Theory of site-specific interactions of the combinatorial transcription factors with DNA

2010 J. Phys. A: Math. Theor. 43 195003

(http://iopscience.iop.org/1751-8121/43/19/195003)

J. Phys. A: Math. Theor. **43** (2010) 195003 (20pp) doi:10.1088/1751-8113/43/19/195003

Theory of site-specific interactions of the combinatorial transcription factors with DNA

R Murugan

Department of Biotechnology, Indian Institute of Technology Madras, Chennai,
Tamil Nadu 600036, India

E-mail: rmurugan@gmail.com

Received 18 December 2009, in final form 24 March 2010
Published 23 April 2010
Online at stacks.iop.org/JPhysA/43/195003

Abstract

We derive a functional relationship between the mean first passage time associated with the concurrent binding of multiple transcription factors (TFs) at their respective combinatorial *cis*-regulatory module sites (CRMs) and the number n of TFs involved in the regulation of the initiation of transcription of a gene of interest. Our results suggest that the overall search time τ_s that is required by the n TFs to locate their CRMs which are all located on the same DNA chain scales with n as $\tau_s \propto n^{\alpha}$ where $\alpha \sim (2/5)$. When the jump size k that is associated with the dynamics of all the n TFs along DNA is higher than that of the critical jump size k_c that scales with the size of DNA N as $k_c \sim N^{2/3}$, we observe a similar power law scaling relationship and also the exponent α. When $k < k_c$, α shows a strong dependence on both n and k. Apparently there is a critical number of combinatorial TFs $n_c \sim 20$ that is required to efficiently regulate the initiation of transcription of a given gene below which $(2/5) < \alpha < 1$ and beyond which $\alpha > 1$. These results seem to be independent of the initial distances between the TFs and their corresponding CRMs and also suggest that the maximum number of TFs involved in a given combinatorial regulation of the initiation of transcription of a gene of interest seems to be restricted by the degree of condensation of the genomic DNA. The optimum number m_{opt} of roadblock protein molecules per genome at which the search time associated with these n TFs to locate their binding sites is a minimum seems to scale as $m_{\mathrm{opt}} \propto Ln^{\alpha/2}$ where L is the sliding length of TFs whose maximum value seems to be such that $L \leqslant 10^4$ bps for the *E. coli* bacterial genome.

PACS numbers: 87.10.−e, 87.15.kj

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

1. Introduction

Site-specific interaction of a protein molecule with the genomic DNA is a fundamental process in biological physics. In prokaryotes, such as bacteria, the initiation of transcription of a given gene of interest occurs upon site-specific interaction of the RNA polymerase (RNAP) enzyme complex with its promoter sequence. Further the initiation of replication of the genomic DNA starts upon the site-specific interaction of the DNA polymerase III enzyme (DNAPIII) complex with the origin of replication sequence element. Eukaryotes such as higher plants and animals differ from prokaryotes in the sense that the genomic DNA that is confined inside the nucleus of the cell is packaged into higher order structures called chromosomes with the aid of non-specifically bound histone particles. This higher order packaging is dynamic since it must be loosened whenever the genes present in a location need to be transcribed into the corresponding mRNA. Then the mRNA that is produced inside the nucleus is transported to the cytoplasm via the nuclear pores present on the nuclear membrane and the translation of mRNA into the corresponding protein polypeptide chain takes place in the cytoplasm of the cell. Hereafter we mainly consider the site-specific interaction of the protein molecules with those loosely packaged regions of the genomic DNA which are actively transcribed and also free from the non-specifically bound histone bodies. Since the length of the genomic DNA is much higher in eukaryotes, the initiation of transcription of a given gene by the RNA polymerase II enzyme (RNAPII) complex at the corresponding promoter sequence additionally requires the interaction of the respective transcription factors (TFs) with the corresponding *cis*-acting binding sites aka *cis*-regulatory modules (CRMs) associated with the gene which are also present on the same DNA chain. These *cis*-acting elements may be present far away from the upstream/downstream of the promoter sequence of the gene of interest. Upon binding with the respective CRMs, these TFs regulate the splicing [1], cell division, development and differentiation of higher eukaryotes. The mechanism of action of these TFs on the initiation of the eukaryotic transcription is not yet understood clearly.

According to the currently accepted picture, these TFs first interact with their respective CRMs to form a complex (we call this complex as ETF complex for convenience). Subsequently this ETF complex stabilizes/destabilizes the interaction between RNAPII and the promoter sequence of the gene of interest via distal action. The mode of this distal action is not clearly understood. There are at least two different school of thoughts [2–4], namely it is mediated either by a one-dimensional (1D) tracking of ETF complex along DNA toward the corresponding promoter sequence of the regulated gene or by looping out of the intervening DNA segment that is present in between the ETF complex and promoter sequence of the gene of interest [2–4]. The efficiency of site-specific binding of TFs with their corresponding CRMs can be characterized by the kinetic affinity (speed) and the specificity (fidelity) of interactions. Here the kinetic affinity indicates how fast the TF molecule locates its specific binding site on the DNA chain which is measured in terms of the site-specific bimolecular association rate (mol^{-1} s^{-1}). The specificity of interactions indicates how best a TF molecule of interest can differentiate its specific binding site from rest of the non-specific binding sites which is measured in terms of differential binding free energy ($\Delta\Delta G_{\mathrm{s-ns}}$ kcal mol^{-1}) which is defined as $\Delta\Delta G_{\mathrm{s-ns}} = |\Delta G_{\mathrm{s}} - \Delta G_{\mathrm{ns}}|$ where ΔG_{ns} is the free energy that is associated with the non-specific binding and ΔG_{s} is the free energy that is associated with the specific binding. When a given set of specific and non-specific binding sites on the same DNA competes for the same pool of TF molecules to bind, one can derive the expression $\Delta\Delta G_{\mathrm{s-ns}} \propto \ln[c_{\mathrm{s}}c_{\mathrm{pns}}/(c_{\mathrm{n}}c_{\mathrm{ps}})]$. Here c_{s} and c_{n} are the concentrations (mol) of the freely available specific and non-specific binding sites, c_{ps} and c_{pns} are the molar concentrations of the specifically and non-specifically bound TFs, $c_{\mathrm{p}} = (p_0 - c_{\mathrm{ps}} - c_{\mathrm{pns}})$ is the concentration of the TFs which are freely available in the bulk solution and p_0 is the total concentration of the TFs in entire system. Since the size of the eukaryotic genomes is larger than that of the prokaryotes, the speed–fidelity negative correlation problem arises [2] in the binding of TFs at their corresponding CRMs in the case of eukaryotes. This means that whenever the CRM–TF system is tuned to achieve a maximum affinity for the binding of TFs with CRMs by decreasing the non-specific-type interactions, the fidelity of binding will be a minimum. On the other hand, whenever the CRM–TF system is manipulated to achieve a maximum fidelity of binding with TFs by enhancing the specific-type interactions, the kinetic affinity will be a minimum. One should note that this affinity–specificity negative correlation arises as a consequence of the concurrent increase (or decrease) in the affinity of the protein molecule of interest toward the non-specific binding sites whenever the affinity for the specific binding site is increased (or decreased) [2]. Since the number of non-specific binding sites in the eukaryotic genomes is much higher than that of the number of specific binding sites, the TF molecule of interest tightly binds with the non-specific DNA sequences under such higher specificity conditions and stays as such for longer periods and it would never find its specific binding site within the physiologically reasonable time scales [2].

It is believed that the problem of speed–fidelity negative correlation can be circumvented by the combinatorial binding and regulation of various TFs at their respective CRMs with cooperative-type interactions among them [2] rather than a one-TF and one-CRM mode in eukaryotes. In the case of combinatorial regulation, instead of a single TF binding site for a given gene there will be a sequence of overlapping and non-overlapping CRMs for many different TFs and there will be cooperative-type interactions between the adjacently binding TFs. Here the cooperative-type interactions indicate the protein–protein interactions between the adjacently binding TFs along the sequentially located CRMs that in turn stabilize the ETF complex. When all these TFs in a given combination assemble at their respective CRMs associated with the gene of interest, the ETF complex is formed and subsequently this ETF complex enhances the initiation of transcription of the associated gene via distal action. Here one should note that the cooperative-type interactions among the adjacently bound TFs occur only upon the arrival of the respective TFs at their CRM binding sites. It has been argued that the combinatorial-type binding of various TFs with cooperative-type interactions between the adjacently bound TFs on the same DNA increases the fidelity of binding without decreasing the kinetic affinity. In other words, different combinatorial subsets of a given set of TFs can efficiently regulate many different genes with the same kinetic affinity as that of a one-TF one-binding-site mode but with higher fidelity than the same. Although these cooperative effects can increase the specificity of binding of TFs at their combinatorial CRMs, depending on the initial positions of these TFs on the DNA chain and the number of such combinatorial TFs involved in the regulatory process, the speed of searching of TFs for their combinatorial binding sites may be significantly retarded.

