protein or a nucleosome. We also assume that the target site is not obstructed by other DNA-bound proteins and remains accessible.

DNA-bound obstacles have two opposite effects on the search process: they affect the search far from the target site (binding to non-specific DNA) and search in immediate vicinity of the target.

The first effect is in sequestration of some fraction of nonspecific DNA. DNA occupied by other proteins reduces the total length of accessible DNA, which can be taken into account by replacing M with φM in equations (2). This also leads to increased duration of 3D diffusion (i.e. search for non-specific DNA) from τ_{3D} to τ_{3D}/φ. Note that according to our assumptions, τ_{1D} is not changed by other DNA-bound proteins: they affect how far a protein can slide (serving as reflective boundaries), but not for how long it stays on DNA. Taken together this transforms original equation (2) into

$$\bar{t}_s^{\text{crowded}} = \frac{\varphi M}{\lambda}(\tau_{3D}/\varphi + \tau_{1D}), \tag{10}$$

where λ is the new effective antenna length and τ_{3D} is the mean duration of the spatial diffusion computed based on the total concentration of DNA. For biologically relevant

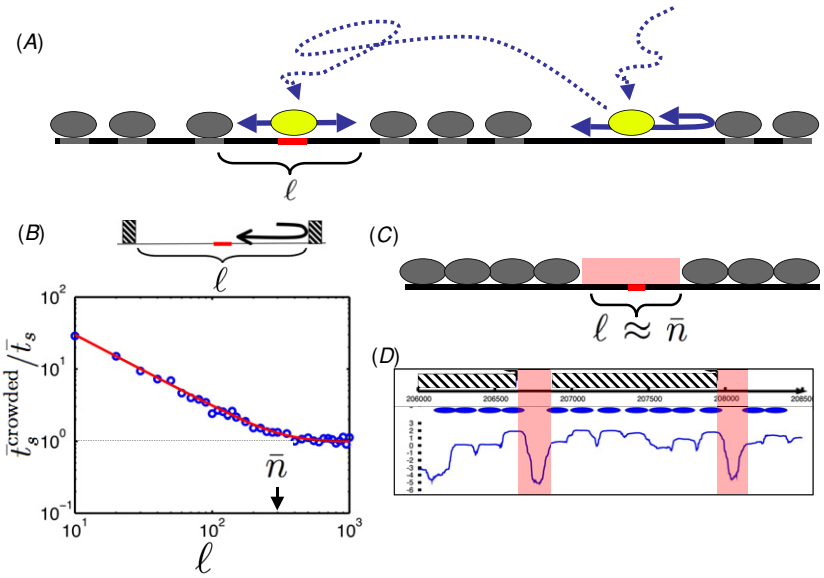

Figure 11. Search on crowded DNA. (*A*) Schematic presentation of the search on DNA which has other proteins/nucleosomes that occlude binding and sliding by the searching protein. (*B*) The increase in the search time as a function of the distance between DNA-bound obstacles ℓ symmetrically placed around the target site. The search time was computed using equations (10) and (11) (solid line) and by direct simulations (symbols). (*C*) Optimal organization of chromatin/nucleosomes on DNA that minimizes the search time (equations (10) and (11)) for proteins poised to bind regulatory regions (shown in red). (*D*) An example of experimentally observed nucleosome structure [52].

$\varphi \approx 0.1$–0.01 and assuming that the majority of time is spent on DNA $\tau_{1D} \gg \tau_{3D}$, the search can get ∼10–100 faster due to sequestration of nonspecific DNA.

The second effect, hampering of the sliding motion, is relevant only for the antenna region leading to its reduction from $\bar{n}$ to λ and slower search. If the distance ℓ between the hampering proteins surrounding the target site is less than $\bar{n}$ then the antenna is reduced to $\lambda \approx \ell$. If, in contrast, $\ell \gg \bar{n}$, then the motion on the antenna is not affected: $\lambda = \bar{n}$. By considering symmetrically bound obstacles as reflective boundaries (figure 11) we calculated the probability of sliding to the target before dissociation and obtained the expression

$$\lambda = \bar{n} \tanh(\ell/\bar{n}) \approx \min\{\ell, \bar{n}\} \tag{11}$$

which has all the right asymptotics discussed above. Figure 11(*B*) illustrates this effect as a function of ℓ (at $\varphi = \text{const}$) and shows a good agreement between this equation and the simulations of search in the presence of site surrounded by two impenetrable symmetrically located obstacles.

