Peer-reviewed version available at *Molecules* **2018**, *23*, 2106; doi:10.3390/molecules23092106

Article

Mechanisms of Protein Search for Targets on DNA: Theoretical Insights

Alexey A. Shvets [1], Maria P. Kochugaeva [2] and Anatoly B. Kolomeisky [3,*]

[1] Institute for Medical Engineering and Science, Massachusetts Institute of Technology; Cambridge, MA 02142, USA; shvets@mit.edu
[2] Department of Biomedical Engineering and System Biology Institute Yale University West Haven, CT, 06516, USA; maria.kochugaeva@yale.edu
[3] Department of Chemistry, Department of Chemical and Biomolecular Engineering, and Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, USA; tolya@rice.edu
* Correspondence: tolya@rice.edu

Abstract: Protein-DNA interactions are critical for the successful functioning of all natural systems. The key role in these interactions is played by processes of protein search for specific sites on DNA. Although it has been studied for many years, only recently microscopic aspects of these processes became more clear. In this work, we present a review on current theoretical understanding of the molecular mechanisms of the protein target search. A comprehensive discrete-state stochastic method to explain the dynamics of the protein search phenomena is introduced and explained. Our theoretical approach utilizes a first-passage analysis and it takes into account the most relevant physical-chemical processes. It is able to describe many fascinating features of the protein search, including unusually high effective association rates, high selectivity and specificity, and the robustness in the presence of crowders and sequence heterogeneity.

1. Introduction

Dynamical nature of underlying processes is what distinguishes the living systems from other processes. [1,2]. Biological processes constantly involve time-dependent fluxes of energy and materials, which makes them strongly deviating from equilibrium as long as organisms are alive. This implies that the concepts of equilibrium thermodynamics have limited applications for biological systems, while the role of methods that study the dynamical transformations is much more important [3]. In this review, we present our theoretical views on dynamic aspects of the protein-DNA interactions, which dominate in biological systems. Our approach is based on explicit calculations of dynamic properties via a first-passage probabilities analysis. The first-passage ideas have been already widely utilized in studies of various complex processes in Chemistry, Physics and Biology [4,5]. We employ these ideas in developing a discrete-state stochastic framework for analyzing the dynamics of protein search for specific targets on DNA.

It is known that the beginning of most biological processes is associated with specific protein molecules binding to specific target sequences on DNA because these events initiate the cascades of corresponding biochemical and biophysical processes [1–3]. For example, to activate or to repress a gene the corresponding transcription factor proteins must bind first to the gene promoter's region [1,2]. This fundamental aspect of protein-DNA interactions has been studied extensively by various experimental and theoretical methods [6–38]. A special attention was devoted to understanding the dynamics of the protein search for specific targets on DNA. Many ideas have been proposed and critically discussed, but only recently a clear molecular picture of the underlying processes started to emerge [11,12,17].

Large amount of experimental observations on protein search phenomena, which mostly come from the single-molecule measurements, suggests that it is a complex dynamic phenomenon which combines three-dimensional (in the bulk solution) and one-dimensional (on the DNA chain) motions

[9–12,16]. But the most paradoxical observation is that, although the protein molecules spend most of the search time (≥90-99%) on the DNA chain where they diffuse very slowly, they still can find the targets very fast, in some cases much faster than the bulk diffusion would allow [10–12]. For example, the measured association rate for *lac*-repressor was $\sim 10^{10} M^{-1} s^{-1}$ (two orders of magnitude faster than the diffusion limit!) [6], and many other experimentally determined protein-DNA association rates were also astonishingly high in comparison to typical biological binding rates. This is known as a *facilitated diffusion*. Several theoretical ideas on the origin of the facilitated diffusion, including lowering of dimensionality, electrostatic effects, correlations between 3D and 1D motions, conformational transitions, bending fluctuations, and hydrodynamics effects have been explored and discussed [10–12]. However, theoretical analysis shows that none of these mechanisms can fully explain the facilitated diffusion in the protein search [17]. To understand the dynamic aspects of protein-DNA interactions, we developed a discrete-state stochastic framework to take into account the most relevant physical-chemical processes in the system. The application of the first-passage probabilities method allows us also to explicitly evaluate the dynamic properties and to clarify dynamic aspects of the protein-DNA interactions.

It is important to note that although there are still different opinions on the theoretical foundations of the protein search phenomena, in this work we mostly present our views on these problems, which, of course, are subjective. In addition, there are many theoretical advances in our understanding of the protein search dynamics, but we will concentrate only on few of them in order to explain better the underlying molecular processes. Furthermore, there is a huge number of investigations on the protein target search phenomena. Our goal is not to cover all studies and all existing views but to present a clear theoretical picture of these processes as we understand it now.

2. Simplest Discrete-State Stochastic Model of the Protein Target Search

Experiments clearly indicate that during the search the protein molecule is alternating between freely diffusing behavior in the solution around the DNA chain and non-specific associations to DNA, which also include scanning the DNA chain [10–12]. The process is completed when the protein molecule reaches the specific target sequence on DNA for the first time. Stimulated by this observations, we start with a simplest minimal model of the protein search as presented in Figure 1. It is important to note that, in contrast to other theoretical approaches [10,11,15,32], this method is based on a discrete-state stochastic description of the system. This is a more realistic view of early stages of protein-DNA interactions because of intrinsically discrete nature of molecular interactions in these systems.

In this simple model, we consider a single protein molecule and a single DNA molecule with a single target site: see Figure 1. The DNA chain is viewed as having L discrete binding sites, and one of them at the position m is considered to be the target for the protein molecule. Because the diffusion of the proteins in the bulk is usually fast, all solutions states for the protein are combined into one state that we label as a state 0 (Figure 1). It is assumed that from the bulk solution the protein molecule can bind with equal probability to any site on DNA, and the total association rate to DNA is equal to k_{on}, while the dissociation rate from DNA is k_{off}. The non-specifically bound proteins can diffuse without bias along the DNA contour in any direction with a rate u (see Figure 1). Since the search process ends as soon as the protein molecule arrives to the specific site for the first time, we introduce a function $F_n(t)$, which is defined as a probability density function of reaching the site m (the target site) for the first time at time t if at $t = 0$ the protein started in the state n ($n = 0$ is the bulk solution, and $n = 1, ..., L$ are the protein-DNA bound states). This function is also known as a first-passage probability density function [4,5]. To compute these first-passage probabilities, we utilize backward master equations that describe the temporal evolution of these quantities [4,5,17],

$$\frac{dF_n(t)}{dt} = u\left[F_{n+1}(t) + F_{n-1}(t)\right] + k_{off}F_0(t) - (2u + k_{off})F_n(t), \quad (1)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2018 doi:10.20944/preprints201807.0606.v1

Peer-reviewed version available at Molecules 2018, 23, 2106; doi:10.3390/molecules23092106

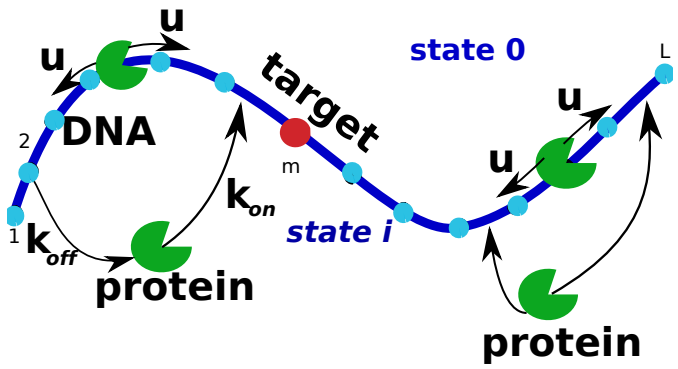


Figure 1. A schematic view of a minimal discrete-state stochastic model of the protein search for targets om DNA. The DNA chain has $L-1$ non-specific binding sites and one specific target site. A protein molecule can diffuse along the DNA segment with a rate u in both directions. It can also associate to DNA from the bulk solution (labeled as state 0) with a rate k_{on} or it can dissociate back to the solution with a rate k_{off}. The search is finished when the protein binds to the target site at the position m for the first time.

for $2 \leq n \leq L-1$, while at the boundaries ($n = 1$ or $n = L$) we have

$$\frac{dF_1(t)}{dt} = uF_2(t) + k_{off}F_0(t) - (u + k_{off})F_1(t), \tag{2}$$

and

$$\frac{dF_L(t)}{dt} = uF_{L-1}(t) + k_{off}F_0(t) - (u + k_{off})F_L(t). \tag{3}$$

For the state $n = 0$, the backward master equation is different,

$$\frac{dF_0(t)}{dt} = \frac{k_{on}}{L}\sum_{n=1}^{L} F_n(t) - k_{on}F_n(t). \tag{4}$$

Here we used the fact that the rate to bind to any site on DNA is k_{on}/L, so that the total association rate is equal to k_{on}. In addition, the initial conditions require that $F_m(t) = \delta(t)$ and $F_{n\neq m}(t = 0) = 0$. This means that if the protein molecule starts at the target site m the search is immediately accomplished.

