Soft Matter

PAPER

View Article Online
View Journal | View Issue

Cite this: *Soft Matter*, 2019, **15**, 5255

Facilitation of DNA loop formation by protein–DNA non-specific interactions†

Jaeoh Shin [a] and Anatoly B. Kolomeisky *[abc]

Complex DNA topological structures, including polymer loops, are frequently observed in biological processes when protein molecules simultaneously bind to several distant sites on DNA. However, the molecular mechanisms of formation of these systems remain not well understood. Existing theoretical studies focus only on specific interactions between protein and DNA molecules at target sequences. However, the electrostatic origin of primary protein–DNA interactions suggests that interactions of proteins with all DNA segments should be considered. Here we theoretically investigate the role of non-specific interactions between protein and DNA molecules on the dynamics of loop formation. Our approach is based on analyzing a discrete-state stochastic model *via* a method of first-passage probabilities supplemented by Monte Carlo computer simulations. It is found that depending on a protein sliding length during the non-specific binding event three different dynamic regimes of the DNA loop formation might be observed. In addition, the loop formation time might be optimized by varying the protein sliding length, the size of the DNA molecule, and the position of the specific target sequences on DNA. Our results demonstrate the importance of non-specific protein–DNA interactions in the dynamics of DNA loop formations.

Received 2nd April 2019,
Accepted 7th June 2019

DOI: 10.1039/c9sm00671k

rsc.li/soft-matter-journal

1 Introduction

Many biological phenomena involve the formation of complex topological structures, which are typically made of protein and nucleic acid biopolymers.[1] In most cases, this is a result of proteins binding simultaneously to spatially distant specific target sites on DNA, which leads to the appearance of DNA loops.[2,3] Specific biological processes with the formation of DNA loops include gene regulation and gene rearrangements *via* site-specific recombination.[4–8] Due to its fundamental importance in natural systems, many theoretical models were proposed to describe the loop formation process in polymer systems.[9–15] It also was extensively studied experimentally using various techniques.[15–19] In addition, many recent investigations considered the loop formation in biologically relevant settings, such as in crowded environment,[19–21] in confined medium[22,23] and in the presence of non-equilibrium fluctuations.[24] However, many aspects of the dynamics of loop formation remain not clarified.

While the molecular mechanism of the DNA loop formation by multi-site proteins is not fully understood, it is reasonable to assume that the protein molecule that has several DNA binding sites first attaches to one of the specific sites on DNA, and subsequently it associates to the other sites. In the majority of previous theoretical studies, it was assumed that the protein interacts only with the specific target sequences on DNA.[25,26] However, as the dominating interaction between the protein and DNA is of the electrostatic origin,[27] it seems reasonable to suggest that the protein–DNA non-specific interactions might also be important. In this scenario, the protein already bound to DNA at one site can bind to a random site of the DNA, forming a transient loop, and the protein then diffuses (slides) along the strand searching for the target site. If the target is not found, the protein dissociates and the process is repeated until the target sequence is located. Indeed, this idea is known as a facilitated diffusion in the process of protein search for a target sequence, and it was shown to be important for single-site proteins that do not form DNA loops. The combination of three-dimensional (3D) diffusion in bulk and one-dimensional (1D) sliding can dramatically enhance the effective protein–DNA association rates.[28–35] The facilitated diffusion in biologically systems has been studied extensively in the past several decades, and it is reviewed, for instance, in ref. 4 and 36–39.

Recently, we theoretically investigated the role of transient DNA looping on the search dynamics for specific targets on DNA by multi-site proteins.[40] It was shown using analytical calculations and computer simulations that the formation of DNA loops might accelerate the overall search process. However, the role of the

[a] *Department of Chemistry, Rice University, Houston, Texas, 77005, USA. E-mail: tolya@rice.edu*
[b] *Department of Chemical and Biomolecular Engineering, Rice University, Houston, Texas, 77005, USA*
[c] *Center for Theoretical Biological Physics, Rice University, Houston, Texas, 77005, USA*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9sm00671k

protein sliding in the context of polymer loop formation has not been studied so far. At the same time, experiments clearly show that proteins might translocate along the DNA chain while being in the looped conformation.[41]

In this paper, we present a theoretical approach to investigate the protein-mediated loop formation kinetics, which also directly incorporates the sliding along the DNA chain. Our main goal is to develop a minimal theoretical model to clarify the role of DNA looping in these complex processes. It is assumed that the protein molecule has two DNA-binding sites, and one of them is already bound to the end of the DNA molecule. It remains bound all the time while the search for the second target sequence is taking place. Because the protein is already bound to DNA at one site, the non-specific protein–DNA interactions depend on the loop size. Therefore, one cannot use theoretical approaches developed for the binding of the single-site protein to target sites.[4,36,38] To explain the dynamics of the system, we take into account the free energy cost of the loop formation. It is found that depending on the protein sliding length, which is the average length that the protein moves along DNA during one binding cycle, the loop formation process shows different dynamic behaviors. Moreover, the loop formation time can be minimized at an intermediate value of the sliding length. The specific location of the target site and the length of the DNA segment also influence the search process. Our results indicate that the non-specific protein–DNA interactions play an essential role in the polymer loop formation.

The paper is organized as follows. The theoretical model is described in the Section 2, and analytic results in limiting cases are presented in Section 3. The general results are presented and discussed in Section 4, and we summarize and conclude in Section 5.

2 Theoretical model

Let us consider a process of the protein searching for a target sequence on DNA as illustrated in Fig. 1 top. It is assumed here that the protein is already bound to one end of the DNA chain (and remains there for a long time) while exploring the space to find the second binding site on the same strand. This is a reasonable assumption because specific protein–DNA interactions are very strong.[4] As we aim to understand the role of non-specific protein–DNA interactions on the DNA loop formation with a minimal model, we have few simplifications of real biological systems. Firstly, we assume that during the sliding motion of the protein along the DNA, the chain segments of the loop can quickly relax to the equilibrium. Since this relaxation time depends on the length of DNA L as $T_r \sim L^2$ for the Rouse chain,[42] this assumption will break down for the long chains. Secondly, we assume that the consecutive non-specific binding sites are uncorrelated as typically done in the literature.[34,40,43–45] This assumption is valid if the chain relaxation time is shorter than the non-specific binding rate $k_{on}(n)$. We take the chain length L and kinetic rates that satisfy these two assumptions. Thirdly, as the protein slides along the DNA helix, it can induce super-coiling and twist of the DNA;[46] however, in our minimal