There are several interesting biological implications of these results. First, having a target site between two nucleosomes spaced at $\ell < \bar{n}$ make the search by a factor of $\bar{n}/\ell$ slower. This effect can be rather significant: ∼5–10 if we assume $\bar{n} \approx 250$–500 and $\ell \approx 50$, a typical spacing between nucleosomes.

Second, since DNA-bound obstacles that hamper kinetics have no effect on equilibrium, while increasing the search time, they also increase the residence time of the protein on its cognate site. Microscopically this works by increasing the probability that a protein which has slid away from its target will re-associate with the target before dissociating from the

DNA altogether, preventing the protein that left its target site to slide away and increasing its probability to re-associate with the target. The presence of DNA-bound proteins close to each other can thus increase the lifetime of the complex by such a 'jamming' mechanism.

Third, the fastest delivery of regulatory DNA-binding proteins (transcription factors) to their targets will be achieved if (i) $\ell \approx \bar{n} \approx$ 250–500 bps of DNA in the vicinity of each target are left unoccupied by DNA-binding proteins and nucleosomes, and (ii) the rest of the DNA gets packed and made inaccessible for binding by regulatory proteins. Figure 11(*C*) illustrates this optimal organization. Interestingly, recently obtained maps of nucleosomes in a variety of organisms are very consistent with this picture (figure 11(*D*)): about 90–99% of DNA is packed [52]and largely inaccessible [54], while regions of ≈300–500 bps in the vicinity of target sites remain available for binding and nucleosome-free [52, 56]. Equations (10) and (11) allow calculation of the speed-up due to such organization of DNA. The search is ∼10–100 times faster as compared to naked DNA (due to sequestration), and about ∼3–5 times faster as compared to random placement of the same number of nucleosomes on DNA, which would lead to ≈50–100 bps spacing.

7. Conclusions

Here we described our recent results that combine theory, simulations and single-molecule experiments. We have presented a simple theory that provides intuition for the mechanism of facilitated diffusion. This theory suggests the conformational transition in the protein–DNA complex (so-called coupling of folding and binding) plays a crucial role, which is consistent with a range of experimental observations. The 1D/3D nature of facilitated diffusion leads to extremely fast searching that can be initiated $\sim 10^2$–10^3 bps away, and thus, facilitated diffusion is critical for expediting the protein-DNA search in bacteria where the search would otherwise be rather slow. Spatial effects can be more accurately taken into account by minimal modifications to the theory. Simulations of the search in a compact DNA globule demonstrate that the theory provides precise answers for both the mean search time and the distribution of search times. Our formalism can be extended to take into account interactions between the searching protein and other DNA-bound proteins, suggesting a significant effect such interactions can have in the cell. Effects of molecular crowding in 3D and on DNA, complex DNA packing, and interactions with other proteins must be studied both theoretically and experimentally to provide a more complete picture of the search process. As it is central to many vital biological processes, the protein–DNA search problem will undoubtedly continue fascinating biophysicists.

Acknowledgments

We are grateful to Mehran Kardar and Antoine van Oijen for insightful discussions of these projects. We also thank the Aspen Center for Physics where parts of this work have been conceived and discussed with colleagues. LM was supported by the NIH-funded National Center for Biomedical Computing *i2b2*. ZW was supported by Howard Hughes Medical Institute Predoctoral Fellowship. JL was supported by NSF.

References

[1] Riggs A D, Bourgeois S and Cohn M 1970 The Lac repressor–operator interaction: 3. Kinetic studies *J. Mol. Biol.* **53** 401–17