It is important to explain the physical meaning of the backward master equations because they differ from classical forward master equations widely employed in Chemical Kinetics. It can be easily seen that all trajectories that start at the state n and finish at the target site m can be divided into several groups. For example, for $2 \leq n \leq L-1$ all trajectories starting at n can be divided into three groups: 1) passing via the state $n-1$, 2) passing via the state $n+1$ or 3) passing via the state 0 in the next time step. The fractions of those trajectories are given by $u/(2u + k_{off})$, $u/(2u + k_{off})$ and $k_{off}/(2u + k_{off})$, respectively. Equation (1) describes this partition of the trajectories in the time-dependent manner because the first-passage probability flux to the target is determined by these trajectories. Thus, the backward master equations reflect the temporal evolution of the first-passage probabilities.

The most convenient way to analyze the dynamics in the system is to use Laplace representations of the first-passage probability functions, $\widetilde{F_n(s)} \equiv \int_0^\infty e^{-st}F_n(t)dt$. Then Equations (1),(2), (3) and (4) can be written as simpler algebraic expressions:

$$(s + 2u + k_{off})\widetilde{F_n(s)} = u\left[\widetilde{F_{n+1}(s)} + \widetilde{F_{n-1}(s)}\right] + k_{off}\widetilde{F_0(s)}; \tag{5}$$

$$(s + u + k_{off})\widetilde{F_1(s)} = u\widetilde{F_2(s)} + k_{off}\widetilde{F_0(s)}; \tag{6}$$

$$(s + u + k_{off})\widetilde{F_L(s)} = u\widetilde{F_{L-1}(s)} + k_{off}\widetilde{F_0(s)}; \tag{7}$$

$$(s+k_{on})\widetilde{F_0(s)} = \frac{k_{on}}{L}\sum_{n=1}^{L}\widetilde{F_n(s)}. \tag{8}$$

In addition, from the initial conditions we have $\widetilde{F_m(s)} = 1$. These equations are solved assuming that the general form of the solution is $\widetilde{F_n(s)} = Ay^n + B$, where the unknown coefficients A, y and B are determined from the initial and boundary conditions [17]. One could argue that the target site m divides the DNA molecule into two homogeneous segments ($1 \le n \le m$ and $m \le n \le L$), which can be considered separately. It was shown [17] that this approach leads to explicit expressions for the first-passage probability functions. Specifically, one obtains

$$\widetilde{F_0(s)} = \frac{k_{on}(k_{off}+s)S_1(s)}{Ls(k_{off}+k_{on}+s)+k_{off}k_{on}S_1(s)}, \tag{9}$$

with an auxiliary function $S_1(s)$ defined as

$$S_1(s) = \frac{y(1+y)(y^{-L}-y^{L})}{(1-y)(y^{1-m}+y^{m})(y^{m-L}+y^{1+L-m})}; \tag{10}$$

and with the parameters y and B given by

$$y = \frac{s+2u+k_{off}-\sqrt{(s+2u+k_{off})^2-4u^2}}{2u}; \tag{11}$$

$$B = \frac{k_{off}\widetilde{F_0(s)}}{(k_{off}+s)}. \tag{12}$$

Explicit expressions for the first-passage probabilities provide a full dynamic description of the protein search processes and any relevant quantities can be easily computed. For example, the mean search time from the bulk solution, which is inversely proportional to the chemical association rate for the specific target site, can be found from[17],

$$T_0 \equiv -\frac{\partial\widetilde{F_0(s)}}{\partial s}\Big|_{s=0} = \frac{1}{k_{on}}\frac{L}{S_1(0)} + \frac{1}{k_{off}}\frac{L-S_1(0)}{S_1(0)}. \tag{13}$$

This result has a very clear physical meaning. Here the parameter $S_1(0)$ describes the average number of distinct sites that the protein molecule scans during each visit to DNA while searching for the single specific site. Then, on average, to find the target the protein must make $L/S_1(0)$ visits to DNA because during every association $S_1(0)$ DNA sites are checked. Each visit, on average, lasts $1/k_{on}$ while the protein scans for the target diffusing along the DNA chain. The protein also makes $L/S_1(0)-1$ dissociations back into the solution. The number of dissociation events is smaller by one than the number of association events because the last binding to DNA leads to finding the specific site.

The results of our calculations for the mean search times are presented in Figure 2. Our main finding here is that there are three dynamic search regimes depending on the values of kinetic parameters. It is convenient to introduce here a scanning length $\lambda = \sqrt{u/k_{off}}$, which gives the average distance that the protein molecule travels on DNA during each search cycle. This quantity is related to the parameter $S_1(0)$, but it is not the same because the protein might visit the same sites several times. If the protein molecule has a strong affinity to bind non-specifically to the DNA molecule (small k_{off}, $\lambda > L$), then there will be only one searching cycle. After binding to DNA the protein will not dissociate until it finds the target. In this case, the mean search time scales as $\sim L^2$ because the DNA-bound protein does a simple unbiased random walk. We call this dynamic phase a random-walk regime. Because of the redundancy of the random walk the search in this regime should be generally slow: many sites are repeatedly visited. In the opposite limit of weak attractions between DNA and

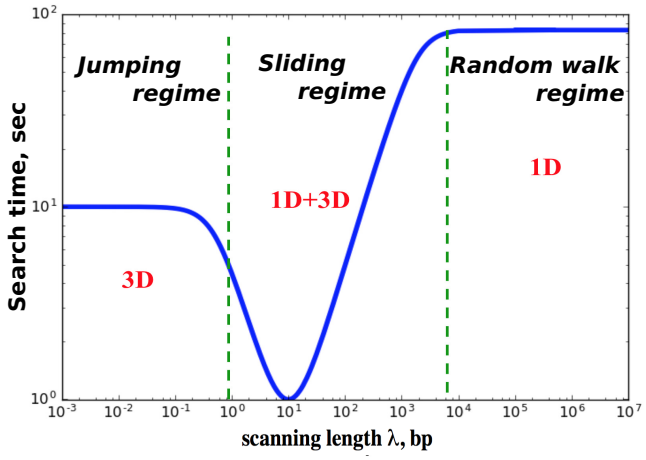


Figure 2. Mean search times as a function of the scanning length parameter $\lambda = \sqrt{u/k_{off}}$. The parameters utilized in calculations are: $L = 10^3$ bp, $u = k_{on} = 10^5\ s^{-1}$, and $m = L/2$. The transition rate k_{off} is varied to change λ.

protein molecules (large k_{off}, $\lambda < 1$), the protein can bind to DNA but it cannot slide because it quickly dissociates back into the solution. The protein on average makes L searching cycles ($T_0 \sim L$). This dynamic regime is called a jumping regime. The search in this regime is generally fast as long as the associations are also fast. The most interesting behavior is observed for the intermediate interactions, which we label as a sliding regime. Here the scanning length λ is larger than one but smaller than the length of DNA L, and the number of searching cycles is also proportional to L. But in this regime the system can reach the most optimal dynamic behavior with the smallest search times. This search facilitation is achieved due to the fact that the fluxes to the target are coming now from both the bulk solution and from the DNA chain. This is one of the main mechanisms of the facilitated diffusion of proteins during the target search, but other processes like inter-segment transfer might also contribute significantly in the facilitated diffusion [27].