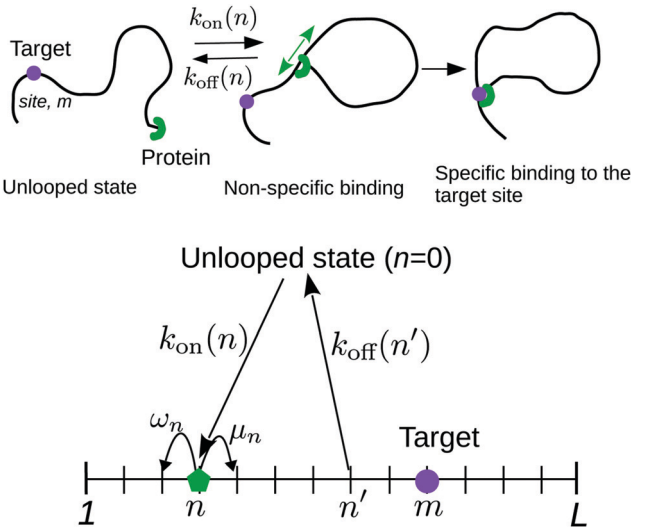

Fig. 1 (top) Schematic view of the DNA looping process. Here the multi-site protein molecule (green), already bound to one end of the DNA, is searching for a target site (violet). (bottom) The discrete-state stochastic model of the search process.

theoretical approach we neglect this. This assumption might be reasonable for some systems such as DNA with a nick,[47,48] and it is also supported by the fact that no supercoiling was observed in experiments on EcoRII proteins.[41] Lastly, we also neglect the twist energy of the DNA, which might be required for the protein to match the binding positions on DNA. The loop formation free energy without twist energy would be the upper bound of the looping time. For more extensive discussions on this issue, we would like to refer, for instance, to ref. 49.

With these assumptions, the system can be viewed as $L + 1$ discrete states, see Fig. 1 bottom. As proteins bind to multiple DNA base-pairs (bp) simultaneously, we set the unit lattice size as 10 bp. This is based on the fact that typical sizes of the specific protein binding sites on DNA are ranging from 6 to 15 base pairs. If the protein is in the state $1 \leq n \leq L$ it means that the DNA loop of size n is formed and the DNA segment of length $L - n$ is free. The final target sequence is in the state $m \neq 0$. The state $n = 0$ corresponds the protein molecule unbound from the DNA chain (but still connected to the DNA end site). The protein can non-specifically associate to the state n with a rate $k_{on}(n)$, while the dissociation rate is equal to $k_{off}(n)$ (Fig. 1). The non-specific binding energy (enthalpic contribution) is given by ε ($\varepsilon < 0$ corresponds to attraction and $\varepsilon > 0$ corresponds to repulsion). This also means that we are neglecting the effect of DNA sequence heterogeneity, although it might be relevant.[43] In the non-specifically bound state, the protein can diffuse along the chain with the position-dependent rates that also depend on the direction of the motion (see Fig. 1). The process of reducing the size of DNA loop is taking place with a rate w_n, while increasing the loop size is associated with a rate μ_n.

Assuming that the relaxation of the DNA chain is taking place faster than any other processes in the system, the dynamics is governed by changes in the free energy. At realistic cellular conditions, a significant fraction of the free energy is due to the

formation and breaking of DNA loops. The free energy cost of forming a loop of size n (in the unit of thermal energy, k_BT) is[40]

$$G_0(n) = \frac{A}{n} + \alpha \log[n]. \quad (1)$$

In this expression, the first term accounts for the polymer bending energy and the second term describes the entropic cost of the loop formation. The coefficient A is proportional to the bending stiffness of the DNA chain. For instance, for the case of a circular loop, $A = 2\pi^2 l_p$, where l_p is the persistence length of DNA ($\simeq$150 bp). The exponent α is related to the scaling exponent for the radius of gyration, and for the ideal Gaussian chain it is equal to $\alpha = 3/2$. Here we ignore the excluded volume effects and the bending stiffness on the corrections to the entropy of loop formation.[25,26,50] Nevertheless, it is expected that our simplified model still should account for the main physical features of the search process with loop formation.

The total free energy cost of loop formation should also include the enthalpic contribution due to the protein–DNA non-specific binding energy, and the final expression is given by

$$G(n) = G_0(n) + \varepsilon = \frac{A}{n} + \alpha \log[n] + \varepsilon. \quad (2)$$

The specific example of the free-energy profile is given in Fig. 2. This allows us to evaluate the position-dependent binding and unbinding rates:

$$k_{on}(n) = k_{on}^{(0)} \exp[-\theta G_0(n)], \quad (3)$$

and

$$k_{off}(n) = k_{off}^{(0)} \exp[(1-\theta) G_0(n)], \quad (4)$$

where $k_{on}^{(0)}$ and $k_{off}^{(0)}$ are association and dissociation rates, respectively, in the absence of loop formation. The parameter $0 \le \theta \le 1$ reflects the relative contribution of free energy changes to the binding and unbinding rates. It also gives the position of the transition state for protein associating to the DNA chain. Since the microscopic details of this process are not well known, we take three different values of θ in our study ($\theta = \{0;0.5;1\}$) to cover all ranges of parameters. Detailed balance arguments suggest that binding/unbinding rates are related to each other as

$$\frac{k_{on}^{(0)}}{k_{off}^{(0)}} = \exp(-\varepsilon), \quad (5)$$

which leads to

$$\frac{k_{on}(n)}{k_{off}(n)} = \exp[-G(n)]. \quad (6)$$

The physical interpretation of eqn (5) and (6) is simple. If the formation of the DNA loop lowers the free energy of the system, then the corresponding association rate is faster and breaking the loop is a slower process. But if the formation of the DNA loop increases the free energy of the system, then the corresponding binding rate is slow while the unbinding transition is fast.

The direction-dependent diffusion of protein along the DNA chain is affected by the free-energy changes associated with varying the size of DNA loops. More specifically, we can write

$$\mu_n = \mu_0 \exp[-\theta_t \Delta G(n+1)]; \quad \omega_n = \mu_0 \exp[(1-\theta_t)\Delta G(n)], \quad (7)$$

where μ_n (ω_n) is the sliding rate that makes the loop size increasing (decreasing) by one unit length, and

$$\Delta G(n) \equiv G(n) - G(n-1) = G_0(n) - G_0(n-1), \quad (8)$$

is the associated free-energy difference. The sliding rate μ_0 describes the diffusion in the absence of the loop formation, *i.e.*, in the flat free-energy profile. In the following calculations we set $\mu_0 = 60\ s^{-1}$ (or 6×10^3 bp^2 s^{-1} in real units) from the experiments on the EcoRII proteins.[41] With this value, the assumption that the sliding motion is slower than the chain relaxation time T_r is valid up to $\simeq$5 kbp long DNA (or $L = 500$). We assume that $\theta_t = 0.5$ for symmetry reason. In addition, the sliding rates are related to each other *via* the detailed balance arguments,

$$\frac{\mu_{n-1}}{\omega_n} = \exp[-\Delta G(n)]. \quad (9)$$

This expression implies that the protein sliding is faster in the direction of lowering the free energy of the system, while the sliding is slower in the direction of increasing the free energy of the system.