One can demonstrate this issue with the following example. Consider a combination of four TF-binding sites ‘1234’ which are all sequentially located along the DNA chain that act as CRMs for the promoter of a given gene of interest. Assume that the corresponding TF protein molecules are ‘a’, ‘b’, ‘c’ and ‘d’. Here the TF protein ‘a’ will bind at the target position ‘1’ and ‘b’ will bind at position ‘2’ and so on. Further, the TF protein ‘a’ can interact with ‘b’, whereas ‘b’ can cooperatively interact with the adjacently bound ‘a’ and ‘c’ and so on. When all these four TFs ‘abcd’ assemble at their respective binding positions ‘1234’, the formation of ETF complex takes place that finally interacts with the RNAPII-promoter complex via distal action which results in the initiation of transcription of the corresponding gene. Assume that already all these TFs are non-specifically bound with the DNA chain and they are currently all searching for their CRMs via 1D diffusion dynamics along DNA. If the interactions between these TFs are cooperative type and the binding energies of all the TFs are identical, then the specificity associated with the interaction of these combinatorial TFs 'abcd' with their binding sites '1234' will be cumulative and it will be at least four times higher than that of a single TF-binding interaction. This follows from our definition of specificity and also from the fact that the free energies associated with the combinatorial binding of all these four TFs with their respective CRMs in the presence of non-specific binding sites are additive and the free energies associated with the cooperative-type protein–protein interactions among TFs will also be added up to this overall free energy of stability of the ETF complex. At the same time, the speed of assembling of such combinatorial complex 'abcd' at sequential locations '1234' strongly depends on the initial positions of the TFs on the DNA chain. When the initial positions of these TFs are in the order such as 'abcd' with respect to their sequential binding sites '1234', the speed will be higher than that of a random initial configuration on the DNA chain such as 'badc'. Clearly a crossing dynamics of TFs over other TFs is required for cases such as 'badc' to form the final complex 'abcd' at the sequential binding sites '1234'. In this particular situation, the TF 'b' must cross 'a' and TF 'd' must cross 'c' to form the final 'abcd' complex on the sequential binding sites '1234'. When the TFs slide on the DNA, crossing dynamics is not allowed and the time that is required for presorting a random initial configuration into an ordered configuration is infinite. This means that hopping, jumping and inter-segmental transfer dynamics of TFs are strictly required for the assembly of TFs at their corresponding CRMs. Let us assume that the initial distances of these TFs from their binding sites are η_a, η_b, η_c and η_d. When a sequential and parallel binding of all the TFs is warranted, the total search time τ_s associated with the finding of the combinatorial positions '1234' by the respective TFs 'abcd' will depend on the longest initial distance of TFs from their CRMs as $\tau_s \propto \max(\eta_a, \eta_b, \eta_c, \eta_d)$. In this regard there are many open questions which need to be answered. (a) How does the mean first passage time (MFPT) associated with the binding of multiple TFs at their corresponding CRMs depend on the number of TFs in that combination and their relative initial positions on the DNA chain? (b) Is there any restriction on the maximum possible number of such TFs in a given combination in eukaryotes? (c) To what extent the spatial organization of the DNA chain can enhance the MFPT associated with the combinatorial binding of TFs? In this paper using a combination of theoretical and simulation tools we will answer these questions in detail.

2. Theory

We assume that the TF molecules locate their respective binding sites on the genomic DNA via a combination of one and three-dimensional diffusion dynamics as that of the standard site-specific DNA–protein interactions [5–7]. Consider a linear DNA of N base pairs (bps) in length with helical ends at the positions $\{0, N\}$ containing a gene of interest (figure 1). We assume that the initiation of transcription of this gene is regulated by the binding of a combination of n TFs $tf = \{tf_1, tf_2, \ldots, tf_n\}$ at the corresponding sequentially located CRMs associated with the gene. The initial positions of the corresponding TF molecule on the DNA chain were $\bar{x}_0 = \{x_{01}, x_{02}, \ldots, x_{0n}\}$ and their current positions are $\bar{x} = \{x_1, x_2, \ldots, x_n\}$ and their corresponding CRMs are located at the positions $\bar{x}_a = \{x_{a1}, x_{a2}, \ldots, x_{an}\}$ which are all such that $x_{a1} < x_{a2} < \cdots < x_{an}$. These CRM sites are acting as absorbing boundaries for the dynamics of the respective TF molecules. Here we also assume that $\{\bar{x}_0, \bar{x}, \bar{x}_a\} \in [0, N]$ and when $i \neq j$ we have $x_i \neq x_j$ due to the excluded volume effect. Upon all the n TFs in a given combination finding their respective CRMs, assembly of the ETF complex completes and subsequently the initiation of transcription of the gene of interest occurs.

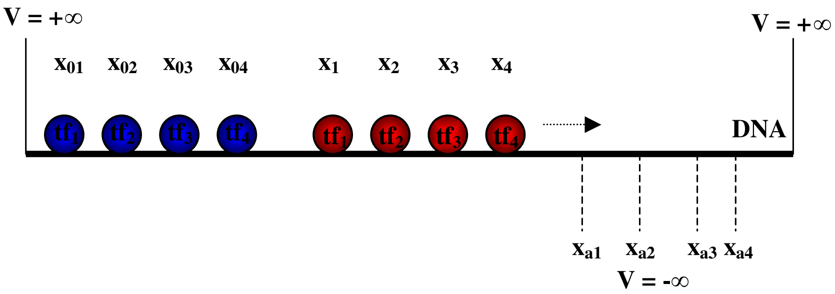

Figure 1. Various initial and boundary conditions used in the text. Here we consider combinatorial binding of four transcription factors $tf = \{tf_1, tf_2, tf_3, tf_4\}$. The initial positions of these TFs were at $\bar{x}_0 = \{x_{01}, x_{02}, x_{03}, x_{04}\}$ and currently they are all undergoing one-dimensional diffusion dynamics along the DNA chain at the positions $\bar{x} = \{x_1, x_2, x_3, x_4\}$ where the positions $\{0, N\}$ are the helical ends of the DNA chain under consideration which are acting as reflecting boundaries for all the TF protein molecules. Whenever all these TFs sequentially assemble at the positions $\bar{x}_a = \{x_{a1}, x_{a2}, x_{a3}, x_{a4}\}$ the formation of the enhancer-TF complex (ETF) completes that subsequently results in the initiation of the transcription of the gene of interest. When the diffusion dynamics of TFs on the DNA chain is characterized by a unit step random walk, the dynamic reflecting boundary condition for tf_1 is $x_l < x_1 < x_2$ where $x_l = 0$ is the left helical end of the DNA under consideration and for tf_2 it is $x_1 < x_2 < x_3$ and so on. As a result, the CRM binding site of tf_1 will be visible to tf_1 only after all the other TFs $\{tf_2, tf_3, tf_4\}$ cross x_{1a}.

The quantity that we want to calculate here is the time τ_s that is required by all the n TFs to find their respective CRMs. When all the TFs search their respective CRMs on DNA via multiple cycles of non-specific association that is followed by a scanning of average L bps and dissociation, the minimum time that is required by these n TFs to assemble at n CRMs will be $\tau_s = NL^{-1}(\tau_{L,n} + \tau_{\mathrm{ns},n})$ where $\tau_{L,n}$ is the average time that is required by all the n TFs to scan L bps of DNA, $\tau_{\mathrm{ns},n}$ is their average re-association time (s) and NL^{-1} is the minimum number of such association–scan–dissociation events that is required by all the n TFs to scan the entire DNA. The non-specific association time τ_{ns} will be such that $(\tau_t/N) \leqslant \tau_{\mathrm{ns},n} \leqslant \{n\tau_t/N\}$ depending on whether all the n TFs bind with DNA at the same time or at different time points where τ_t (bps s) is the 3D diffusion controlled bimolecular association time. When all the n TFs scan the entire DNA, the probability that is associated with the TFs to locate their CRMs on DNA is 1. When $n = 1$, we find that $\tau_{L,1} \approx (6D)^{-1}L^2$ (s) is the mean time [7] that is required by a single TF molecule to scan L bps of DNA where D (bps^2 s^{-1}) is the 1D diffusion coefficient associated with the dynamics of the TF molecule on DNA and $\tau_{\mathrm{ns},n} = (\tau_t/N)$. When all these n TFs independently scan the DNA chain in a synchronized manner, we find that $\tau_{L,n} \geqslant \tau_{L,1}$. This inequality will be true when all the TFs bind with DNA at the same time but at different locations and independently scan an average length of L bps and then they dissociate at the same time. This is an extreme situation where the association–scan–dissociation cycles of all the n TFs are temporally synchronized. Here the increase in $\tau_{L,n}$ is mainly the consequence of retarding effects of adjacently moving TFs on the dynamics of a given TF due to spatial confinement and dynamic reflections. On the other hand when all the n TFs scan the entire DNA in an asynchronous manner we find that $\tau_{L,n} \leqslant \{n\tau_{L,1}\}$ and $\tau_{\mathrm{ns},n} \leqslant (n\tau_t/N)$. This inequality will be true when all the n TFs bind with DNA at different locations at different time points and scan L bps at non-overlapping time intervals and they dissociate at different time points. This is another extreme situation where the association–scan–dissociation cycle of one TF molecule is temporally not overlapping with that of another TF molecule. Upon combining both these inequalities, we find that

$\tau_{L,1} \leqslant \tau_{L,n} \leqslant \{n\tau_{L,1}\}$. In the following sections we will show that this inequality is indeed true and we also derive an expression for $\tau_{L,n}$.