3. The Effect of Multiple Targets and Traps

The advantage of the discrete-state stochastic framework with the first-passage analysis presented above is that it can be extended and generalized to more realistic biological situations. This allows us to investigate important questions related to the mechanisms of the protein target search on DNA. Let us present several specific examples, although many more results have been obtained.[17–29] We start with the problem of how the presence of multiple target sites or multiple semi-specific trap sites affect the dynamics of the protein search.

It is known that in eukaryotic cells multiple target sites are available on the accessible DNA fragments [1–3,40]. The protein search is accomplished in these systems when the protein molecules finds for the first time *any* of the target sites. It has been argued that the mean search time in this system might not decrease proportionally to the number of targets as one would naively expect from simple-minded applications of chemical kinetics [18]. This is due to the complex mechanism of the protein search that involves both 3D and 1D motions [18]. Applying our discrete-state stochastic framework to this problem, we consider a model with multiple targets at arbitrary locations as presented in Figure 3. To describe the search dynamics in this system, we again introduce the first-passage probability function $F_n(t)$ of finding *any* of the targets at time t if the process started at $t = 0$ at the site n. Targets are dividing the DNA chain into several homogeneous segments, and this

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2018 doi:10.20944/preprints201807.0606.v1

Peer-reviewed version available at Molecules 2018, 23, 2106; doi:10.3390/molecules23092106

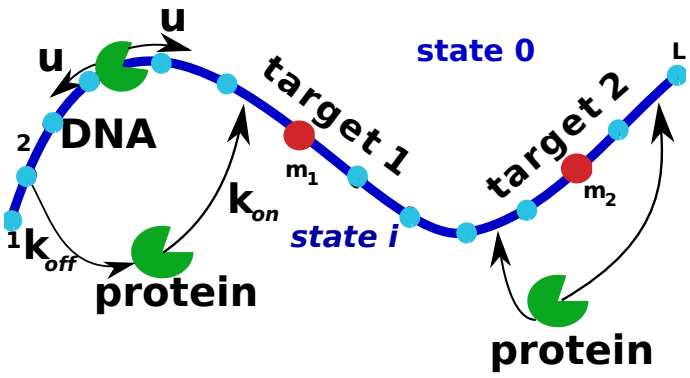


Figure 3. A schematic view of the discrete-state stochastic model of the protein search with multiple specific sites. Targets are located at the sites m_1 and m_2.

allows us to solve the corresponding backward master equations as explained in Section 2. This leads to the following explicit expression for the mean search time for any number of targets [18],

$$T_0 = \frac{1}{k_{on}} \frac{L}{S_i(0)} + \frac{1}{k_{off}} \frac{L - S_i(0)}{Si(0)}, \tag{14}$$

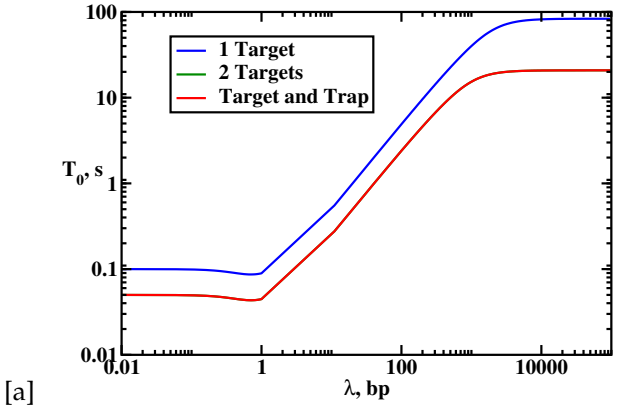
with a function $S_i(0)$ describing the average number of distinct sites scanned by the protein on DNA with i targets. This formula is a generalization of Equation (13) when there is only one target ($i = 1$). Specific expressions for $S_i(0)$ for various numbers of randomly distributed targets have been obtained [18]. For example, for $i = 2$ it was shown that

$$S_2(s) = \frac{(1+y)\left[2(1-y^{2L+m_1-m_2}) + (1-y^{m_2-m_1})(y^{2m_1-1} + y^{1+2(L-m_2)})\right]}{(1-y)(1+y^{2m_1-1})(1+y^{1+2(L-m_2)})(1+y^{m_2-m_1})}, \tag{15}$$

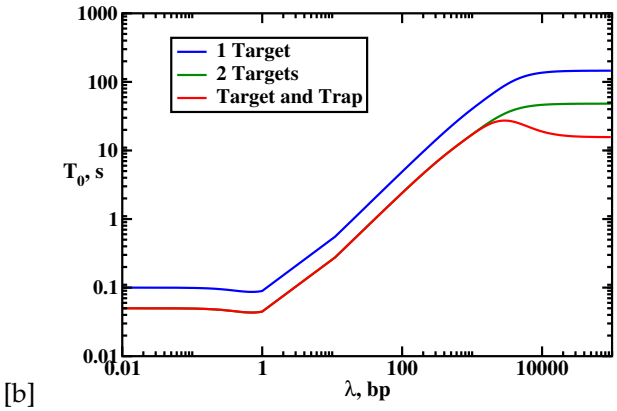
where the parameter y is given in Equation 11.

To understand the effect of multiple targets on the protein search dynamics, we analyze the results of explicit calculations for mean search times as presented in Figure 4. It is found that the presence of multiple targets does not affect the overall dynamic phase diagram as compared with the single-target case: three search regimes are again observed depending on the size of the scanning length, the target size and the size of the DNA segment. Generally, the search is faster in the multiple-target systems. However, surprisingly, increasing the number of specific sites might not always accelerate the search. To quantify this effect, we introduced an acceleration parameter, $a_n = T_0(1)/T_0(n)$, where $T_0(n)$ is the mean search for the system with n targets. This ratio gives a numerical value of how faster the search is in the presence of n targets in comparison with the single-target system. It is illustrated in Figure 5. One can see that there is a range of parameters when the search dynamics in the system with two targets can be slower than the dynamics in the system with one target. This happens in the effectively 1D search regime (random-walk dynamic phase) when the single target is located in the middle of the DNA chain, while two targets are close to each other and located near one of the ends of the DNA segment. In this case, for the protein molecule the two targets are viewed as effectively a single target site (with the size equal to two target sites) because they are so close to each other. But it is faster to find the target located in the middle of the chain than the target positioned near the ends.[17] This is the main reason why having multiple targets does not always lead to decrease in the search times. Thus, our theoretical analysis predicts that the degree of acceleration due to the presence of multiple targets depends on the nature of the dynamic search phase and on the location of the specific sites with respect to each other and with respect to the middle point of DNA [18].