To analyze the dynamics of the polymer loop formation by the multi-site protein, a method of first-passage probabilities, which have been successfully employed in studies of various protein search processes for target sites,[34,40,43–45] is utilized. We define a first-passage time probability density function $F(n,t)$, which describes the probability to reach the target site m at time t given that it was at the site n at $t = 0$. The state $n = 0$ is the unbound state (see Fig. 1). The temporal evolution of the first-passage probabilities $F(n,t)$ follows the backward master equations,[34,40]

$$\frac{\partial F(n,t)}{\partial t} = -[\mu_n + \omega_n + k_{off}(n)]F(n,t) + \mu_n F(n+1,t) + \omega_n F(n-1,t) + k_{off}(n)F(0,t), \quad (10)$$

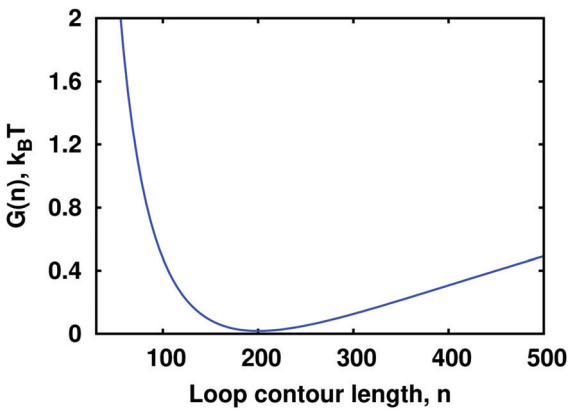

Fig. 2 Free-energy cost of the DNA loop formation as a function of the loop contour length n. In our calculations, we set $A = 300$ (3 kbp) and $\alpha = 3/2$.

for $n \neq m$. For $n = 0$ state we have,

$$\frac{\partial F(0,t)}{\partial t} = -F(0,t)\sum_{n=1}^{L} k_{on}(n) + \sum_{n=1}^{L} k_{on}(n)F(n,t). \quad (11)$$

Additionally, the initial condition implies that $F(m,t) = \delta(t)$, which means that if the protein is at the site m at time $t = 0$, the process will end immediately. Calculating explicitly these first-passage probabilities should provide a full dynamic description of the system.[34,40]

3 Dynamics in limiting cases

Although we were not able to determine the first-passage probabilities explicitly in general situations, there are several limiting cases that can be solved analytically. They provide important physical insights on the role of non-specific interactions in DNA loop formation.

3.1 No desorption limit, $\mu_0 \gg k_{off}(n)$

If the DNA looped states are energetically strongly favorable ($G(n) \ll -1\ k_BT$), then the protein will bind to DNA and it will not dissociate until the target site is found. It can be realized, for example, if the protein–DNA non-specific interactions are strongly attractive. In this case, we can approximate eqn (10) as

$$\frac{\partial F(n,t)}{\partial t} = -(\mu_n + \omega_n)F(n,t) + \mu_n F(n+1,t) + \omega_n F(n-1,t), \quad (12)$$

and we call it a "no desorption limit". In order to solve it together with eqn (11), we apply the Laplace transformations, $\tilde{F}(n,s) \equiv \int_0^\infty F(n,t)\exp(-st)\mathrm{d}t$, where s is the Laplace variable. Then eqn (12) transforms into

$$(s + \mu_n + \omega_n)\tilde{F}(n,s) = \mu_n\tilde{F}(n+1,s) + \omega_n\tilde{F}(n-1,s) \quad (13)$$

Correspondingly, eqn (11) now can be written as

$$\left[s + \sum_{n=1}^{L} k_{on}(n)\right]\tilde{F}(0,s) = \sum_{n=1}^{L} k_{on}(n)\tilde{F}(n,s) \quad (14)$$

The most relevant quantity to describe the dynamics in the system is the mean search time T_n, which is defined as the average time to reach the target site m when the initial binding site is at n,

$$T(n) = \int_0^\infty tF(n,t)\mathrm{d}t = -\left.\frac{\partial\tilde{F}(n,s)}{\partial s}\right|_{s=0} \quad (15)$$

Correspondingly, the mean search time from the unbound state, which we label as a looping time, is given by

$$T = \int_0^\infty tF(0,t)\mathrm{d}t = -\left.\frac{\partial\tilde{F}(0,s)}{\partial s}\right|_{s=0}. \quad (16)$$

With the help of eqn (14) it can be found that

$$T = \frac{1}{k_{on}^{(S)}} + \sum_{n=1}^{L}\left(\frac{k_{on}(n)}{k_{on}^{(S)}}\right)T(n), \quad (17)$$

where $k_{on}^{(S)} \equiv \sum_{n=1}^{L} k_{on}(n)$ is the total binding rate of the protein molecule to all DNA sites. Since the rate of the chemical reaction between the protein and DNA molecules is expected to be proportional to the number of binding sites on DNA, the total association rate $k_{on}^{(S)}$ to DNA should also increase with the DNA length, at least for not too long DNA chains. The physical meaning of eqn (17) is the following. The total mean search time to reach the target from the unbounded state is a sum of two terms. The first terms describes the average time to bind to any site on DNA, while the second term is the average time to reach the target from the site n, $T(n)$ multiplied by the probability that the protein will associate to the site n from the unbounded state. The coefficient $\frac{k_{on}(n)}{k_{on}^{(S)}}$ gives this probability.

To evaluate the looping time we need to calculate $T(n)$. This can be done in the following way. In this limit, the search process in the looped conformation can be viewed as a one-dimensional inhomogenous random walk, for which the first-passage times have been explicitly analyzed in terms of position-dependent hopping rates.[51] We utilize these results for calculating $T(n)$ in eqn (17).