The presorted initial configuration of TFs as $x_{01} < x_{02} < \cdots < x_{0n}$ is necessary when all the TFs scan the DNA chain by sliding dynamics where a given TF is not allowed to jump across other TFs and therefore the time that is required to sort all the n TFs starting from a random initial configuration to the right initial order is infinite. We define this presorted initial configuration as $\bar{x}_0 \to \tilde{x}_0$. The presorted initial condition may not be true in real *in vivo* situations since the initial positions of the TF protein molecules of our interest in a given combination on the genomic DNA will be a random one. One should also note that in real situations the DNA chain will be in a condensed state that allows various facilitated dynamics such as hopping and inter-segmental transfers via ring-closure events which in turn allow the crossing dynamics of TFs over other TFs. Prior to the assembly of these TFs at their respective CRMs, these TF molecules independently undergo many cycles of association–scan–dissociation events. This means that though all the n TFs start their search for their combinatorial CRMs with a presorted initial order, such order will be lost in the subsequent cycles of association–scan–dissociation events and the overall search time will be almost independent of their initial configuration as well as their arrival times on DNA.

The cooperative-type protein–protein interactions between the adjacently binding TFs can occur only after the arrival of the respective TFs at their corresponding CRMs. This means that the search time (inverse of this search time is the ‘on rate’) associated with the binding of TFs at their respective CRMs is independent on the cooperative-type interactions between the adjacently binding TFs. However, the residence time (inverse of this residence time is the ‘off rate’) associated with the site-specifically bound TFs will be strongly influenced by the cooperative-type interactions between the adjacently binding TFs that in turn enhance the specificity of the site-specific interactions. Here we mainly consider the search times associated with the assembly of all the combinatorial regulatory TFs at their respective CRMs and therefore we can ignore the cooperative-type interactions among the adjacently binding TFs. In other words, this corresponds to a maximum specificity condition. One should note that this assumption is not valid when the cooperative-type protein–protein interactions occur among the combinatorial TFs before they arrive at their corresponding CRMs. Apparently the cooperative-type interactions are not favored among the combinatorial TFs before they arrive at their CRMs since such interactions would increase the size of the one-dimensionally as well as three-dimensionally diffusing protein–protein complexes of TFs. When the dynamics of such larger complexes is driven by the thermal energy, the increase in size of the complexes might in turn increase the search times associated with the finding of CRMs mainly by decreasing the overall 1D and 3D diffusion coefficients. When such interactions between these combinatorial TFs prior to their arrival at the respective CRMs are mandatory for the biological functions, such interactions as well as the searching dynamics of these larger TF-complexes for their CRMs will be coupled to an active transport such actin–myosin system which in turn is driven by the external free energy input in the form of ATP hydrolysis rather than the thermal energy.

With this background, the concurrent dynamics of n TFs which are all present on the same DNA chain can be well described by a set of generalized Langevin equations as $\{d_t x_i = \sqrt{D_i}\xi_{i,t}\}$ where x_i is the position of the ith TF molecule on the DNA chain and $i = 1, 2, \ldots, n$. Here $\xi_{i,t}$ are the delta-correlated Gaussian white noises with means as $\langle \xi_{i,t} \rangle = 0$ and variances as $\langle \xi_{i,t}\xi_{j,t'} \rangle = \delta_{ij}\delta(t - t')$ for all $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n$ where δ_{ij} is defined in such a way that $\delta_{ij} = 0$ for $i \neq j$, $\delta_{ij} = 1$ for $i = j$ and D_i is the 1D diffusion coefficient associated with the dynamics of the ith TF molecule along DNA. We define the average 1D diffusion coefficient as D. Here we have assumed that the mean force originating from the resultant non-specific electrostatic interactions [1, 2, 5–7] between the positively charged amino acid side chains of the DNA binding domains (DBDs) of TFs and the negatively charged phosphate backbone of the DNA helix will be comparable [7] with that of the thermal free energy (∼0.591 kcal mol^{-1} at 298 K) in the presence of intervening water molecules at the DNA–protein interface. The dynamics of the protein molecule will be generally confined within the capturing domain that is formed by this electrostatic attractive force field. As a result, one can assume the helical ends of the DNA chain as reflecting boundaries for the 1D diffusion dynamics of the TF molecule. The corresponding Fokker–Planck equation (FPE) associated [7, 8] with the set of such coupled Langevin equations that describe the temporal evolution of the probability density function $P_n(\bar{x}, t|\bar{x}_0, 0)$ which is associated with the simultaneous observation of these n TFs at the DNA positions $\bar{x}$ at time t, which were all started from the DNA positions $\bar{x}_0$ at time $t = 0$, can be written as follows:

$$\partial_t P_n(\bar{x}, t|\bar{x}_0, 0) = \sum_{i=1}^{n} (D_i/2)\partial_{x_i}^2 P_n(\bar{x}, t|\bar{x}_0, 0). \tag{1}$$

Here the initial condition is $P_n(\bar{x}, 0|\bar{x}_0, 0) = \prod_{i=1}^{n} \delta(x_i - x_{0i})$ and the boundary conditions vary depending on the type of TF dynamics. When the TF molecules search for their CRMs via sliding dynamics on the DNA chain, the presorted initial condition $\bar{x}_0 = \tilde{x}_0$ is necessary and the boundary conditions are given as follows where $i = 1, 2, \ldots, n$:

$$[P_n]_{\bar{x}=\bar{x}_a} = \left[\partial_{x_1} P_n\right]_{x_1=0} = \left[\partial_{x_n} P_n\right]_{x_n=N} = \left[\partial_{x_i} P_n\right]_{x_i=x_{i-1}, i>1} = \left[\partial_{x_i} P_n\right]_{x_i=x_{i+1}, i<n} = 0. \tag{2}$$

Here the absorbing boundary condition is defined such that whenever all the n TFs simultaneously find their respective CRMs as $\bar{x} \to \bar{x}_a$ where $x_i = x_{ai}$ for all $i = 1, 2, \ldots, n$, we have $P_n = 0$. The MFPT $T_n(\bar{x}_0)$ associated with the simultaneous finding of all the combinatorial CRMs $\bar{x}_a$ by the respective n TFs which all started from the presorted positions $\bar{x}_0$ on the DNA chain obeys the following backward-type FPE with similar boundary conditions [7, 8] as given by equation (2).

$$\sum_{i=1}^{n} (D_i/2)\partial_{x_i}^2 T_n(\bar{x}) = -1. \tag{3}$$

Here the boundary conditions for equation (3) can be explicitly given from equation (2) as follows:

$$[T_n]_{\bar{x}=\bar{x}_a} = \left[\partial_{x_1} T_n\right]_{x_1=0} = \left[\partial_{x_n} T_n\right]_{x_n=N} = \left[\partial_{x_i} T_n\right]_{x_i=x_{i-1}, i>1} = \left[\partial_{x_i} T_n\right]_{x_i=x_{i+1}, i<n} = 0. \tag{4}$$

The general solution to equation (3) can be written as follows:

$$T_n(\bar{x}) = -\left(x_1^2/(2D_1)\right)\left(2 + \sum_{i=2}^{n} \alpha_i D_i\right) + (1/2)\sum_{i=2}^{n} \alpha_i x_i^2 + (1/2)\sum_{i=1}^{n} (\beta_i x_i + \gamma_i). \tag{5}$$