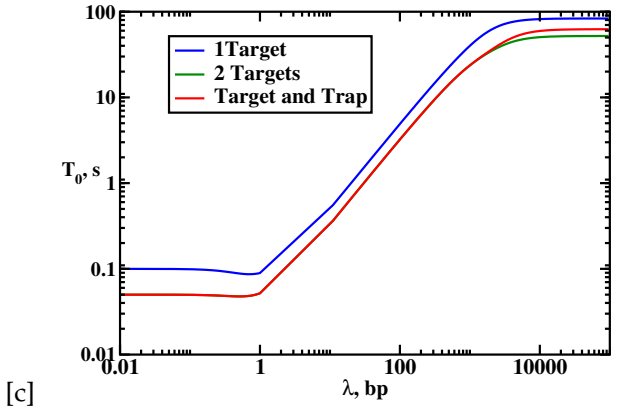
Another important factor that might affect the protein search dynamics is the existence of so-called semi-specific sites, or decoys, on DNA. These sites have a chemical composition very similar to the



[a]



[b]



[c]

Figure 4. Dynamic phase diagrams for the protein search on DNA with one target at the position m, with two targets at the positions m_1 and m_2 and with the target and the trap at the positions m_1 and m_2, respectively. Parameters used for calculations are: $k_{on} = u = 10^5$ s^{-1} and $L = 10000$. a) $m = L/2$, $m_1 = L/4$ and $m_2 = 3L/4$; b) $m = L/4$, $m_1 = L/4$ and $m_2 = L/2$; and c) $m = L/2$, $m_1 = L/2$ and $m_2 = L$. Adapted with permission from Ref. [19].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2018 doi:10.20944/preprints201807.0606.v1

Peer-reviewed version available at Molecules 2018, 23, 2106; doi:10.3390/molecules23092106

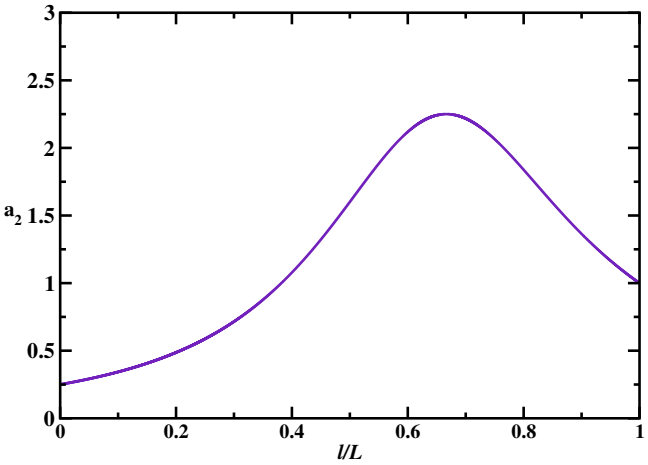


Figure 5. Ratio of the mean search times as a function of the normalized distance between the targets for single-target and two-target systems (l is the distance between between targets, L is the DNA length). The single target is in the middle of the chain. In the two-target system, one of the specific sites is fixed at the end and the position of the second one is varied. The parameters used in calculations are: $u = k_{on} = 10^6$ s^{-1}; $k_{off} = 10^{-4}$ s^{-1}; and $L = 10000$. Adapted with permission from Ref. [18].

specific targets with differences in only one or few nucleotides. The protein molecule can be trapped in these sites, and this should influence the search for real targets. To analyze this effect, we can extend the simplest model to include the possibility of traps, assuming that associations to these semi-specific sites are effectively irreversible [19]. This assumption is reasonable because the search times in many systems are relatively short and the experimental observations also limited in time. Thus the bindings to decoys can be viewed as effectively irreversible. The first-passage analysis can be applied here, but we have to notice that only a fraction of trajectories will reach the correct target site. Then the main quantity of our calculations, the first-passage probability function $F_n(t)$, is now a *conditional* probability for the protein molecules not captured by the trap to find the target site.

Let us consider a system consisting of a single target at the site m_1 and a single trap at the site m_2 on the DNA molecule with L sites [19]. The scheme presented in Figure 3 is also a correct representation of this system with the correction that instead of the second target there is a trap in the site m_2, and the successful search corresponds to the protein molecule finding the specific site m_1. Following our theoretical method, the corresponding backward master equations can be solved and they yield the Laplace transform of the first-passage probability function to find the target if the protein starts from the bulk solution [19],

$$\widetilde{F_0(s)} = \frac{k_{on}(k_{off}+s)S_0(s)}{Ls(k_{off}+k_{on}+s)+k_{off}k_{on}S_2(s)}, \tag{16}$$

with

$$S_0(s) = \frac{(1+y)(1-y^{m_1+m_2-1})}{(1-y)(1+y^{2m_1-1})(1+y^{m_1-m_2})}, \tag{17}$$

and the parameters y and S_2 given in Equations (11) and (15), respectively. This allows us to evaluate all dynamic properties in the system and to test the effect of traps.

The probability to reach the target (i.e., the fraction of the successful trajectories) is now given by a so-called splitting probability function [4,5],

$$\Pi \equiv \widetilde{F_0(s=0)} = \frac{S_0(0)}{S_2(0)}. \tag{18}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2018 doi:10.20944/preprints201807.0606.v1

Peer-reviewed version available at Molecules 2018, 23, 2106; doi:10.3390/molecules23092106

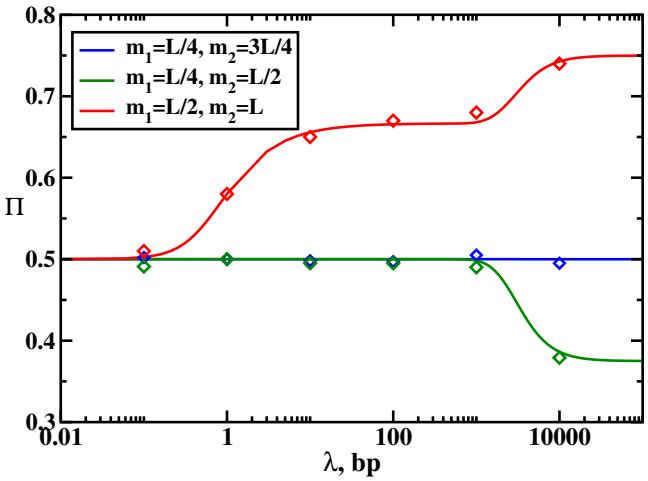


Figure 6. Probability to reach the target as a function of the scanning length for different distributions of the target and trap sites. Parameters used for calculations are: $k_{on} = u = 10^5$ s^{-1}, $L = 10000$ and k_{off} is changing. Symbols are from Monte Carlo computer simulations. Adapted with permission from Ref. [19]

The mean search time, which is the conditional mean first-passage time to reach the target, can be estimated by averaging over the successful trajectories, producing

$$T_0 \equiv -\frac{\left.\frac{\partial \widetilde{F_0(s)}}{\partial s}\right|_{s=0}}{\Pi} = \frac{1}{k_{on}}\frac{L}{S_2(0)} + \frac{1}{k_{off}}\frac{L - S_2(0)}{S_2(0)} + \Pi \frac{\partial}{\partial s}\left[\frac{S_2(s)}{S_0(s)}\right]\Bigg|_{s=0}. \quad (19)$$

Let us analyze this expression. On the left side, the division by the splitting probability emphasizes the fact that this is the conditional mean search time. It is also interesting to note that the first two terms on the right side of the equation is exactly the mean search time for the system with two targets and no traps (at the sites m_1 and m_2) as we discussed above [18], while the third term is a correction which accounts for the fact that the site at m_2 is actually the trap. The main reason for this is the observation that the sites m_1 and m_2 are special locations where all trajectories are end up in both systems, with two targets and with the target and the trap. For the two-target case the mean search times are averaged over all trajectories to both sites, while for the target and the trap system the mean search times are obtained only by considering the trajectories finishing at the target [19].

The results of calculations for the dynamic properties of the protein search in the presence of traps are presented in Figures 4 and 6. Again, three dynamic search phases are observed, but adding the trap generally facilitates the search dynamics, which is a counter-intuitive result: see Figure 4. However, this acceleration (in comparison with the single-target system) is always associated with lowering of the probability of reaching the specific target, as shown in Figure 6. This means that the protein molecules might reach the target faster in the presence of the traps, but the fraction of such events is decreasing. In addition, the search dynamics is sensitive to the nature of the dynamic phase. The strongest effect due to the presence of the trap is observed in the effective 1D random-walk regime (because it has only one searching cycle) where the locations of the target and the trap strongly influence the search. In other dynamic regimes, the effect is smaller.