3.2 No sliding limit, $\mu_0 \ll k_{off}(n)$

Another situation that can be solved analytically corresponds to the limiting case when the protein can form the transient DNA loops, but it cannot slide in the looped states. This can be associated with a very large free energy for being in the looped state ($G(n) \gg 1\ k_BT$), and it might be realized for strong non-specific protein–DNA repulsive interactions. In this case, we can approximate eqn (10) as

$$\frac{\partial F(n,t)}{\partial t} = -k_{off}(n)F(n,t) + k_{off}(n)F(0,t), \quad (18)$$

and we call it a "no sliding limit". Since this case has been fully analyzed previously,[40] here we briefly recapitulate the main results. Eqn (10) in this limit is written as

In the Laplace domain, it transforms into

$$[s + k_{off}(n)]\tilde{F}(n,s) = k_{off}(n)\tilde{F}(0,s). \quad (19)$$

With eqn (14) and the initial condition $\tilde{F}(m,s) = 1$, one can obtain the following expression,

$$\tilde{F}(0,s) = \frac{k_{on}(m)}{s + f(s)}, \quad (20)$$

where the auxiliary function $f(s)$ is given by

$$f(s) \equiv k_{on}(m) + \sum_{i\neq m}\frac{sk_{on}(i)}{s + k_{off}(i)}. \quad (21)$$

Then the mean search time T can be easily computed, yielding

$$T = \frac{1 + \sum_{i\neq m}\frac{k_{on}(i)}{k_{off}(i)}}{k_{on}(m)}. \quad (22)$$

This results underlines the fact that, on average, the protein should visit every site on DNA before the target can be found.

3.3 No looping effect limit, $G_0(n) = 0$

There is one more limiting case that can be explicitly analyzed. If the free-energy associated with the formation of loops are relatively small, $|G_0(n)| \leq k_BT$, then the search process is taking place in effectively flat free-energy profile. This was extensively investigated before for describing the single-site protein search.[34,43,45,52] Because in this case the transient formation of loops does not influence much the free energy of the system, we call it a "no looping effect limit".

In this case, all transition rates become position independent, $k_{on}(n) = k_{on}$, $k_{off}(n) = k_{off}$ and $\mu_n = \omega_n = \mu$. Then it can be shown that the mean search time is given by

$$T = \frac{1}{k_{on}}\frac{L}{S} + \frac{1}{k_{off}}\left(\frac{L}{S} - 1\right), \quad (23)$$

where a new parameter S describes the number of sites visited during each binding event, and it depends on transition rates k_{off} and μ, see ref. 34 and 52 for more details. Eqn (23) also has a clear physical meaning. There are L/S protein bindings to DNA ($1/k_{on}$ is the time for each event), and there are $L/S - 1$ unbindings ($1/k_{off}$ is the time for each event). The number of dissociations is less than the number of associations by one because the after last binding event the target will be found.

4 Results

Now let us consider a general search problem for the two-site protein molecule already bound to DNA at the end of the chain to locate the second target sequence. We investigate it using Monte Carlo computer simulations with the Gillespie algorithm for various sets of parameters.[53] To describe the dynamics in the system, we introduce a new parameter $\lambda_0 \equiv \sqrt{\mu_0/k_{off}}$, which we call a scanning length. It corresponds to a distance that the protein would explore while sliding along the DNA chain if diffusion rate at all sites will be the same and equal to μ_0 and the dissociation rate will be the same and equal to k_{off}. The actual scanning length depends on the position of the binding, but it is always proportional to λ_0. Thus, the parameter λ_0 is a convenient measure of non-specific protein–DNA interactions as well as the measure of the stability of the transient loop formation. The larger the scanning length, the stronger is non-specific protein–DNA interaction and the longer the system is found in the looped conformation.

The results of Monte Carlo computer simulations, as well as analytical predictions in limiting cases, are shown in Fig. 3, where the looping time as a function of the scanning length is presented for different values of the parameter θ. One can see that in most cases, varying θ does not much influence the dynamics of the loop except modifying the position of the most optimal looping times. However, changing θ might affect the looping dynamics in some cases, as shown in Fig. 7 in the Appendix.

Analyzing Fig. 3, three dynamic regimes can be identified. If the scanning length is very small, $\lambda_0 < 1$, the protein occasionally binds to the DNA chain, but it cannot slide. This is a 3D search

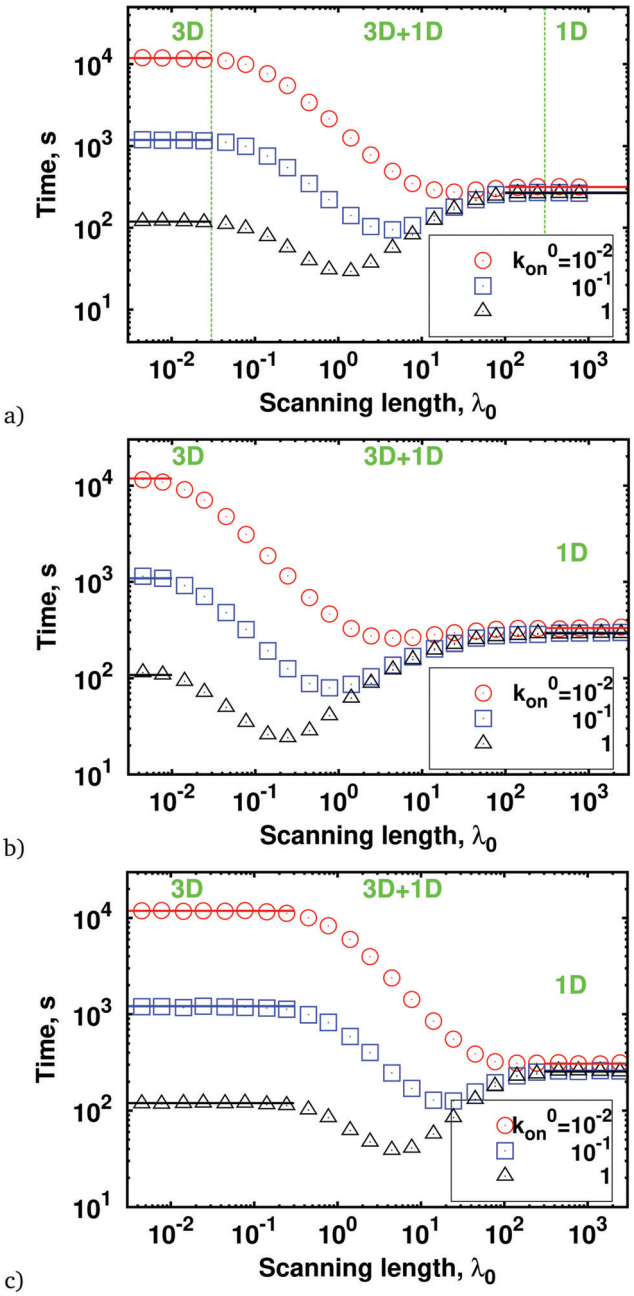

Fig. 3 Looping time T as a function of the scanning length λ_0 for three different values of $k_{on}^{(0)}$ and for different values of the parameter θ: (a) $\theta = 0.5$, (b) $\theta = 0$ and (c) $\theta = 1$. The target is located at the end of the chain at $n = L$. Simulation data are shown in symbols and the solid lines are from theoretical predictions. For calculations we use $\mu_0 = 60\ s^{-1}$ and $L = 300$ (3 kbp).