Here α_i, β_i and γ_i are arbitrary constants which need to be determined from the appropriate boundary conditions given by equation (4). When there is only one TF in the system which searches for its CRM binding site via sliding dynamics along the DNA chain, $n = 1$ and the particular solution to equation (3) can be given as $T_1(x_{10}) = -x_{10}^2 D_1 + \beta_1 x_{10}/2 + \gamma_1$ where $\beta_1 = 0$ and $\gamma_1 = \left(x_{1a}^2/D_1\right)$. When $n > 1$, the general properties of this particular solution to equation (3) can be derived as follows. Since we have $\bar{x}_0 < \bar{x} < \bar{x}_a$ with a presorted $\bar{x}_0 = \tilde{x}_0$, the TF tf_1 can find its CRM binding site that is located at x_{a1} only after all the other TFs have already crossed x_{a1}. As a result, the total time that is required by all the n TFs to cross x_{a1} and subsequently assemble at their corresponding CRMs which are all located at the positions $\bar{x}_a$ starting from $\bar{x}_0$ should be such that $T_n(\bar{x}_0) \leqslant \{nT_1(x_{01})\}$. This follows from the fact that the initial distances $\eta_i = |x_{0i} - x_{ai}|$ of the TFs from their CRM binding sites are all the same as

$\eta_i = \eta_j$ in the current setting. Here one should note that $T_n(\bar{x}_0)$ is the time that is required by the n TFs to scan a DNA segment of length η_i and we have not considered the pure enhancing effects of other $(n-1)$ TFs on tf_n and the pure retardation effects of other $(n-1)$ TFs on tf_1 in this calculation [9]. When $\eta_i = L$, we find $T_n(\bar{x}_0) = \tau_{L,n}$. Consequently $T_n(\bar{x}_0) \leqslant \{nT_1(x_{01})\}$ will be modified to a generalized inequality as $T_n(\bar{x}_0) \leqslant \{n^\alpha T_1(x_{01})\}$ where α is an exponent such that $0 < \alpha \leqslant 1$. This result is in line with our predicted inequality $\tau_{L,1} \leqslant \tau_{L,n} \leqslant \{n\tau_{L,1}\}$ based on scaling arguments. Using stochastic simulation of equation (3) we will show in the following sections that this inequality is indeed true when $n \leqslant n_c$ where $n_c \sim 20$ is some critical number of TFs in a given combination and when $n > n_c$ the exponent becomes as $\alpha > 1$.

3. Results

For the simulation purpose, we consider a linear DNA chain with a length of $N = 150$ bps. To start with, we assume that there is a non-specifically bound TF protein molecule located at the left helical end $x_{01} = 0$ at time $t = 0$ and currently it is searching for its CRM binding site that is located at the position $x_{a1} = 25$ via 1D diffusion dynamics. We set this 1D diffusion coefficient associated with dynamics of the TF molecule as $D_1 = 1$. This is a typical unit-step 1D one-walker problem and the corresponding equation (3) can be solved exactly as follows. When there is only one TF in the system, we have $n = 1$ and equation (3) can be written as $d_{x_1}^2 T_1(x_1) = -2$ with the absorbing boundary condition at the position of CRM as $[T]_{x_1 = x_{1a}} = 0$ and the reflecting boundary conditions [7, 8] at the helical end of the DNA chain as $\left[d_{x_1} T_1\right]_{x_1=0} = \left[d_{x_1} T_1\right]_{x_1=N} = 0$. The MFPT associated with this single TF to locate its CRM which is present at x_{a1} starting from x_{01} by sliding dynamics can be given as $T_1(x_{01}) = \left(x_{a1}^2 - x_{01}^2\right)$. Here the MFPT is measured in terms of the dimensionless number of steps taken by the TF molecule to locate its CRM. In the present case we have $T_1(0) = 625$ since we have set the initial position of TF on DNA as $x_{01} = 0$. One can also derive a similar expression whenever the TF molecule starts the search for its CRM by sliding on DNA anywhere from the interval (x_{a1}, N) as $T_1(x_{01}) = \left(x_{a1}^2 - x_{01}^2\right) + 2N(x_{01} - x_{a1})$. With this background we assume that there are n TFs which are all located at the sequential initial positions on the same DNA as $x_{01}, x_{01} + 1, \ldots, x_{01} + n$ and currently trying to locate their respective CRMs by sliding dynamics which are all located sequentially on the same DNA chain as $x_{a1}, x_{a1} + 1, \ldots, x_{a1} + n$ in such a way that the inequalities $(x_{a1} + n) \leqslant N$ and $(x_{01} + i) < (x_{a1} + i)$ are true for all $i = 1, 2, \ldots, n$. As we have shown in the previous section, the overall MFPT associated with the finding of all the n CRMs by all the n TFs by sliding dynamics should be an increasing function of n that is mainly due to the dynamic reflections and spatial confinement which are imposed on the dynamics of a given TF molecule at the boundaries of other adjacently diffusing TFs along the same DNA chain. To check this proposition, we carried out the stochastic numerical simulations of the system of equations (1)–(3) with the settings as ($x_{01} = 0, x_{a1} = 25, N = 150$ and $D_i = 1$ for all the values of $i = 1, 2, \ldots, n$ and unit base pair step size as $k = 1$ for a sliding dynamics) at various n values.

Results of this simulation study showed the following scaling relationship (figure 2) for the overall MFPT ($T_n(\bar{x}_0)$) that is associated with the binding of all the n TFs with their respective CRMs:

$$T_n(\bar{x}_0) \approx n^\alpha T_1(x_{01}). \tag{6}$$

One should note that equation (6) is in line with the results obtained in the theory section that is based on the scaling arguments. Here the MFPTs at various n values were computed by

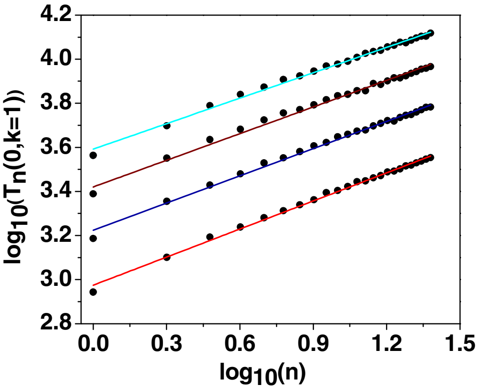

Figure 2. Dependence of the mean first passage time $T_n(0, 1)$ (measured as the dimensionless number of steps) associated with the finding of n binding sites by n TFs on n when the jump size $k = 1$. The simulation settings are as follows. Here the size of DNA template is $N = 150$, initial positions of TF proteins are set as $x_{0i} = i$ and the corresponding CRM binding sites are set at $x_{ai} = \chi + i$ where $i = 1, 2, \ldots, n$ and $\chi = \{30, 40, 50, 60\}$ so that the initial distances of TFs from their CRM binding sites take the values $\mu = \{30(\text{red/bottom}), 40(\text{royal/next to bottom}), 50(\text{wine/below top}), 60(\text{cyan})/\text{top}\}$. The MFPT was computed by averaging over 10^5 trajectories of all TFs. For $n = 1$ and the initial distance $\mu = 50$, we observed the MFPT of $T_1(0, 1) \sim 50^2$. For other values of n we observed the scaling relationship $T_n(0, 1) = T_1(0, 1)n^{\alpha}$. We observed the exponent $\alpha \approx (2/5)$ with respect to n irrespective of the initial distance μ. Solid lines are the linear least-squares fitting with the log–log transformed data ($R^2 = 0.99$) that yielded the exponent $\alpha \approx 0.39 \pm 0.005$ for $\mu = 50$.

averaging over 10^5 trajectories of all the n TFs on the same DNA chain. When $n = 1$ and the initial position of the TF molecule on the DNA chain is set as $x_{01} = 0$ in equation (6), we recover the result for the one-walker problem as $T_1(0) = x_{a1}^2 = 625$. The linear least-squares fitting of the log-transformed MFPT data at various log-transformed n values yielded the parametric estimate for the exponent α as $\alpha \sim (2/5)$. This value of the exponent α seems to be independent (figure 2) of the initial positions $\bar{x}_0$ and the distances $\mu_i = |x_{0i} - x_{ai}|$ of various combinatorial TF molecules from their corresponding CRMs which are all located on the same DNA chain.