[2] Fersht A 1999 *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (New York: Freeman)
[3] Richter P H and Eigen M 1974 Diffusion controlled reaction rates in spheroidal geometry. application to repressor–operator association and membrane bound enzymes *Biophys. Chem.* **2** 255–63
[4] Berg O G and Blomberg C 1976 Association kinetics with coupled diffusional flows. Special application to the lac repressor–operator system *Biophys. Chem.* **4** 367–81
[5] Berg O G, Winter R B and von Hippel P H 1981 Diffusion-driven mechanisms of protein translocation on nucleic acids: 1. Models and theory *Biochemistry* **20** 6929–48
[6] Revzin A 1990 *The Biology of Nonspecific DNA–Protein Interactions* (Boca Raton, FL: CRC Press)
[7] Halford S E and Marko J F 2004 How do site-specific DNA-binding proteins find their targets? *Nucl. Acids Res.* **32** 3040–52
[8] Coppey M, Benichou O, Voituriez R and Moreau M 2004 Kinetics of target site localization of a protein on DNA: a stochastic approach *Biophys. J.* **87** 1640–9
[9] Tao Hu, Grosberg A Y and Shklovskii B I 2006 How proteins search for their specific sites on DNA: the role of DNA conformation *Biophys. J.* **90** 2731–44
[10] Tao Hu and Shklovskii B I 2006 How does a protein search for the specific site on DNA: the role of disorder *Phys. Rev.* E **74** 021903
[11] Flyvbjerg H, Jülicher F, Ormos P and David F (ed) 2002 *Physics of bio-molecules and cells Les Houches* vol 75 (Heidelberg: Springer) chapter 1
[12] Lomholt M A, van den Broek B, Kalisch S-M J, Wuite G J L and Metzler R 2009 Facilitated diffusion with DNA coiling *Proc. Natl. Acad. Sci. USA* **106** 8204–8
[13] Lomholt M A, Ambjornsson T and Metzler R 2005 Optimal target search on a fast-folding polymer chain with volume exchange *Phys. Rev. Lett.* **95** 260603
[14] Zhou H-X and Szabo A 2004 Enhancement of association rates by nonspecific binding to DNA and cell membranes *Phys. Rev. Lett.* **93** 178101
[15] Loverdo C, Benichou O, Voituriez R, Biebricher A, Bonnet I and Desbiolles P 2009 Quantifying hopping and jumping in facilitated diffusion of DNA-binding proteins *Phys. Rev. Lett.* **102** 188101
[16] Bonnet I, Biebricher A, Porte P-L, Loverdo C, Benichou O, Voituriez R, Escude C, Wende W, Pingoud A and Desbiolles P 2008 Sliding and jumping of single Ecorv restriction enzymes on non-cognate DNA *Nucl. Acids Res.* **36** 4118–27
[17] Slutsky M and Mirny L A 2004 Kinetics of protein–DNA interaction: facilitated target location in sequence-dependent potential *Biophys. J.* **87** 4021–35
[18] Kolesov G, Wunderlich Z, Laikova O N, Gelfand M S and Mirny L A 2007 How gene order is influenced by the biophysics of transcription regulation *Proc. Natl. Acad. Sci. USA* **104** 13948–53
[19] Wunderlich Z and Mirny L A 2008 Spatial effects on the speed and reliability of protein–DNA search *Nucl. Acids Res.* **36** 3570–8
[20] Kim J G, Takeda Y, Matthews B W and Anderson W F 1987 Kinetic studies on Cro repressor–operator DNA interaction *J. Mol. Biol.* **196** 149–58
[21] Elf J, G-W Li and Xie X S 2007 Probing transcription factor dynamics at the single-molecule level in a living cell *Science* **316** 1191–4
[22] Gowers D M, Wilson G G and Halford S E 2005 Measurement of the contributions of 1d and 3d pathways to the translocation of a protein along DNA *Proc. Natl. Acad. Sci. USA* **102** 15883–8
[23] McKinney K, Mattia M, Gottifredi V and Prives C 2004 p53 linear diffusion along DNA requires its c terminus *Mol. Cell* **16** 413–24
[24] Winter R B, Berg O G and von Hippel P H 1981 Diffusion-driven mechanisms of protein translocation on nucleic acids: 3. The *Escherichia coli* lac repressor–operator interaction: kinetic measurements and conclusions *Biochemistry* **20** 6961–77
[25] Gorman J and Greene E C 2008 Visualizing one-dimensional diffusion of proteins along DNA *Nat. Struct. Mol. Biol.* **15** 768–74
[26] Blainey P C, van Oijen A M, Banerjee A, Verdine G L and Xie X S 2006 A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA *Proc. Natl. Acad. Sci. USA* **103** 5752–7
[27] Gorman J, Chowdhury A, Surtees J A, Shimada J, Reichman D R, Alani E and Greene E C 2007 Dynamic basis for one-dimensional DNA scanning by the mismatch repair complex msh2-msh6 *Mol. Cell* **28** 359–70
[28] Graneli A, Yeykal C C, Robertson R B and Greene E C 2006 Long-distance lateral diffusion of human rad51 on double-stranded DNA *Proc. Natl. Acad. Sci. USA* **103** 1221–6
[29] Wang Y M, Austin R H and Cox E C 2006 Single molecule measurements of repressor protein 1d diffusion on DNA *Phys. Rev. Lett.* **97** 048302