dynamic regime from the point of view of the protein molecule although it is always connected to DNA. It was explicitly investigated before.[40] This also corresponds to the no sliding limit, considered above. Excellent agreement between analytical results and computer simulations in this regime shows that our theoretical arguments correctly capture the main physics in this regime. In the opposite limit of $\lambda_0 > L$ (L is the length of the DNA chain), once the protein binds to the DNA, it remains on it until it reaches the target

site. This is effectively a 1D dynamic process, and the search time T is insensitive to the binding rate k_{on} because the association occurs only once. Our analytical predictions also perfectly agree here with computer simulations. It is interesting to note that the dynamics in this regime might be faster or slower in comparison with $\lambda_0 < 1$ regime, depending on the association rates. If the binding rats are slow, then the 1D search is faster than 3D search because it needs only one binding event to reach DNA. However, when the binding rates are fast 3D search is more efficient since in the 1D regime the protein might be trapped by repeatedly moving over the sites that are far away from the target.

The most interesting behavior is observed in the intermediate dynamic regime for $1 < \lambda_0 < L$, which we label as 3D + 1D search (see Fig. 3). In this case, the protein binds to DNA, slides some distance and dissociates, and then the cycle is repeated several times until the target is found. Computer simulations show that the search dynamic can be optimized in this dynamic phase. The minimum in the search time is observed for some intermediate scanning lengths. This physically corresponds to the situation when the protein is not trapped for a long time in sliding but can dissociate to start the search at a new location, but, at the same time, it is not doing too many binding/unbinding events that might slow down the dynamics. It seems that this regime is the most realistic for typical biological systems.

Our theoretical approach allows us to quantify the role of transient loop formation in the overall search process. We compare the search time in the presence and in the absence of loop formation free energy $G_0(n)$ as a function of the scanning length λ_0 in Fig. 4. We consider two target positions, $m = 200$, where the free energy is minimal, and $m = 50$ where the free energy is much higher (see Fig. 2). For the case of $m = 200$, the search time T (shown in blue squares) decreases in comparison with the case in the absence of the loop formation (blue dashed line). The main reason is that after the protein binds anywhere on DNA, its motion to the target is accelerated because it always involves moving down along the free-energy profile. In addition, the direct binding to the target site at the minimum of the free-energy surface is also the fastest, as one can see from eqn (3). These processes facilitate the dynamics significantly for all the search regimes. On the other hand, for the case of $m = 50$ (red circles), both the binding rate and the sliding rate toward that the target are lower. Binding of the protein to any site $n > m$ means that near the target the sliding will be very slow due to moving against the free-energy potential. Therefore the search time increases compared to the case without loop formation (red line). These findings indicate that the loop formation might be an important tool for controlling the target search kinetics.

Because the free-energy profile generally is strongly position-dependent (see Fig. 2), it is reasonable to expect that the search dynamics will be sensitive to the location of the target. We investigated this effect, and the results are presented in Fig. 5 for different scanning lengths. As expected, the looping times depend on the target position m, however, this dependence is also determined by the nature of the dynamic regime. For small scanning lengths ($\lambda_0 < 1$, 3D search regime) the protein does not slide along the DNA chain and the probability of reaching the specific site on DNA is fully determined by the free-energy profile as given by eqn (6). The sites that are closer to the free-energy minimum are more probable to be explored first. For this reason, the dependence of the search time in 3D dynamic regime follows almost exactly the free-energy profile in Fig. 2. A different behavior is observed for large scanning lengths ($\lambda_0 \geq L$, 1D search regime) when the protein associates only once with the DNA chain. In this case, the target can be achieved mainly *via* 1D diffusion. Then the average distance between the target and the location where the protein binds first to DNA determines the overall search time. For this reason, the minimum search time is closer to $m = L/2$ position due to symmetry. For the intermediate 3D + 1D dynamic regime, the overall search is faster and the dependence on m is weaker.

In our system, the process is taking place *via* the formation of transient polymer loops. But it is easier to form the loop for

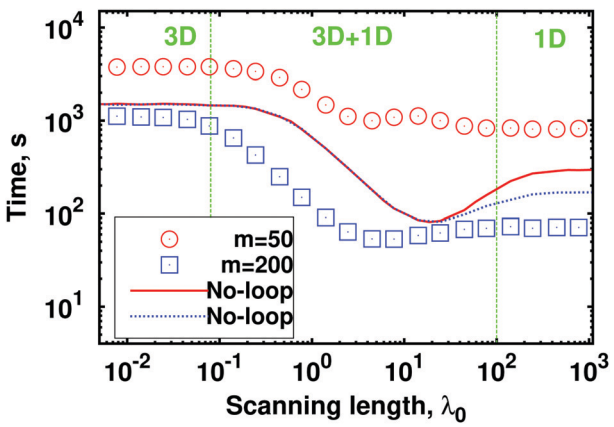

Fig. 4 Looping time T as a function of the scanning length λ_0 in the presence (symbols) and in the absence (lines) of looping free energy contribution. The target is located at $m = 50$ (red) and $m = 200$ (blue). For calculations we use $k_{on}^{(0)} = 0.1\ s^{-1}$, $\mu_0 = 60\ s^{-1}$, $\theta = 0.5$, $L = 300$ (3 kbp). In Fig. 7 in the Appendix, we show the results with $\theta = 0$ and 1.