To understand the effect of the spatial organization of the template DNA on $T_n(\bar{x}_0)$, we carried out the simulations of equations (1)–(3) at various jump size values ($k > 1$). When $k > 1$, the presorted initial configuration of TFs is not necessary. Here the unbiased random jump size k means that the random walker which started from the lattice position x can be found (jump) anywhere inside the interval $(x - k, x + k)$ in the next step with equal probabilities as $w_i = 1/(2k)$. Under this condition D_1 becomes [7–10] as $D_1 = \sum_{i=-k}^{k} i^2 \big/ (2k)$ in the dimensionless form and for an arbitrary jump size distribution function we find that $D_1 = \sum_{i=-k}^{k} i^2 w_i$. Clearly when $k = 1$, we have $D_1 = 1$ and when $k > 1$, for $w_i = 1/(2k)$ we find that $D_1 = 6^{-1}(k + 1)(2k + 1)$. We will show in the later sections that the equal probability assumption $w_i = (2k)^{-1}$ for the distribution of hopping lengths is indeed valid under *in vivo* conditions when the jump size k is less than that of the critical jump size value as $k \leqslant k_c$ where k_c scales with N as $k_c \propto N^{2/3}$ [9, 10]. This scaling law is true whenever the electrostatic attractive field is strong enough to keep the TF molecule under non-specifically bound conditions until it scans the entire DNA. When the TF molecule scans only L bps and then dissociates, the scaling law becomes as $k_c \propto (NL)^{1/3}$. Since D_1 increases with k as $D_1 \propto k^2$, one can conclude that the scan time for L bps of DNA decreases with k. We also learn from the earlier studies [9, 10] that the search time or the scan time for L bps cannot be enhanced further by increasing the jump size beyond this critical value k_c. Here one should recall the fact that the average jump size k is positively correlated with the spatial condensation and degree of super-coiling of DNA. This means that the search time that is required by a TF molecule to locate its CRM on DNA can be enhanced by increasing the degree of condensation of DNA only within certain limit. The boundary conditions which were used for the numerical simulation of the system of equations (1)–(3) under random jump conditions for all $i = 1, 2, \ldots, n$ can be given as follows:

$$\left.\begin{array}{l}[P_n]_{\bar{x}=\bar{x}_a} = \left[\partial_{x_1} P_n\right]_{x_i=0} = \left[\partial_{x_n} P_n\right]_{x_i=N} = \left[\partial_{x_i} P_n\right]_{x_i=x_j, i\neq j} \\ [T_n]_{\bar{x}=\bar{x}_a} = \left[\partial_{x_1} T_n\right]_{x_i=0} = \left[\partial_{x_n} T_n\right]_{x_i=N} = \left[\partial_{x_i} T_n\right]_{x_i=x_j, i\neq j}\end{array}\right\} = 0. \tag{7}$$

Apart from the reflecting boundary conditions in equation (7), the TF molecules are also allowed to jump across other TF molecules provided that there is no other TF molecule present at the target position of the jump event. Assume that there are only two TF molecules in the systemnamely ‘a’ and ‘b’. Consider that the TF molecule ‘a’ is located at the boundary of other TF molecule ‘b’ where the positions of ‘a’ and ‘b’ on DNA are respectively x and $x + 1$. When the jump size is k, the allowed target positions of the jump events for the TF molecule ‘a’ with respect to the ‘b’ molecule are such that $x - k, x - k + 1, \ldots, x - 1$ and$x + 2, x + 3, \ldots, x + k$. The position of ‘b’ acts as a reflecting boundary for the transition such as $x \to x + 1$ of ‘a’. These conditions further ensure that two TF molecules never occupy the same location on DNA (excluded volume effect). When the random jumps were allowed in the dynamics of all the n TFs with a jump size of k bps, we found from numerical simulations that the exponent α and the pre-exponential term in the expression of MFPT as given by equation (6) were strongly dependent on k as follows:

$$T_n(\bar{x}_0, k) \approx T_1(x_{01}, k) n^{\alpha_k}. \tag{8}$$

Here one should note from equation (6) that $\alpha_1 \sim (2/5)$ for $k = 1$. When $n = 1$, we find the limit for the pre-exponential term from the earlier studies [9] as $\lim_{k\geqslant k_c} T_1(x_{01}, k) \to N$, where $k_c \sim 2N^{2/3}$ is the critical jump size associated with the dynamics of an individual TF molecule. When $n > 1$ and k is such that $k \geqslant k_c$, our simulation results show (figure 3) that the scaling exponent α_k in equation (8) is almost independent of k and we observed a limiting relationship as $T_n(\bar{x}_0, k_c) \approx N n^{\alpha_1}$. The linear least-squares fitting of the log–log transformed MFPT data at various n values again yielded a parametric estimate of this critical exponent as $\alpha_{k_c} = \alpha_1 \sim (2/5)$. When $k < k_c$, the scaling exponent α_k in equation (8) seems to be strongly dependent on k and also the number of TFs n in a complicated manner. When $k < k_c$, we observed a point of inflection in $T_n(\bar{x}_0, k)$ with respect to n such that (figure 4) when $n < n_c$ where n_c is some critical number of TF molecules in the combinatorial regulation, the exponent α was such that $(2/5) < \alpha < 1$. When $n > n_c$, we observed the exponent such that $\alpha > 1$. The derivative plot (figure 5) of $T_n(\bar{x}_0, k)$ with respect to n clearly demonstrated this inflection behavior. It appears that irrespective of the jump size k, the first derivative of $T_n(\bar{x}_0, k)$ showed a monotonic decrease until the critical value of n is reached as $n_c \sim 5$ and the point of inflection occurred in a broad range as $5 < n < 20$. When $n > n_c$, then $d_n T_n(\bar{x}_0, k)$ showed a further increase (when $k < k_c$) or a decrease (when $k \geqslant k_c$) depending on the jump size k. The critical number of TF molecules in the combinatorial regulation n_c seems to be independent of the initial distances μ_i and also the total size of the DNA chain N under consideration (figure 5).

It follows from equations (6) and (8) that $\tau_{L,n} \sim n^{\alpha}\tau_{L,1}$. This means that $\tau_{L,1} \leqslant \tau_{L,n} \leqslant \{n\tau_{L,1}\}$ as we have predicted in the theory section using hand waving arguments. When there is a coherence in the dynamics of all the n TFs, while they are non-specifically interacting with

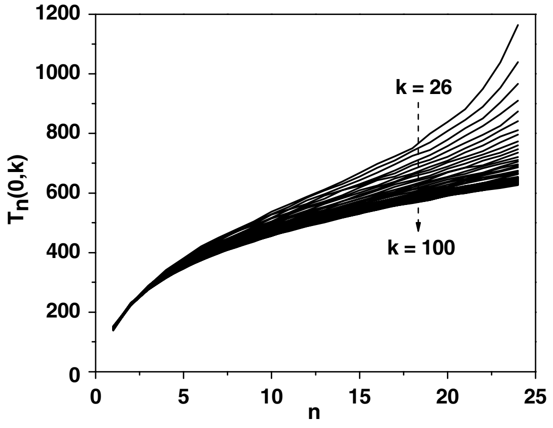

Figure 3. Dependence of the mean first passage time $T_n(0, k)$ (measured as the dimensionless number of steps) that is associated with the finding of n CRM binding sites which are all located sequentially on the same DNA chain by n TFs on the jump size k. Here the size of DNA template is $N = 150$, initial positions of TFs on DNA are set at $x_{0i} = i$ and the corresponding binding sites are located at $x_{ai} = 25 + i$ where $i = 1, 2, \ldots, n$. The critical jump size for the DNA length of $N = 150$ is $k_c = 2 \times 150^{2/3} \approx 57$. When the jump size k was such that $k > k_c$, we observed the scaling relationship $T_n(0, k_c) \approx Nn^{\alpha}$ where the exponent was $\alpha \sim (2/5)$. The MFPT was computed by averaging over 10^5 trajectories of TFs. The linear least-squares fitting ($R^2 = 0.99$) of the log–log transformed MFPT data for jump size $k = 58$ yielded the exponent $\alpha \approx 0.38 \pm 0.005$. For the jump size $k < k_c$ we observed a point of inflection such that when $n < 20$ we observed the exponents in the range of $0.4 < \alpha < 1$ and when $n > 20$ we observed the exponent $\alpha > 1$.

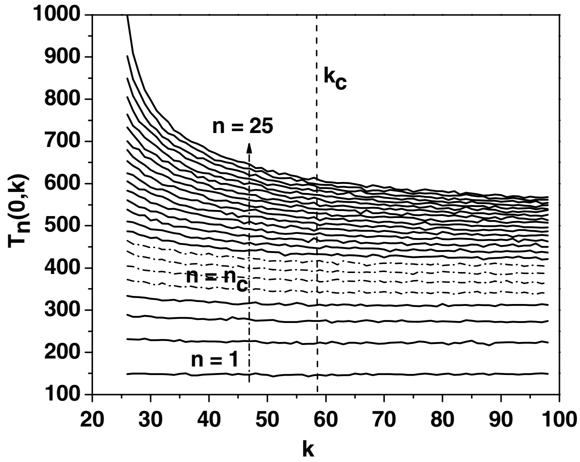

Figure 4. Dependence of the mean first passage time $T_n(0, k)$ (measured as the dimensionless number of steps) that is associated with the finding of n CRM binding sites by n TFs on the jump size k at various n values. Here the size of DNA template is $N = 150$, initial positions of TFs on DNA are at $x_{0i} = i$ and the corresponding binding sites are located at $x_{ai} = 25 + i$ where $i = 1, 2, \ldots, n$. The MFPT was computed by averaging over 10^5 trajectories of TFs. Here the point of inflection occurs in a broad range of n values as $5 < n_c < 20$.