4. Sequence heterogeneity

Real DNA molecules are heterogeneous polymers consisting of several types of subunits. This means that the interactions between protein and DNA molecules depend on the DNA sequence at the location where they meet. It is reasonable to expect that this sequence dependence in the interaction

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2018 doi:10.20944/preprints201807.0606.v1

Peer-reviewed version available at Molecules 2018, 23, 2106; doi:10.3390/molecules23092106

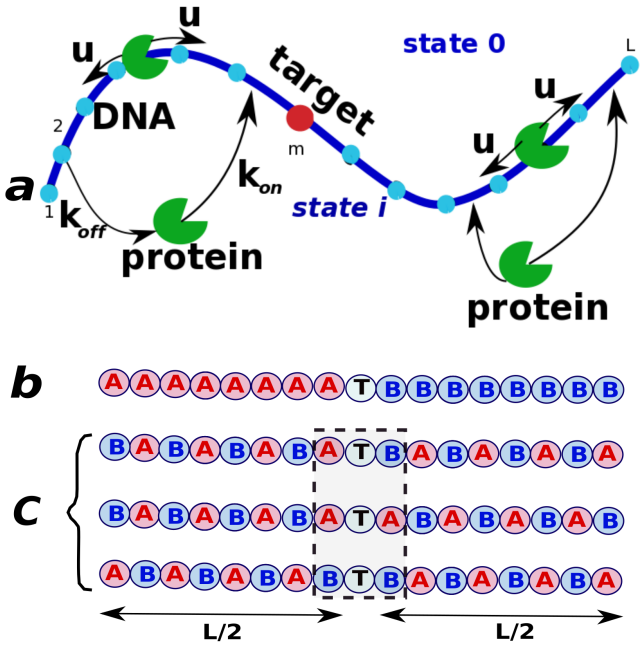


Figure 7. A simplified view on the protein search on DNA with two different types of subunits, A and B. a) A general scheme; b) DNA is viewed as a symmetric block copolymer with the target in the middle of the chain; c) DNA is viewed as alternating copolymer with different compositions of the subunits flanking the target in the middle of the chain. Adapted with permission from Ref. [20]

strength should affect the protein search dynamics because the diffusion rate for the non-specifically bound proteins will be position-dependent [3,11,41]. Similarly, association and dissociation rates should also depend on the location of the protein molecule on DNA. In addition, recent theoretical investigations suggested that different DNA sequence symmetries might lead to additional effective interactions between protein and DNA molecules [43–46]. The discrete-state stochastic framework with the first-passage analysis is a convenient tool to investigate the effect of DNA sequence heterogeneity and symmetry on the protein search dynamics [20].

Our goal here is clarify the molecular origin of how the sequence heterogeneity influences the protein target search. We assume here a simplified picture of DNA, in which each monomer can be one of two chemical species, A or B, as presented in Figure 7 [20]. When the protein is bound to the subunit A (B), it interacts with energy ε_A (ε_B), and the difference between interaction energies is given by a parameter $\varepsilon = \varepsilon_A - \varepsilon_B \geq 0$. This means that the protein attracts stronger to the B sites than to the A sites. The protein molecule can diffuse along DNA with a rate $u_A \equiv u$ or $u_B = ue^{-\varepsilon}$, where ε is measured in k_BT units. This reflects the assumption that if the protein interacts stronger with the DNA at given location then it will move out of this site slower. In addition, we assume that, independently of the chemical nature of the neighboring sites, sliding out of the sites A is characterized by the rate u_A, while the diffusion out of the sites B is given by u_B. From the bulk solution the protein might associate to any site A or B on DNA with the corresponding rates $k_{on}^A = k_{on}$ or $k_{on}^B = k_{on}e^{-\theta\varepsilon}$. Note that for convenience the on-rates defined here as the rates per unit site, in contrast to our definitions in the previous sections. Similarly, the dissociations from the DNA chain are described by the rates $k_{off}^A = k_{off}$ and $k_{off}^B = k_{off}e^{(\theta-1)\varepsilon}$. Here, the parameter $0 \leq \theta \leq 1$ specifies how the protein-DNA interaction energy is distributed between the association and dissociation transitions [20]. The physical meaning of this parameter is that the protein molecule tends to bind faster and to dissociate slower from the stronger attracting sites B, as compared with the weaker attracting A sites. The parameter θ accounts for these effects.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2018 doi:10.20944/preprints201807.0606.v1

Peer-reviewed version available at Molecules 2018, 23, 2106; doi:10.3390/molecules23092106

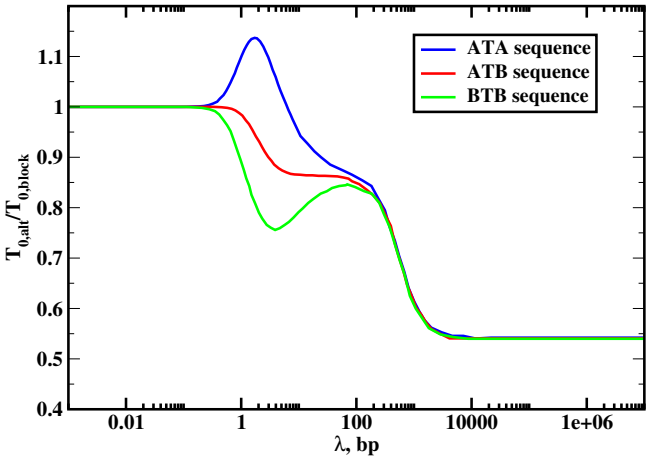


Figure 8. The ratio of the mean search times for the alternating DNA sequences and for the block copolymer DNA sequences as a function of the scanning length $\lambda = \sqrt{u/k_{off}}$. Three different chemical compositions near the target (T) are distinguished, namely, ATA, ATB, BTB. The transition rates are $u = 10^5\, s^{-1}$ and $k_{on} = 0.1\, s^{-1}$. The DNA length is $L = 1000$, the loading parameter is $\theta = 0.5$, and the energy difference of interactions for the protein with A and B sites is $\varepsilon = 5\, k_B T$. Adapted with permission from Ref. [20].

To quantify the role of sequence heterogeneity, we consider the DNA molecule with a fixed chemical composition (the fractions of A and B monomers are the same), but with different arrangements of subunits. Two limiting cases are specifically analyzed. One of them views the DNA molecule as two homogeneous segments of only A and only B subunits separated by the target in the middle of the chain (Figure 7). Another one is the DNA chain with the alternating A and B sites. The block copolymer has two homogeneous sequence segments, while the alternating polymers are more heterogeneous. It is important to note that in both cases, the overall interaction between the protein and DNA is the same (because the overall chemical composition in both cases is identical), and thus our analysis probes only the effect of the heterogeneity and symmetry in the subunit positions, in contrast to other theoretical treatments [42].

Applying again the first-passage approach and solving the corresponding equations leads to the explicit expressions for mean search times for all situations shown in Figure 7 [20]. For example, for the block copolymer DNA sequences, we obtain

$$T_0 = \frac{k_{off} + k_{on}\left[(L/2 - P_A) + e^{\varepsilon}(L/2 - P_B)\right]}{k_{on}k_{off}(1 + P_A + e^{\theta\varepsilon}P_B)}, \tag{20}$$

where

$$P_i = \frac{x_i^{1-L/2} - x_i^{1+L/2}}{(1 - x_i)(x_i^{1+L/2} + x_i^{L}/2)}, \tag{21}$$

$$x_i = \frac{2u_i + k_{off}^{(i)} - \sqrt{(2u_i + k_{off}^{(i)})^2 - 4u_i^2}}{2u_i}, \tag{22}$$

for $i = A$ or B. The expressions for the mean search time for alternating sequences are quite bulky and can be found in Ref. [20].