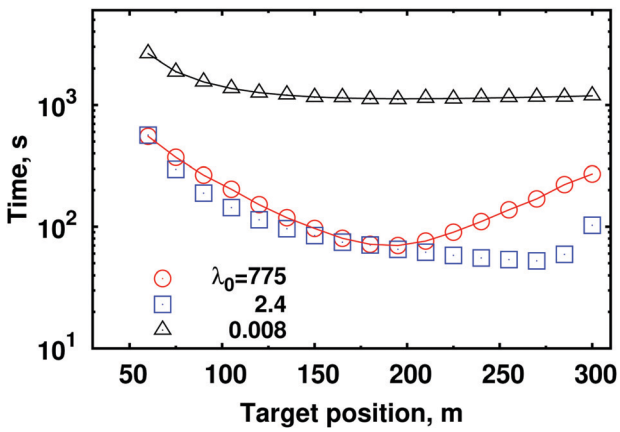

Fig. 5 Looping time T as a function of the target position m for different values of the scanning length λ_0. Simulation data are shown as symbols and theoretical predictions are shown as lines. For calculations we use $k_{on}^{(0)} = 0.1\ s^{-1}$, $\mu_0 = 60\ s^{-1}$, $\theta = 0.5$, and $L = 300$ (3 kbp). In Fig. 8 in the Appendix, we show the results with $\theta = 0$ and 1.

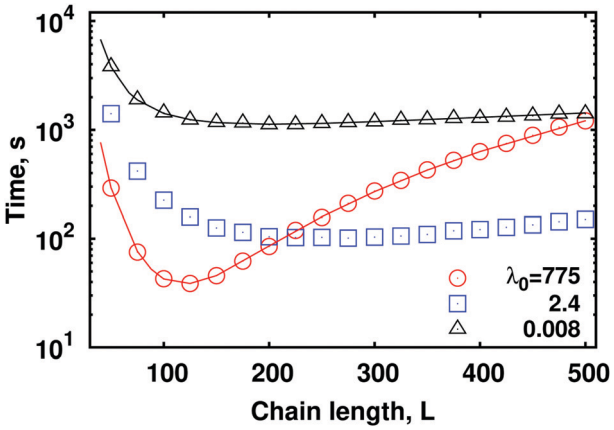

Fig. 6 Looping time T as a function of the chain length L for different values of λ_0. Simulation data are shown in symbols, and theoretical predictions are shown in solid lines. Here we take the parameters $k_{on}^{(0)}$ = 0.1 s^{-1}, μ_0 = 60 s^{-1}, and the target is located at the end of the chain $m = L$.

longer DNA segments than for the shorter chains. These arguments suggest that the DNA length L might also be an important factor in the overall search process. We tested this idea, and the results are presented in Fig. 6. Here we show the looping time T for three different values of λ_0. For all three cases, the looping time T showed a minimum when the chain length corresponds to the loop size of the minimum in the free energy profile. The analytical theory for the 3D search (black line) matches excellently with the simulation data. The theory of 1D search (red line) is also in a good agreement with Monte Carlo simulations. The presented results clearly show that the looping dynamics can be optimized by varying the DNA chain length.

Although our aim was to develop a minimal theoretical model to describe the role of DNA looping in protein search phenomena, it is important to discuss the relevance of our theoretical calculations for real biological systems. First of all, we chose the sliding rate in the looped state to be μ_0 = 60 s^{-1}, which is comparable to the experimental value of the diffusion constant of the EcoRII protein, $D \simeq 7.2 \times 10^{-4}$ μm^2 s^{-1}, as measured in recent experimental studies.[41] In addition, in the same experiments,[41] the DNA fragment of 810 bp size was considered, and in our calculations we only looked at the DNA chains less than 5000 base pairs. Since other important rates, such as the association and dissociation rates k_{on} and k_{off} are not available yet, we used a range of parameters in our calculations. Furthermore, the search times calculated in our model for realistically most relevant parameters (3D + 1D regime) are of the order 10–100 seconds, which again agrees well with experimental observations.[41] These arguments show that the parameters chosen in our theoretical framework probably are not very different from the parameters found in biological systems. Then, our model can make several quantitative predictions that can be tested in experiments. To be more specific, Fig. 3 predicts how the protein search times change for different sets of association and dissociation rates, which can be changed, for example, by varying the ionic strength. Fig. 5 shows that changing the position of the target will affect the search times. Fig. 6 gives the prediction on how the search dynamics is influenced by varying the DNA chain length.

5 Summary and conclusions

We presented a theoretical analysis of the formation of a protein–DNA complex with a loop using analytical calculations and Monte Carlo computer simulations. We specifically considered two-site proteins that are already bound to DNA at one site that are searching for the second target site. A discrete-state stochastic model that takes into account the free-energy cost of the transient loop formation is utilized in our analysis. It is found that the non-specific protein–DNA interactions strongly influence the loop formation in the final complex. Three different dynamic regimes are identified depending on the relative sliding lengths and the size of the DNA chain. When the protein cannot slide along the DNA, the search is effectively three-dimensional with the formation and breaking of transient loops at each site. This corresponds to weak protein–DNA non-specific interactions. In the opposite limit of very strong non-specific interactions, after the first association to the DNA chain the protein slides continuously until the target is found. This is effectively a one-dimensional search. For the intermediate range of protein–DNA interactions, the slidings alternate with breaking and making transient polymer loops. It is found that the dynamics can be optimized (fastest) in this 3D + 1D search regime. Our analysis shows the importance of the transient loop formation, and there is a range of parameters when it can even show faster dynamics in comparison with the case without loop formation. We also found that due to the free-energy changes associated with the formation of transient loops at different sites, the location of the target sequence affects the dynamics. In addition, the length of the DNA segment is another important factor in the formation of protein–DNA complexes due to different free-energy cost of making loops of different sizes. All these observations clearly show that the non-specific protein–DNA interactions are important in the formation of protein–DNA complexes with topological features such as loops.

Our theoretical approach is able to describe the main features of the non-specific interaction assisted DNA looping by multi-site proteins. However, it is important to discuss its limitations. Here we do not take into account the sequence heterogeneity of the DNA segments, while the previous study showed that this can strongly affect the protein search dynamics without loops formation.[43] Besides, our theoretical model neglects protein and DNA conformational fluctuations that can also play an important role in the search process.[45,54–56] Furthermore, real cellular systems are very crowded, and the presence of other molecules bound to DNA could prevent the search dynamics, and it is not accounted for in our current model. Including those effects would be necessary to fully understand real biological systems, and will be important directions of the future study. Despite these limitations, it is reasonable to say that our theoretical method provides a consistent physical picture of the DNA loop formation with the help of

non-specific protein–DNA interactions. The main advantage of our approach is quantitative predictions that can be tested in experiments. Therefore, it will be important to validate our results using various experimental techniques. For instance, it would be possible to control the protein–DNA interactions by changing the salt concentration.[57] We expect that for low concentrations the 1D sliding would dominate, whereas for high concentrations the 3D search will be the most important part of the looping mechanism.

Conflicts of interest

There are no conflicts to declare.