DNA, then we find $\tau_{\mathrm{ns},n} = \tau_{\mathrm{ns},1}$. Upon substituting the expression for $\tau_{L,n}$ in the expression for τ_s we find that $\tau_s \sim NL^{-1}(\tau_{L,1}n^{\alpha} + \tau_{\mathrm{ns},1})$. This result clearly suggests that the presence of many TFs on the same DNA chain ultimately decreases the overall 1D diffusion coefficient

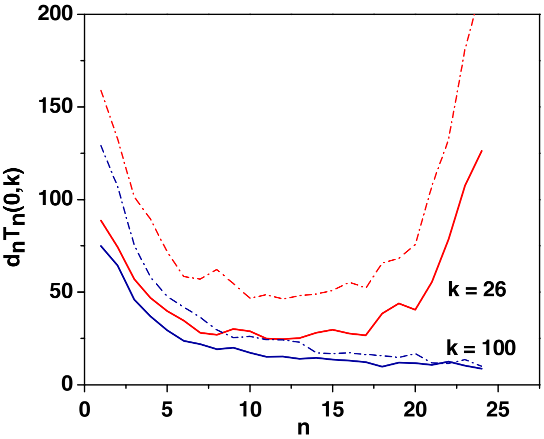

Figure 5. Derivative plot of function $T_n(0, k)$ (measured as the dimensionless number of steps) with respect to n that clearly shows the point of inflection. Simulation settings for the solid lines are as follows. Here the size of the DNA template is $N = 150$, initial positions of TFs are at $x_{0i} = i$ and the corresponding CRM binding sites are located at $x_{ai} = 25+i$ where $i = 1, 2, \ldots, n$ so that the initial distance is $\mu = 25$. Simulation settings for the dotted lines are as follows. Here the size of DNA template is $N = 250$, initial positions of TFs are at $x_{0i} = i$ and the corresponding binding sites are located at $x_{ai} = 50+i$ where $i = 1, 2, \ldots, n$ so that the initial distance is $\mu = 50$. The MFPTs were computed by averaging over 10^5 trajectories of TFs. The critical jump size for $N = \{150, 250\}$ was $k_c = 2N^{2/3} \sim \{57, 81\}$ and therefore two different jump sizes k were tried as $k = \{26 < k_c, 100 > k_c\}$. This result clearly demonstrates that the inflection point is independent of the distance between the initial and CRM binding positions of TFs and also the size of the DNA chain under consideration.

associated with the dynamics of TFs due to confinement of the search space and dynamic reflections at the boundaries of adjacently diffusing other TFs [9] on the same DNA chain that results in an increase in the overall target finding time irrespective of the jump size. Upon solving the equation $\partial_L \tau_s = 0$ for L, we find the optimum sliding length that is required by the TFs to achieve the overall minimum [9, 10] search time as $L_{\text{opt}} = \sqrt{6D\tau_{\text{ns},1}n^{-\alpha}}$. Upon substituting this back into τ_s we find the required overall minimum search time as $\tau_{s,\min} \sim 2N\sqrt{\tau_{\text{ns},1}n^{\alpha}/(6D)}$. When $k < k_c$ and $n < n_c$, the maximum value of α is $\alpha = 1$ and the maximum value of $\tau_{\text{ns},n}$ is $\tau_{\text{ns},n} = n\tau_{\text{ns},1}$. This means that the maximum value of the search time that is optimized for the sliding length scales as $\tau_{s,\max} \propto n$. Using these results one can derive the scaling relationships for the number of TFs in the combinatorial regulation as $\tau_{s,\min} \propto n^{\alpha/2}$ and $L_{\text{opt}} \propto n^{-\alpha/2}$.

To investigate the anomalous behavior of the system of n TFs, we plotted the normalized positional variances σ_x^2 at various values of τ_B where τ_B is the ratio between the number of random walk steps and the number of steps required to attain the 'steady-state' positional variance $\sigma_{x,s}^2$. Here we call it as 'steady state' since the system under consideration is a non-equilibrium one and the positional variance σ_x^2 starts to decline beyond this steady state due to absorption of the TF molecules at their corresponding CRM binding sites. Figure 6 clearly suggests the presence of anomalous-type diffusion when the jump size is $k = 1$ which is somewhat similar to that of the single file diffusion problem addressed in the literature [11]. To simplify the current problem we denote the TF molecule that is closest to the combinatorial binding sites as the 'outer' one and the remaining TFs are the 'inner' ones. Results suggest that except the 'outer' TF molecule all the other 'inner' TF molecules show a sub-diffusion

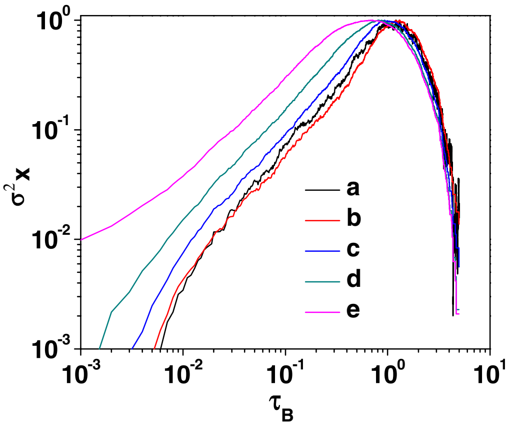

Figure 6. Anomalous-type diffusion observed when the jump size associated with the dynamics of TFs toward their combinatorial binding sites on DNA is $k = 1$. Here the total size of the DNA is 150 bps, σ_x^2 is the normalized positional variance and τ_B is the ratio between the number of random walk steps and the number of steps that is required to attain the steady-state positional variance. There are five TF molecules, namely $tf = \{a, b, c, d, e\}$, whose initial positions were at $x = \{1, 2, 3, 4, 5\}$ and the combinatorial binding sites are at $x = \{25, 26, 27, 28, 29\}$ whereas the helical ends $x = \{0, 150\}$ are reflecting boundaries. Here the TF molecule 'e' is the outer one and all others are inner ones. Clearly the outer one shows a super-diffusion $\sigma_x^2 \propto \tau_B^\chi$ where $\chi > 1$ and all the other inner ones show a sub-diffusion-type dynamics $\sigma_x^2 \propto \tau_B^\chi$ where $\chi < 1$.

which means that $\sigma_x^2 \propto \tau_B^\chi$ where $\chi < 1$ and the outer one shows the super-diffusion which means that $\sigma_x^2 \propto \tau_B^\chi$ where $\chi > 1$. This anomalous-type behavior decreases as the jump size k increases (figure 7). This is reasonable since as k increases the confinement effects of molecular crowding and the dynamic reflections which are the sources of anomalous behavior decrease which in turn results in the normal-type diffusion which means that $\sigma_x^2 \propto \tau_B^\chi$ where $\chi = 1$.

4. Discussion

So far we have assumed that the system contains only the set of combinatorial TFs of our interest which is not true in the real *in vivo* situation. There will be different classes of other protein molecules concurrently undergoing 1D diffusion dynamics along with the TFs of our interest on the same DNA chain. As a result, the dynamical trajectories of the set of TFs of our interest will be always interfered by other classes of protein molecules (roadblocks). Recently effects of such roadblock protein molecules on the dynamics of a given TF protein molecule have been studied in detail [12]. The main conclusion of this theoretical study is that the presence of such roadblock protein molecules induces more association–scan–dissociation events in the dynamics of the TFs, that in turn results in the existence of an optimum number of such roadblock protein molecules m_{opt} per genomic DNA, at which the overall minimum search time associated with the site-specific binding of these TF molecules with the CRMs is attained. Further calculations showed that [12] m_{opt} should be in the order of $m_{\text{opt}} \leqslant 10^4$ for the *E. coli* genome of size $N \sim 4.6 \times 10^6$ bps. This result agrees well [13] with the total number of protein molecules $\sim 3 \times 10^4$ that is found on the genomic DNA of *E. coli* in the log-phase of the growth kinetics.