The results of our calculations are presented in Figure 8, where the ratio of the mean search times for the block copolymer and alternating sequences are plotted. The analysis of this figure produces several interesting observations. First, we see that three dynamic search regimes are also found in this system and the effect of sequence heterogeneity on protein search dynamics depends on the nature of

Peer-reviewed version available at Molecules 2018, 23, 2106; doi:10.3390/molecules23092106

the dynamic phase. In the jumping regime when the protein does not slide along the DNA contour ($\lambda < 1$), the symmetry of the sequence does not play any role. This is because in this case the process is taking place only via associations and dissociations (3D search), and the structure of the DNA chain is not important. The situation is different for the intermediate sliding regime (3D+1D search, $1 < \lambda < L$) where in most cases, the search on alternating sequences is faster. This can be explained by noticing that the search time in this dynamic phase is proportional to L/λ, which gives the average number of cycles before the protein can find the target. In the block copolymer sequence, the protein mostly comes to the target from the B segment because of stronger interactions with these sites, i.e, it comes from one side of the DNA molecule. In the alternating sequences, the protein can reach the target from both sides of DNA, and this lowers the overall search time. It can be shown analytically that the scanning length on the alternating segment is larger than the scanning length for the B segment, i.e., $\lambda_{AB} > \lambda_B$ [20]. Then the search is faster for the alternating sequences because $L/\lambda_{AB} < L/\lambda_B$, i.e., the number of searching cycles is lower for the alternating sequences, which helps to find the target faster. The only deviation from this picture is found for ATA sequences, which corresponds to having two A sites around the target site, where for the small range of parameters the search is slower than in the block copolymer sequence. This effect can be explained by the fact that the protein does not sit at A sites for the long time and it moves quickly away, effectively increasing the barrier to enter the target via DNA [20]. Thus, our theory predicts that the composition of the DNA flanking sites around the target sequences might affect the dynamics of reaching them. It is interesting to note that recent experiments are consistent with our theoretical predictions [47].

In the random-walk regime (1D search, $\lambda > L$), the effect of the sequence heterogeneity is even stronger. The protein molecule finds the specific binding site up to 2 times faster for more heterogeneous alternating DNA sequences. To understand this behavior, we note that in this case the mean first-passage time to reach the target is a sum of residence times on the DNA sites since the protein will not dissociate until the target is located so that all trajectories to the target are one-dimensional. Because the target is in the middle of the chain, the mean time to reach the target from the block copolymer sequence can be approximated as $T_0 \simeq (L/4)\tau_B$, where τ_B is the average residence time on any site B. The protein prefers to start the search at any position on the B segment with equal probability, i.e., the distance to the target varies from 0 to $L/2$. Then, the average starting position of the protein is $L/4$ sites away from the target. For the alternating sequences, the average distance to the target is approximately the same ($L/4$), but the chemical composition of intermediate sites on the path to the target is different, yielding, $T_0 \simeq (L/8)\tau_A + (L/8)\tau_B$ (τ_A is the residence time on A sites). The protein spends much less time on A subunits, and this leads to faster search for the alternating DNA sequences. For $\tau_A \ll \tau_B$, this also explains the factor of 2 in the search speed. In this case, the B subunits can be viewed as effective traps that slow down the search dynamics. Thus, our theoretical calculations make surprising predictions that the sequence heterogeneity almost always lead to faster protein search for targets on DNA despite the fact that it lowers the effective protein-DNA binding affinity [43–46]. And the stronger the contribution of the 1D search modes, the more relevant will be the effect of sequence heterogeneity.

5. The Effect of Crowding on DNA in the Protein Target Search

Living cells are typically crowded with a large number of molecules, and many of them are attached to the DNA chains [1,2]. This should prevent the fast protein search for targets on DNA, and earlier theoretical studies supported this prediction [49]. However, surprisingly, experiments show that crowding on DNA does not affect much the effectiveness of the protein target search [33,34], and this was also found in MD simulations [48]. By applying the discrete-state stochastic approach, we were able to clarify the role of the crowding on DNA in the protein target search.

To analyze this problem, the model illustrated in Figure 9 is considered. There is a single DNA molecule with $L + 1$ binding sites, and one of them is the target (at the site m). On the DNA chain there is also a crowding particle that can diffuse with a rate u_{ob}, but it cannot leave DNA. A single protein

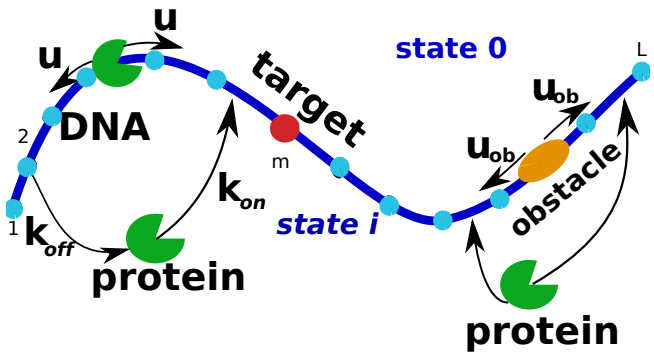


Figure 9. A schematic view of the protein target search in the presence of a moving obstacle on DNA. The crowding particle cannot dissociate from DNA, while the protein molecule can dissociate into the solution, labeled as state 0, and return back to the DNA chain.

molecule starts from the solution (state 0) and it can bind to any site on DNA that is not occupied by the crowder with a rate k_{on} (rate per site). The bound protein molecule can diffuse with a rate u, and there is an exclusion interaction between the protein and the crowder. Finally, the protein molecule can dissociate from DNA to the bulk solution with a rate k_{off}: see Figure 9.

Investigating the model with the mobile crowding particle on DNA first using Monte Carlo computer simulations, it is found that there are three search regimes depending on the main length scales in the system. This is shown in Figure 10 for the mean search times to find the target as a function the scanning length λ. We can understand the complex dynamics in this system using the following arguments. If the diffusion rate of the crowder is much smaller than other rates ($u_{ob} \ll u$, k_{on} and k_{off}), then the protein molecule will find the target before the crowding particle can move away from its original location. But we already explicitly solved the problem of the protein target search with static obstacles using the same discrete-state stochastic approach with the first-passage analysis [23]. Then the mean search time in the system with movable crowder can be approximated as the average over all possible static locations of the crowding particle [21], yielding

$$\langle T_0 \rangle \simeq \frac{1}{L}\left(\sum_{l_{ob}=1}^{m-1} T_{ob}(l_{ob}) + \sum_{l_{ob}=1}^{L-m} T_{ob}(l_{ob})\right), \tag{23}$$

where

$$T_{ob} = \frac{k_{off} + k_{on}(L - S_{ob}(0))}{k_{on}k_{off}S_{ob}(0)}, \tag{24}$$

is the mean search time with the static obstacle located at a distance l_{ob} from the target. An auxiliary function S_{ob} is given by [23]

$$S_{ob}(s) = \frac{y(y^{-m} - y^{m})}{(1-y)(y^{m} + y^{1+m})} + \frac{y(1 - y^{2l_{ob}-2})}{(1-y)(1 + y^{2l_{ob}-1})} \tag{25}$$

with the parameter y specified in Equation 11.

This simple approximate theory works quite well in the dynamic regimes where 3D pathways are important for the search ($\lambda < L$). However, theoretical arguments fail in the random-walk regime where 1D dynamics dominate the search. These results are expected. The protein molecule that collides with the crowding particle on DNA in dynamic regimes with 3D pathways will have the opportunity to dissociate into the bulk solution and to avoid the blocking effect. But in the random-walk regime (1D search) there is no such opportunity, and the search times will definitely increase. Computer simulations also indicate that the search times in this regime depend on the diffusivity of the crowding particle. The search is faster for more mobile crowders: see Figure 10.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 July 2018 doi:10.20944/preprints201807.0606.v1

Peer-reviewed version available at Molecules 2018, 23, 2106; doi:10.3390/molecules23092106

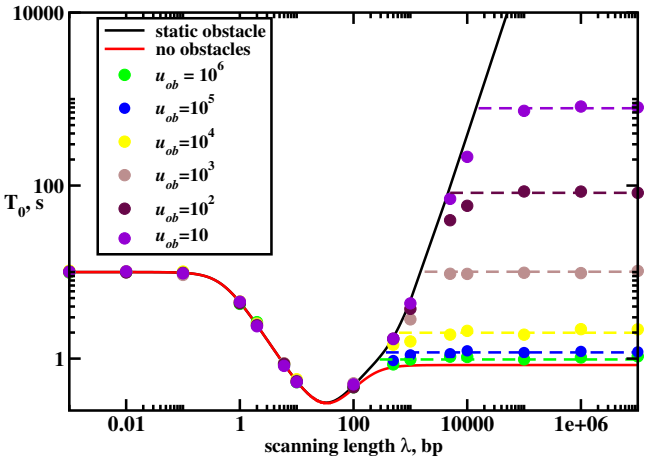


Figure 10. Mean search times to find the target in the system with a mobile crowder on DNA. The DNA chain has $L = 1000$ sites, and the target is in the middle of the chain, $m = L/2$. Parameters used for calculations are $k_{on} = 0.1$ s^{-1}, $u = 10^5$ s^{-1} and variable u_{ob}. Solid curves correspond to analytical results for DNA without obstacles and for DNA with a static obstacle, which are averaged over all initial positions of the crowder. Symbols correspond to Monte Carlo computer simulations. Dashed lines describe the approximate theory, as explained in the text. Adapted with permission from Ref. [21].