[30] Tafvizi A, Huang F, Leith J S, Fersht A R, Mirny L A and van Oijen A M 2008 Tumor suppressor p53 slides on DNA with low friction and high stability *Biophys. J.* **95** L01–L03
[31] Mueller F, Wach P and McNally J G 2008 Evidence for a common mode of transcription factor interaction with chromatin as revealed by improved quantitative fluorescence recovery after photobleaching *Biophys. J.* **94** 3323–39
[32] Slutsky M, Kardar M and Mirny L A 2004 Diffusion in correlated random potentials, with applications to DNA *Phys. Rev.* E **69** 061903
[33] Berg O G and von Hippel P H 1987 Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters *J. Mol. Biol.* **193** 723–50
[34] Iwahara J, Zweckstetter M and Clore G M 2006 NMR structural and kinetic characterization of a homeodomain diffusing and hopping on nonspecific DNA *Proc. Natl. Acad. Sci. USA* **103** 15062–7
[35] Gutin A, Sali A, Abkevich V, Karplus M and Shakhnovich E I 1998 Temperature dependence of the folding rate in a simple protein model: search for a glass transition *J. Chem. Phys.* **108** 6466–83
[36] Shakhnovich E 2006 Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet *Chem. Rev.* **106** 1559–88
[37] Slutsky M 2005 Protein–DNA interaction, random walks and polymer statistics *PhD Thesis* MIT
[38] Spolar R S and Record M T 1994 Coupling of local folding to site-specific binding of proteins to DNA *Science* **263** 777–84
[39] Kalodimos C G, Biris N, Bonvin A M, Levandoski M M, Guennuegues M, Boelens R and Kaptein R 2004 Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes *Science* **305** 386–9
[40] Viadiu H and Aggarwal A K 2000 Structure of Bamhi bound to nonspecific DNA: a model for DNA sliding *Mol. Cell* **5** 889–95
[41] Erie D A, Yang G, Schultz H C and Bustamante C 1994 DNA bending by cro protein in specific and nonspecific complexes: implications for protein site recognition and specificity *Science* **266** 1562–6
[42] Kalodimos C G, Boelens R and Kaptein R 2002 A residue-specific view of the association and dissociation pathway in protein–DNA recognition *Nat. Struct. Biol.* **9** 193–7
[43] Gerland U, Moroz J D and Hwa T 2002 Physical constraints and functional characteristics of transcription factor–DNA interaction *Proc. Natl. Acad. Sci. USA* **99** 12015–20
[44] Maerkl S J and Quake S R 2007 A systems approach to measuring the binding energy landscapes of transcription factors *Science* **315** 233–7
[45] Sheinman M and Kafri Y 2009 The effects of intersegmental transfers on target location by proteins *Phys. Biol.* **6** 16003
[46] Redner S 2001 *A Guide to First-Passage Processes* (Cambridge: Cambridge University Press)
[47] Neidhardt F C and Curtiss R 1996 *Escherichia Coli and Salmonella: Cellular and Molecular Biology* 2nd edn (Washington, DC: ASM Press)
[48] Kosmrlj A and Mirny L A 2009 Protein search in compact DNA, in preparation
[49] Vologodskii A V and Cozzarelli N R 1993 Monte Carlo analysis of the conformation of DNA catenanes *J. Mol. Biol.* **232** 1130–40
[50] Grosberg A Yu and Khokhlov A R 1994 *Statistical Physics of Macromolecules* (New York: AIP)
[51] Emanuel M, Radja N H, Henriksson A and Schiessel H 2009 The physics behind the larger scale organization of DNA in eukaryotes *Phys. Biol.* **6** 25008
[52] Lee W, Tillo D, Bray N, Morse R H, Davis R W, Hughes T R and Nislow C 2007 A high-resolution atlas of nucleosome occupancy in yeast *Nat. Genet.* **39** 1235–44
[53] Segal E and Widom J 2009 What controls nucleosome positions? *Trends Genet* **25** 335–43
[54] Sabo P J *et al* 2006 Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays *Nat. Methods* **3** 511–8
[55] Dion M F, Kaplan T, Kim M, Buratowski S, Friedman N and Rando O J 2007 Dynamics of replication-independent histone turnover in budding yeast *Science* **315** 1405–8
[56] Mavrich T N *et al* 2008 Nucleosome organization in the drosophila genome *Nature* **453** 358–62