One can also derive this result in a way different from [12] as follows. Consider the 1D diffusion-mediated search dynamics of a single TF protein molecule for its CRM in

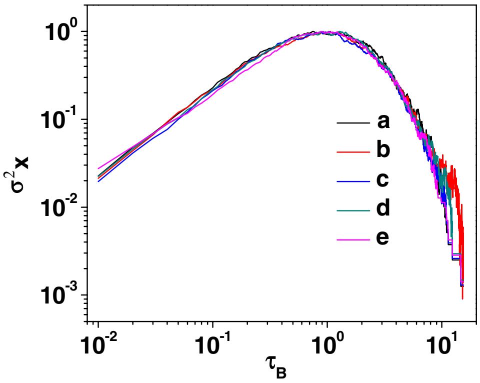

Figure 7. Normal-type diffusion observed when the jump size associated with the dynamics of TFs toward their combinatorial binding sites on DNA is $k \gg 1$. Here the total size of the DNA is 150 bps, σ_x^2 is the normalized positional variance and τ_B is the ratio between the number of random walk steps and the number of steps that is required to attain the steady-state positional variance. There are five TFs, namely $tf = \{a, b, c, d, e\}$, whose initial positions were at $x = \{1, 2, 3, 4, 5\}$ and the combinatorial binding sites are at $x = \{25, 26, 27, 28, 29\}$ whereas the helical ends $x = \{0, 150\}$ are reflecting boundaries. Here the TF molecule 'e' is the outer one and others are inner ones. Clearly both the outer and inner ones show a normal diffusion $\sigma_x^2 \propto \tau_B^\chi$ where $\chi = 1$.

the presence of m one-dimensionally diffusing other classes of roadblock protein molecules on the same DNA chain. Under this condition the 1D sliding lengths L of the TFs of our interest rescales [14] as $L \to Lm^{-1}$. As a consequence, the expression for the overall search time τ_s that is associated with the binding of all the TFs of interest with the CRMs that we have derived in the previous sections becomes $\tau_s = NmL^{-1}(\tau_{L,1}m^{-2} + \tau_{\mathrm{ns},1})$. Figure 8 shows the plot of this overall search time $\tau_s \to \tau_s(m)$ as a function of m at various values of the sliding lengths L for $n = 1$ and figure 9 shows the 3D surface plot of the search time $\tau_s \to \tau_s(m, L)$ as a function of both the variables m and L. When $n = 1$, upon solving $\partial_m \tau_s = 0$ for the variable m, we find the optimum number of other classes of roadblock protein molecules that is required to minimize the search time associated with the binding of the TF of our interest with its corresponding CRM as $m_{\mathrm{opt}} = \sqrt{\tau_{L,1}/\tau_{\mathrm{ns},1}}$. Upon substituting the expressions $\tau_{L,1} = (6D)^{-1}L^2$ and $\tau_{\mathrm{ns},1} = (\tau_t/N)$ in m_{opt} we find that $m_{\mathrm{opt}} = L\sqrt{N(6D\tau_t)^{-1}}$. When $n > 1$ we find that $m_{\mathrm{opt}} = \sqrt{\tau_{L,n}/\tau_{\mathrm{ns},n}}$. When all these n TF molecules non-specifically bind with DNA at different time points, we find $\tau_{\mathrm{ns},n} \approx n\tau_{\mathrm{ns},1}$ and $\tau_{L,n} \approx n^\alpha \tau_{L,1}$. This means that $m_{\mathrm{opt}} \propto n^{(\alpha-1)/2}$. On the other hand, when all these n TF molecules non-specifically bind with DNA at the same time points, we find $\tau_{\mathrm{ns},n} \approx \tau_{\mathrm{ns},1}$ and this means that $m_{\mathrm{opt}} \propto n^{\alpha/2}$ under such conditions.

From the literature we find that the volume of an *E. coli* bacterial cell is $V_e \sim 10^{-18}$ m^3 [15], the size of the *E. coli* genome is $N \approx 4.6 \times 10^6$ bps [16] and there are at least N non-specific binding sites for the TF molecule of our interest whose copy number tf_c inside the *E. coli* cell is in the order of $tf_c \sim 10^2$. Using these values, the three-dimensional diffusion controlled bimolecular collision time τ_t (mol s) can be calculated as follows. The concentration of a single non-specific binding site on the DNA chain as well as a single TF molecule inside the *E. coli* cell volume is $\sim 2 \times 10^{-9}$ M. When there is only one binding site on the DNA chain and only one TF protein molecule inside a cellular volume of $V_e \sim 10^{-18}$ m^3, the maximum achievable 3D diffusion controlled collision rate between a single non-specific binding site (bps) of the DNA chain and a single TF protein molecule of our interest can

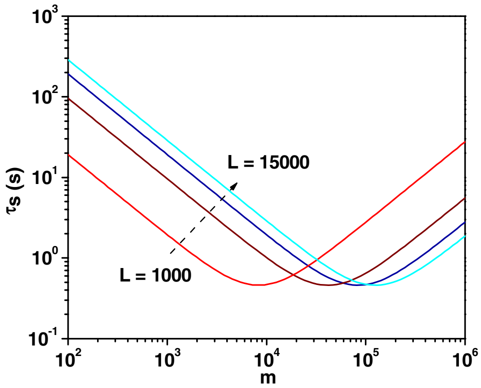

Figure 8. Variation of the search time τ_s (measured in seconds) associated with the finding of the CRM binding sites by transcription factor (TF) of our interest as a function of the m (dimensionless numbers) roadblock protein molecules present on the same DNA chain at various one-dimensional sliding lengths $L = \{1, 5, 10, 15\} \times 10^3$ bps. In the case of the bacterium *E. coli*, there are at least $N \sim 4.6 \times 10^6$ non-specific binding sites for a given TF molecule whose copy numbers inside the *E. coli* cell volume of $V_e \sim 10^{-18}$ m^3 will be in the order of $tf_c \sim 10^2$. From our theory we find $\tau_s = NmL^{-1}(\tau_{L,1}m^{-2} + \tau_{\text{ns},1})$ where $\tau_{\text{ns},1} \sim 6 \times 10^{-9}$ (s) is the time required by the TF molecule of interest to make a non-specific contact with the DNA chain via three-dimensional diffusion controlled routes and $\tau_{L,1} = L^2/(6D)$ is the time required by the TF molecule to scan L bps of the DNA chain. Here $D \sim 4 \times 10^5$ bps^2 s^{-1} [17, 18] is the one-dimensional phenomenological diffusion coefficient associated with the dynamics of the TF protein molecule on the DNA chain. Using these numerical values in the expression for the target finding search time τ_s we find that $\tau_s \approx \text{mL}^{-1}(0.21\ \text{L}^2\text{m}^{-2} + 0.028)$ (seconds).

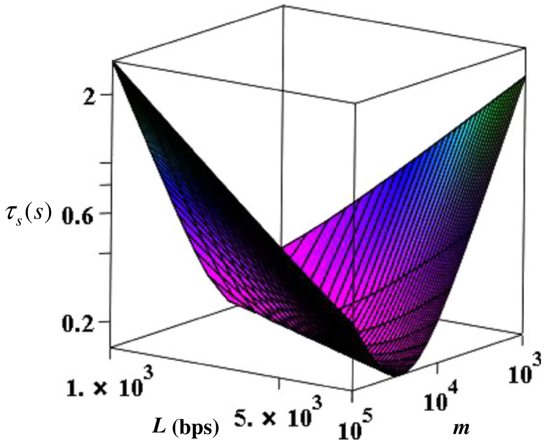

Figure 9. Thee-dimensional surface plot of the overall search time τ_s (seconds) associated with the finding of the CRM binding site by the TF protein molecule of interest on the DNA chain as a function of the numbers m (dimensionless numbers) of a roadblock protein molecule present on the same DNA as well as the sliding length L (bps). For an *E. coli* cell volume of $V_e \sim 10^{-18}$ m^3 the explicit expression for the search time becomes as $\tau_s \approx \text{mL}^{-1}(0.21\ \text{L}^2\text{m}^{-2} + 0.028)$ (seconds). In this calculation we have assumed that there are at least $N \sim 4.6 \times 10^6$ non-specific binding sites for a given TF molecule whose copy numbers inside the *E. coli* cell volume will be in the order of $tf_c \sim 10^2$.