Appendix

In this Appendix, we present supplementary figures on the effects of different θ values. In Fig. 7, the looping time as a function of λ_0 is shown for two target positions m with $\theta = 0$ (top) and $\theta = 1$ (bottom). For the case of $\theta = 1$, the behavior is similar to $\theta = 0.5$ case. However, for the case of $\theta = 0$ and $m = 50$, the looping time shows a maximum, instead of a minimum, at an intermediate value of λ. This unusual behavior can also be noticed in the case of $\theta = 0.5$ in Fig. 4, although to a much smaller degree. However, the detailed investigation of this observation is out of the scope of this work and it will be a future direction of study.

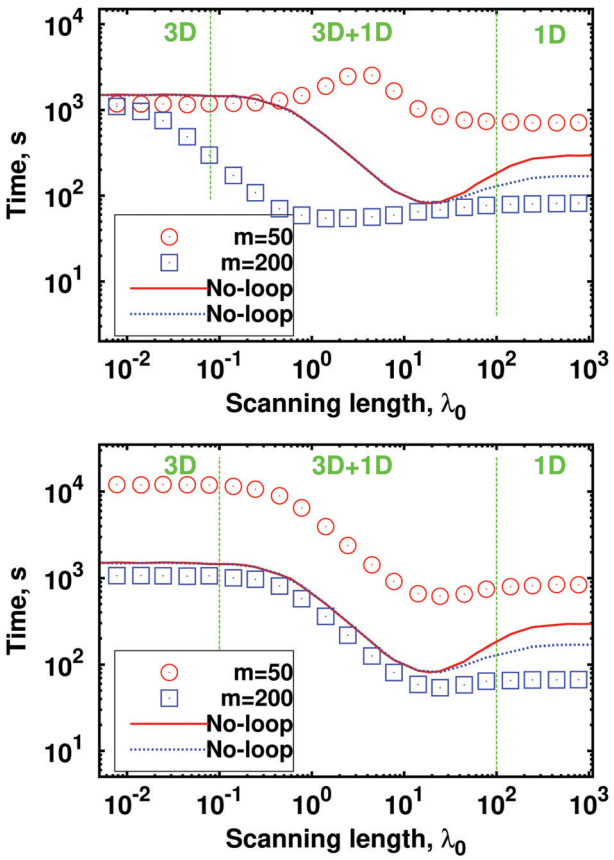

Fig. 7 Looping time T as a function of the scanning length λ_0 in the presence (symbols) and in the absence (lines) of looping free energy contribution. The target is located at $m = 50$ (red) and $m = 200$ (blue). The simulation data with $\theta = 0$ is shown in the top and $\theta = 1$ is shown in the bottom. For calculations we use $k_{on}^{(0)} = 0.1\ s^{-1}$, $\mu_0 = 60\ s^{-1}$, and $L = 300$.

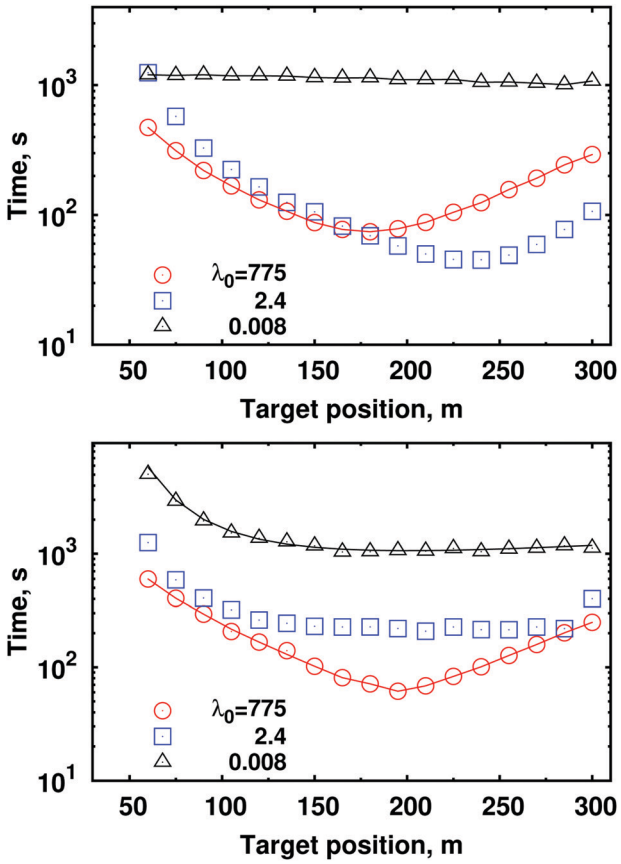

Fig. 8 Looping time T as a function of the target position m for different values of the scanning length λ_0. The simulation data with $\theta = 0$ is shown in the top and $\theta = 1$ is shown in the bottom. We use $k_{on}^{(0)} = 0.1\ s^{-1}$, $\mu_0 = 60\ s^{-1}$, and $L = 300$ (3 kbp).

In Fig. 8 we show the target position dependent looping time for different values of the parameter θ. In this case, the trend remains the same as the $\theta = 0.5$ case.

Acknowledgements

This work was supported by the Welch Foundation (C-1559), by the NSF (CHE-1664218), and by the Center for Theoretical Biological Physics sponsored by the NSF (PHY-1427654). We thank two anonymous referees whose comments helped improve and clarify this manuscript.

References

1 B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Molecular Biology of the Cell*, Garland Science, 4th edn, 2002.
2 K. Matthews, *Microbiol. Rev.*, 1992, **56**, 123–136.