The dynamics in the random-walk regime can be explained using the following arguments. The overall search can be viewed as consisting of two terms,

$$\langle T_0^{ob} \rangle \simeq T_0 + \langle T_{bl} \rangle, \tag{26}$$

where T_0 is the search in the random-walk regime without any crowders, and it is given in Equation 13. The second term is the average time it takes for the crowder to diffuse away and clear the path for the protein to reach the target without interference [21]. It was shown that this blocking time T_{bl} depends on the location of the target and the diffusion rate of the crowding particle u_{ob} [21],

$$\langle T_{bl} \rangle = \frac{m^4 + (L-m)^4}{16 u_{ob}(L^2 + m^2 - mL)}. \tag{27}$$

This simple theoretical arguments show excellent agreement with Monte Carlo computer simulations: see dashed lines in Figure 10. But more importantly, they provide a clear molecular picture on the role of the crowding on DNA in the protein target search. If the protein search is dominated by 1D pathways and the mobility of the crowder is low the search dynamics will be significantly slowed down. But if the search involves mostly 3D pathways and the crowder is mobile the mean search times will not be affected much. It seems that real biological systems operate in 3D+1D regime, and crowding particles diffuse with the rates comparable to the searching proteins ($u \sim u_{ob}$) [3]. Then one might conclude that the effect of the crowders on DNA should be minimal. This fully agrees with experimental observations and with results from MD simulations [34,48].

6. Conclusions and Future Directions

Although protein search for targets on DNA is a very complex phenomenon that involves multiple biochemical and biophysical processes, significant advances in our understanding of the underlying molecular mechanisms have been achieved in recent years. A major role in this success is due to analysis of the systems using the discrete-state stochastic framework supplemented by explicit calculations via the first-passage probabilities method. In this review, we presented and explained this theoretical approach by considering the protein target search in various systems. It is important to

emphasize that the main advantage of our theoretical approach is the ability to obtain analytical results that clarify the physics of the underlying processes. In addition, the method can be easily extended in many directions, as shown in this work, as well as in other cases which we did not discuss in this work, such as the role of conformational transitions [24] and the effect of DNA loop formation in the protein target search [23]. Furthermore, our theoretical calculations using this theoretical framework were successful in explaining the experimental observations on homology search by RecA protein filaments [25] and the dynamics of CRISPR genome interrogation [29].

Several important dynamic features of the protein search for targets on DNA have been identified from theoretical analysis. It is found that the dynamic phase diagram of the protein target search always shows thee dynamic regimes, which are determined by the three relevant length scales in the system: the size of DNA, the average scanning length of the non-specifically bound proteins, and the size of the target sequence. Depending on the dynamic phase, the search is dominated by the 3D motions (jumping regime), 1D motions (random-walk regime) or a combination of 3D and 1D motions in the sliding regime. The analysis shows that the most optimal search dynamics can be achieved in the dynamic regime when the protein molecules explore both 1D and 3D pathways during the search. In this case, the protein can reach the target by sliding from the DNA chain or by directly binding from the solutions. Theoretical calculations also indicate that the presence of several target sites influences the search dynamics differently depending on the locations of the targets on DNA and distances between them. Surprising observations are found in the system with semi-specific sites, which are viewed as effective traps. It is shown that the search dynamics can be faster in this case, but it comes with the price of lowering the yield of the protein molecules reaching the target. We also investigated the effect of sequence heterogeneity and symmetry in the protein search dynamics. Our calculations indicate that the search is faster for more heterogeneous sequences, and the chemical composition around the target is also an important factor in this process. Furthermore, our method allowed us to probe the effect of crowding on DNA in the protein target search. It is shown that it depends on the dynamic phase and on the mobility of the crowding particles. The crowders influence the protein search stronger when 1D pathways dominate and when the diffusivity of the crowding particle is small enough so that the protein will be frequently blocked during the process. Increasing the mobility of the crowders and/or increasing the contribution of 3D search pathways lowers the effect of the crowding. These theoretical arguments fully agree with experimental observations and MD computer simulations.

Despite tremendous progress in theoretical understanding of the protein target search phenomena, there are many questions remain on the molecular mechanisms of these processes. It is still unclear what is the nature of protein-DNA interactions in the regions surrounding the target sequences. Is the effective potential created by these interactions drives the protein molecule to the target like a funnel or is it completely random? How large is the size of the flanking segments that affect the finding of the target? What is the role of DNA topology in the protein target search? This is especially important for proteins that have several binding sites for DNA which can form DNA loops and other complex structures. Another interesting question is the role of various DNA and protein conformations in these processes. It is clear that further progress in understanding protein target search phenomena depends on combining theoretical, computational and experimental methods.

Author Contributions: Conceptualization, A.K., A.S. and M.K.; Methodology, A.K.; Software, A.S. and M.K.; Validation, A.S. and M.K.; Formal Analysis, A.K., A.S. and M.K.; Investigation, A.K., A.S. and M.K.; Resources, A.K.; Data Curation, A.K.; Writing — Original Draft Preparation, A.K., A.S. and M.K.; Writing — Review & Editing, A.K.; Supervision, A.K.; Project Administration, A.K.; Funding Acquisition, A.K.