3 R. Schleif, *Annu. Rev. Biochem.*, 1992, **61**, 199–223.
4 S. E. Halford and J. F. Marko, *Nucleic Acids Res.*, 2004, **32**, 3040–3052.
5 N. D. Grindley, K. L. Whiteson and P. A. Rice, *Annu. Rev. Biochem.*, 2006, **75**, 567–605.
6 R.-S. Mani and A. M. Chinnaiyan, *Nat. Rev. Genet.*, 2010, **11**, 819.
7 F. Bushman, M. Lewinski, A. Ciuffi, S. Barr, J. Leipzig, S. Hannenhalli and C. Hoffmann, *Nat. Rev. Microbiol.*, 2005, **3**, 848.
8 A. Cournac and J. Plumbridge, *J. Bacteriol.*, 2013, **195**, 1109–1119.
9 G. Wilemski and M. Fixman, *J. Chem. Phys.*, 1974, **60**, 866–877.
10 G. Wilemski and M. Fixman, *J. Chem. Phys.*, 1974, **60**, 878–890.
11 A. Szabo, K. Schulten and Z. Schulten, *J. Chem. Phys.*, 1980, **72**, 4350–4357.
12 N. M. Toan, G. Morrison, C. Hyeon and D. Thirumalai, *J. Phys. Chem. B*, 2008, **112**, 6094–6106.
13 T. Guérin, O. Bénichou and R. Voituriez, *Nat. Chem.*, 2012, **4**, 568.
14 L. Saiz and J. M. Vilar, *Curr. Opin. Struct. Biol.*, 2006, **16**, 344–350.
15 J.-F. Allemand, S. Cocco, N. Douarche and G. Lia, *Eur. Phys. J. E: Soft Matter Biol. Phys.*, 2006, **19**, 293–302.
16 L. Finzi and J. Gelles, *Science*, 1995, **267**, 378–380.
17 G. Bonnet, O. Krichevsky and A. Libchaber, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 8602–8606.
18 Y.-F. Chen, J. Milstein and J.-C. Meiners, *Phys. Rev. Lett.*, 2010, **104**, 048301.
19 O. Stiehl, K. Weidner-Hertrampf and M. Weiss, *New J. Phys.*, 2013, **15**, 113010.
20 J. Shin, A. G. Cherstvy and R. Metzler, *Soft Matter*, 2015, **11**, 472–488.
21 J. Shin, A. G. Cherstvy, W. K. Kim and R. Metzler, *New J. Phys.*, 2015, **17**, 113008.
22 A. Amitai and D. Holcman, *Phys. Rev. Lett.*, 2013, **110**, 248105.
23 J. Shin, A. G. Cherstvy and R. Metzler, *ACS Macro Lett.*, 2015, **4**, 202–206.
24 J. Shin and W. Sung, *J. Chem. Phys.*, 2012, **136**, 045101.
25 Y.-J. Chen, S. Johnson, P. Mulligan, A. J. Spakowitz and R. Phillips, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 17396–17401.
26 P. J. Mulligan, Y.-J. Chen, R. Phillips and A. J. Spakowitz, *Biophys. J.*, 2015, **109**, 618–629.
27 P. L. Privalov, A. I. Dragan and C. Crane-Robinson, *Nucleic Acids Res.*, 2010, **39**, 2483–2491.
28 A. D. Riggs, S. Bourgeois and M. Cohn, *J. Mol. Biol.*, 1970, **53**, 401–417.
29 O. G. Berg and C. Blomberg, *Biophys. Chem.*, 1976, **4**, 367–381.
30 R. B. Winter and P. H. Von Hippel, *Biochemistry*, 1981, **20**, 6948–6960.
31 P. H. von Hippel and O. G. Berg, *J. Biol. Chem.*, 1989, **264**, 675–678.
32 M. Coppey, O. Bénichou, R. Voituriez and M. Moreau, *Biophys. J.*, 2004, **87**, 1640–1649.
33 M. A. Lomholt, B. van den Broek, S.-M. J. Kalisch, G. J. Wuite and R. Metzler, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 8204–8208.
34 A. Veksler and A. B. Kolomeisky, *J. Phys. Chem. B*, 2013, **117**, 12695–12701.
35 A. Esadze, C. A. Kemme, A. B. Kolomeisky and J. Iwahara, *Nucleic Acids Res.*, 2014, **42**, 7039–7046.
36 L. Mirny, M. Slutsky, Z. Wunderlich, A. Tafvizi, J. Leith and A. Kosmrlj, *J. Phys. A: Math. Theor.*, 2009, **42**, 434013.
37 O. Bénichou, C. Loverdo, M. Moreau and R. Voituriez, *Rev. Mod. Phys.*, 2011, **83**, 81.
38 A. B. Kolomeisky, *Phys. Chem. Chem. Phys.*, 2011, **13**, 2088–2095.
39 M. Sheinman, O. Bénichou, Y. Kafri and R. Voituriez, *Rep. Prog. Phys.*, 2012, **75**, 026601.
40 A. A. Shvets and A. B. Kolomeisky, *J. Phys. Chem. Lett.*, 2016, **7**, 5022–5027.
41 J. L. Gilmore, Y. Suzuki, G. Tamulaitis, V. Siksnys, K. Takeyasu and Y. L. Lyubchenko, *Biochemistry*, 2009, **48**, 10492–10498.
42 M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics*, Oxford University Press, 1988, vol. 73.
43 A. A. Shvets and A. B. Kolomeisky, *J. Chem. Phys.*, 2015, **143**, 245101.
44 J. Shin and A. B. Kolomeisky, *J. Phys. Chem. B*, 2018, **122**, 2243–2250.
45 J. Shin and A. B. Kolomeisky, *J. Chem. Phys.*, 2018, **149**, 174104.
46 E. G. Marklund, A. Mahmutovic, O. G. Berg, P. Hammar, D. van der Spoel, D. Fange and J. Elf, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 19796–19801.
47 R. Yuan, D. L. Hamilton and J. Burckhardt, *Cell*, 1980, **20**, 237–244.
48 N. Crampton, M. Yokokawa, D. T. Dryden, J. M. Edwardson, D. N. Rao, K. Takeyasu, S. H. Yoshimura and R. M. Henderson, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 12755–12760.
49 Y. Zhang, A. E. McEwen, D. M. Crothers and S. D. Levene, *Biophys. J.*, 2006, **90**, 1903–1912.
50 A. Hanke and R. Metzler, *Biophys. J.*, 2003, **85**, 167–173.
51 X. Li and A. B. Kolomeisky, *J. Chem. Phys.*, 2013, **139**, 144106.
52 M. Lange, M. Kochugaeva and A. B. Kolomeisky, *J. Chem. Phys.*, 2015, **143**, 09B605.
53 D. T. Gillespie, *J. Phys. Chem.*, 1977, **81**, 2340–2361.
54 J. I. Friedman, A. Majumdar and J. T. Stivers, *Nucleic Acids Res.*, 2009, **37**, 3493–3500.
55 A. B. Kochaniak, S. Habuchi, J. J. Loparo, D. J. Chang, K. A. Cimprich, J. C. Walter and A. M. van Oijen, *J. Biol. Chem.*, 2009, **284**, 17700–17710.
56 C. L. Vestergaard, P. C. Blainey and H. Flyvbjerg, *Nucleic Acids Res.*, 2018, **46**, 2446–2458.
57 A. Tafvizi, F. Huang, J. S. Leith, A. R. Fersht, L. A. Mirny and A. M. Van Oijen, *Biophys. J.*, 2008, **95**, L01–L03.