Funding: This research was funded by Welch Foundation grant number [C-1559], NSF grant number [CHE-1664218] and from the Center for Theoretical Biological Physics sponsored by the NSF grant number [PHY-1427654].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alberts, B. et al., *Molecular Biology of Cell*, 6th ed., Garland Science, New York, 2014.
2. Lodish, H. et al., *Molecular Cell Biology*, 6th ed., W. H. Freeman, New York, 2007.
3. Phillips, R.; Kondev, J.; Theriot, J. *Physical Biology of the Cell*, 2nd ed., Garland Science, New York, 2012.
4. Van Kampen, N.G. *Stochastic Processes in Physics and Chemistry*, 3rd ed., North Holland, Amsterdam, 2007.
5. Redner, S. *A Guide to First-Passage Processes*, Cambridge University Press, Cambridge, 2001.
6. Riggs, A.D.; Bourgeois, S.; Cohn, M. The lac-represser-operator interaction: IIIKinetic studies. *J.Mol. Biol.* **1970**, *53*, 401-417.
7. Berg, O.G.; Winter, R.B.; von Hippel, P.H. Diffusion-driven mechanisms of protein translocation on nucleic acids: I. Models and theory. *Biochemistry* **1981**, *20*, 6929-6948.
8. Berg, O.G.; von Hippel, P.H. Diffusion-controlled macromolecular interactions. *Annu. Rev. Biophys. Biophys. Chem.* **1985**, *14*, 131-160.
9. Gowers, D.M.; Wilson, G.G.; Halford, S.E. Measurements of the contributions of 1D and 3D pathways to the translocation of protein along DNA. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15883-15888.
10. Halford, S.E.; Marko, J.F. How do site-specific DNA-binding proteins find their targets? *Nucl. Acids Res.* **2004**, *32*, 3040-3052.
11. Mirny, L.; Slutsky, M.; Wunderlich, Z.; Tafvizi, A.; Leith, J.; Kosmrlj, A. How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J. Phys. A: Math. Theor.* **2009**, *42*, 434019.
12. Kolomeisky, A.B. Physics of protein-DNA interactions: Mechanisms of facilitated target search. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2088-2095.
13. Hu, T.; Grosberg, A.Y.; Shklovskii, B.I. How proteins search for their specific sites on DNA: the Role of DNA conformations. *Biophys. J.* **2006**, *90*, 2731-2744.
14. Hu, L.; Grosberg, A.Y.; Bruinsma, R. Are DNA transcription factor proteins maxwellian demons? *Biophys. J.* **2008**, *95*, 1151-1156.
15. Bauer, M.; Metzler, R. Generalized facilitated diffusion model for DNA-binding proteins with search and recognition states. *Biophys. J.* **2012**, *102*, 2321-2330.
16. Sheinman, M.; Benichou, O.; Kafri, Y.; Voituriez, R. Classes of fast and specific search mechanisms for proteins on DNA. *Rep. Progr. Phys.* **2012**, *75*, 026601.
17. Veksler, A.; Kolomeisky, A.B. Speed-selectivity paradox in the protein search for targets on DNA: Is it real or not? *J. Phys. Chem. B* **2013**, *117*, 12695-12701.
18. Lange, M.; Kochugaeva, M.; Kolomeisky, A.B. Protein search for multiple targets on DNA. *J. Chem. Phys.* **2015**, *143*, 105102.
19. Lange, M.; Kochugaeva, M.; Kolomeisky, A.B. Dynamics of the protein search for targets on DNA in the presence of traps. *J. Phys. Chem. B* **2015**, *119*, 12410-12416.
20. Shvets, A.A.; Kolomeisky, A.B. Sequence heterogeneity accelerates protein search for targets on DNA. *J. Chem. Phys.* **2015** *143*, 245101.
21. Shvets, A.A.; Kolomeisky, A.B. Crowding on DNA in protein search for targets. *J. Phys. Chem. Lett.* **2016**, *7*, 2502-2506.
22. Shvets, A.A.; Kochugaeva, M.; Kolomeisky, A.B. The role of static and dynamic obstacles in the protein search for targets on DNA. *J. Phys. Chem. B* **2015**, *120*, 5802-5809.
23. Shvets, A.A.; Kolomeisky, A.B. The role of DNA looping in the search for specific targtes on DNA by multisite proteins. *J. Phys. Chem. Lett.* **2016**, *7*, 5022-5027.
24. Kochugaeva, M.P.; Shvets, A.A.; Kolomeisky, A.B. How conformational dynamics influences the protein search for targets on DNA. *J. Phys. A: Math. Theor.* **2016**, *49*, 444004.
25. Kochugaeva, M.P.; Berezhkovskii, A.A.; Kolomeisky, A.B. Optimal length of conformational transitions region in the protein search for targets on DNA. *J. Phys. Chem. Lett.* **2017**, *8*, 4049-4054.
26. Shin, J.; Kolomeisky, A.B. Surface-assisted dynamic search processes. *J. Phys. Chem. B* **2018**, *122*, 2243-2250.
27. Esadze, A.; Kemme, C.A.; Kolomeisky, A.B.; Iwahara, J. Positive and negative impacts of nonspecific sites during targte location by a sequence-specific DNA-binding protein: Origin of the optimal search at physiological ionic strength. *Nucl. Acids Res.* **2014**, *42*, 7039-7046.
28. Kochugaeva, M.P.; Shvets, A.A.; Kolomeisky, A.B. On the mechanism of homology search by ReacA protein filaments. *Biophys. J.* **2017** *112*, 859-867.

29. Shvets, A.A.; Kolomeisky, A.B. Mechanism of genome interrogation: How CRISPR RNA-guided Cas9 proteins locate specific targets on DNA. *Biophys. J.* **2017**, *112*, 1416-1424.
30. Tafvizi, A.; Huang, F.; Fersht, A.R.; Mirny, L.A.; van Oijen, A.M. A single-molecule characterization of p53 search on DNA. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 563-568.
31. Slutsky, M.; Mirny, L.A. Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophys. J.* **2004**, *87*, 4021-4035.
32. Benichou, O.; Kafri, Y., Sheinman, M.; Voituriez, R. Searching fast for a target on DNA without falling to traps. *Phys. Rev. Lett.* **2009**, *103*, 138102-138104.
33. Hammar, P.; Leroy, P.; Mahmutovic, A.; Marklund, E.G.; Berg, O.G.; Elf, J. The *lac* Repressor displays facilitated Diffusion in Living Cells. *Science* **2012**, *336*, 1595-1598.
34. Mahmutovic, A;, Berg, O.G.; Elf, J. What matters for Lac repressor search in vivo - sliding, hopping, intersegment transfer, crowding on DNA or recognition? *Nucl. Acids Res.* **2015**, *43*, 3454-3464.
35. Cuculis, L.; Abil, Z.; Zhao, H.; Schroeder, C.M. Direct observation of TALE protein dynamics reveals a two-state search mechanism. *Nat. Commun.* **2015**, *6*.
36. Zandarashvili, L.; Esadze, A.; Vuzman, D.; Kemme, C.A.; Levy, Y.; Iwahara, J. Balancing between affinity and speed in target DNA search by zinc-finger proteins via modulation of dynamic conformational ensemble. *Nucl. Acid Res.* **2015**, *112*, E5142-E5149.
37. Reingruber, J.; Holcman, D. Transcription factor search for a DNA promoter in a three-state model. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **2011**, *84*, 020901-020904.
38. Koslover, E. F.; de la Rosa, M. A. D.; Spakowitz, A. J. Theoretical and computational modeling of target-site search kinetics in vitro and in vivo. *Biophys. J.* **2011**, *101*, 856-865.
39. Chu, X.; Liu, F.; and Maxwell, B.A.; Wang, Y.; Suo, Z.; Wang, H.; Han, W.; Wang, J. Dynamic conformational change regulates the protein-DNA recognition: an investigation on binding of a Y-family polymerase to its target DNA. *PLoS Comput. Biol.* **2014**, *10*, e1003804.
40. Townson, S.A.; Samuelson, J.C.; Bao, Y.; Xu, S.-Y.; Aggarwal, A.K. BstYI Bound to Noncognate DNA Reveals a "Hemispecific" Complex: Implications for DNA Scanning. *Structure* **2007**, *15*, 449-459.
41. Bauer, M.; Rasmussen, E.S.; Lomholt, M.A.; Metzler, R. Real sequence effects on the search dynamics of transcription factors on DNA. *Sci. Rep.* **2015**, *5*, 10072.
42. Brackley, A.A.; Cates, M.A.; Marenduzzo, D. Facilitated diffusion on mobile DNA: Configurational traps and sequence heterogeneity. *Phys. Rev. Lett.* **2012**, *109*, 168103.
43. Afek, A.; Sela, I., Musa-Lempel, N.; Lukatsky, D.B. Nonspecific transcription-factor-DNA binding influences nucleosome occupancy in yeast. *Biophys. J.* **2011**, *101*, 2465-2475.
44. Afek, A.; Lukatsky, D.B. Nonspecific protein-DNA binding is widespread in the yeast genome. *Biophys. J.* **2012**, *102*, 1881-1888.
45. Afek, A.; Lukatsky, D.B. Positive and negative design for nonconsensus protein-DNA binding affinity in the vicinity of functional binding sites. *Biophys. J.* **2013**, *105*, 1653-1660.
46. Afek, A.; Schipper, J.L.; Horton, J.; Gordan, R.; Lukatsky, D.B. Protein-DNA binding in the absence of specific base-pair recognition. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 17140-17145.
47. Le, D.D.; Shimko, T.C.; Aditham, A.K.; Keys, A.M.; Longwell, S.A.; Orenstein, Y.; Fordyce, P.M. Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E3702-E3711.
48. Marcovitz, A.; Levy, Y. Obstacles may facilitate and direct DNA search by proteins. *Biophys. J.* **2013**, *104*, 2042-20152.
49. Gomez, D.; Klumpp, S. Facilitated diffusion in the presence of obstacles on the DNA. *Phys. Chem. Chem. Phys.* **2016**, *18*, 11184-11192.