be given as $k_t \sim 0.4$ (bps^{-1} s^{-1}), where we have used the value of 3D diffusion controlled collision rate limit when there are molar concentrations of both the protein molecule and its respective binding site on the DNA chain as $\sim 10^8$ (mol^{-1} s^{-1}). This limiting value transforms under the nano-molar (nM) concentrations of the reactants as $\sim 10^{-1}$ (nM^{-1} s^{-1}). Since the intracellular concentrations of a single DNA binding site and a single TF protein molecule are in the range of $\sim$2 nM, without loss of generality one can write nM $\rightarrow$ bps and we finally arrive at the result $k_t \sim 0.4$ (bps^{-1} s^{-1}). Upon rescaling this three-dimensional diffusion controlled collision rate as $k_{to} \rightarrow (k_t \times tf_c \times N)$ (s^{-1}) for $tf_c \sim 10^2$ copies of TF molecules and $N \sim 4.6 \times 10^6$ non-specific binding sites which are present on the genomic DNA inside a cellular volume of V_e, one finds that $k_{\text{to}} \sim 1.8 \times 10^8$ (s^{-1}). Using this value, one can compute the time $\tau_{\text{ns},1}$ that is required for the non-specific binding of a given TF molecule with the genomic DNA inside the *E. coli* cell volume via 3D routes as $\tau_{\text{ns},1} = (k_{\text{to}})^{-1} \sim 6 \times 10^{-9}$(s). From the earlier studies [17], we find the 1D diffusion coefficient that is associated with the dynamics of TF on the genomic DNA as $D \sim 0.046\ \mu\text{m}^2\ \text{s}^{-1} \sim 4 \times 10^5\ \text{bps}^2\ \text{s}^{-1}$ where we used the transformation rule 1 bps $\approx 3.4 \times 10^{-10}$ m and 1 μm $\approx$ 2941 bps. Upon substituting these numerical values in the expression for m_{opt}, we finally find that $m_{\text{opt}} \sim 8L$.

One should note that the experimentally observed *in vitro* sliding length L that is associated with the diffusion of *lac* repressor protein on a stretched DNA chain [18] ranges from $\sim$120 nm to $\sim$2920 nm. This corresponds to a range of DNA length of $L \sim$ (35–8588) bps. Using these results one finds the optimum number of roadblock protein molecules which is required to attain the minimum search time associated with the binding of a given TF of our interest with its corresponding CRM that is also present on the genomic DNA of *E. coli* as $m_{\text{opt}} \sim 7 \times 10^4$ bps. This result is in line with the recent theoretical studies which state that the optimum number of roadblock protein molecules per genomic DNA of *E. coli* should be in the order of $\sim 10^4$ [12]. The observed sliding length mentioned so far is from the *in vitro* studies and one also should note that the sliding length is strongly dependent on the ionic strength of the medium. Higher ionic strengths weaken the non-specific electrostatic attractive force that is present at the DNA–protein interface leading to lower sliding lengths. The ionic strength under *in vitro* conditions will be much lower than that of the *in vivo* conditions. As a consequence, the approximate sliding length under *in vivo* conditions seems to be in the order as $L \sim 10^2$ bps and corresponding to this sliding length we find $m_{\text{opt}} \sim 10^3$. When the existing number of roadblocks that is present on the genomic DNA of *E. coli* during the log-phase of the growth kinetics is $\sim 3 \times 10^4$, the maximum achievable 1D sliding length $L_{\max}$ of the TF molecule of our interest on the genomic DNA can be computed from the inequality $(3 \times 10^4) \geqslant \{8L_{\max}\}$ as $L_{\max} \leqslant 10^4$ (bps). This is in line with the maximum value of L that is obtained for a stretched DNA by single molecule *in vitro* experiments [18]. When we consider the assembly of all the n combinatorial TFs at their corresponding sequentially located CRMs on the same DNA chain in the presence of roadblocks, it follows from equation (6) that the time that is required for the 1D scanning of L bps of DNA by all these n TFs rescales with n in that combination as $\tau_{L,n} \rightarrow (\tau_{L,1} n^{\alpha})$. Under this condition we find the expression for optimum m as $m_{\text{opt}} = n^{(\alpha-1)/2}\sqrt{\tau_{L,1}/\tau_{\text{ns},1}}$. This also means that $m_{\text{opt}} \approx 8Ln^{(\alpha-1)/2}$ and $L_{\max} \leqslant \{10^4 n^{-(\alpha-1)/2}\}$. These results also put a limit on the maximum possible number of combinatorial TF molecules that can be associated with the regulation of the initiation of transcription of a given gene of interest which is located on the genomic DNA of *E. coli* as $n \sim 1$. This result in turn directly follows from the inequality condition $L_{\max} \leqslant 10^4$.

What are all the consequences of the results given by equation (8)? Here one should recall the fact that the jump size k is directly proportional to the degree of condensation of the DNA lattice. Whenever the degree of condensation of DNA is not enough to achieve the critical jump size k_c, our theory suggests that the number of TFs in a combination should be limited in the range of $1 < n < 20$ to avoid the exponentially growing MFPT with the increasing value of n. It is also interesting [1, 2] to note that the number of TFs involved in the combinatorial regulation of the initiation of transcription of most of the eukaryotic genes is in the range of $1 < n < 20$. This observation is in line with our theoretical results. Our theoretical and simulation results suggest that the number of TFs in a combination should fall in the range of ∼(5–20) for an efficient TF-mediated combinatorial regulation of the initiation of transcription of various genes under all situations. In this range of values of n, our results suggested that $d_nT_n(\bar{x}_0, k)$ will be a minimum (figure 5) especially when the jump size k is such that $k < k_c$. When $k > k_c$, our results show that $d_nT_n(\bar{x}_0, k) \to 0$. Here one should note that when all the TFs are similar and also bind with the same CRM, the resultant MFPT deceases [14] as n increases with the scaling relationship $T_n(\bar{x}_0) \propto n^{-2}$. The origin of this scaling relationship is as follows. The time that is required by the ith TF to scan L bps of DNA by sliding dynamics without collisions with other $(n-1)$ such similar TFs those are present on the same DNA chain is given as $\tau_{L,1} \sim L^2/(6D)$. When there are n such TFs simultaneously performing one-dimensional scanning for the same CRM site on DNA, we arrive at the scaling law $\tau_{L,1} \propto L^2n^{-2}$ since the sliding length L rescales as $L \to (L/n)$. One can generalize these scaling arguments to the situation where there are multiple copies of each of n TF molecules searching for sequentially located n different CRMs binding sites as follows. When there are θ_i copies of TF molecule tf_i, totally there are various $\bar{m} = \sum_{i=1}^{n}\theta_i$ TFs. Since only one molecule of tf_i binds with its CRM among θ_i such tf_i molecules at any time, the resultant scaling for $\tau_{L,n}$ will become as $\tau_{L,n} \propto \tau_{L,1}n^{\alpha}\bar{m}^{-2}$. Here one should note that the rescaling $L \to (L/\bar{m}^2)$ or $L \to (L/n)$ is valid only when the dynamics of TFs along the DNA lattice is mainly via the 3D routes since it does not consider the retarding effects due to the dynamic reflections [9] at the boundaries of other adjacently diffusing TFs on the TF of interest when all these TFs are concurrently searching for the same binding site on DNA via one-dimensional routes.

Recently many groups tried to compute the distribution of jump lengths associated with the dynamics of protein molecules on DNA [19, 20]. Using detailed experimental studies Broek *et al* [first one in 20] has shown that the 3D excursions of the protein molecules could give to the distribution of effective jump lengths. One should note that the effective jump length is not only influenced by the 3D excursions but also by the inter-segmental transfer dynamics via ring closure events which is strongly influenced by the dynamics of the DNA chain. Loverdo *et al* [19] has derived an expression for the probability distribution function w_k of the hopping distances k (bps) associated with the dynamics of the protein molecule on a stretched DNA chain. Especially when k is large, their expression can be written as $w_k \propto (k\ln^2(k))^{-1}$. This wide tail jump size distribution seems to fit well with the experimental observations on the 1D diffusion dynamics of the EcoRV molecule on the elongated DNA under lower salt concentrations [21]. Lomholt *et al* (third one in [20]) has shown that the distribution of effective jump lengths may even be of power-law form on long flexible DNA. One should note that these distribution functions do not consider the various facilitating processes such as inter-segmental transfers due to ring closure events and also the concurrent dynamics of the DNA chain. The ring closure events are in turn driven by the condensation and spatio-temporal dynamics of the DNA chain which in turn bring two different distal segments of the same DNA chain closer together so that the TF molecule of our interest can be transferred from one segment to another without getting released in the bulk solution. Under such conditions, although the expression for the probability distribution function associated with the hopping distances is not known, one can derive it by numerically simulating such systems. For example one can use [22] the analogy between the self-intersection loop lengths in the theory of random walks and the ring closure events in the theory of site-specific DNA–protein interactions. Stochastic numerical simulations (figure 10), in line with [22], suggest that whenever the jump size k

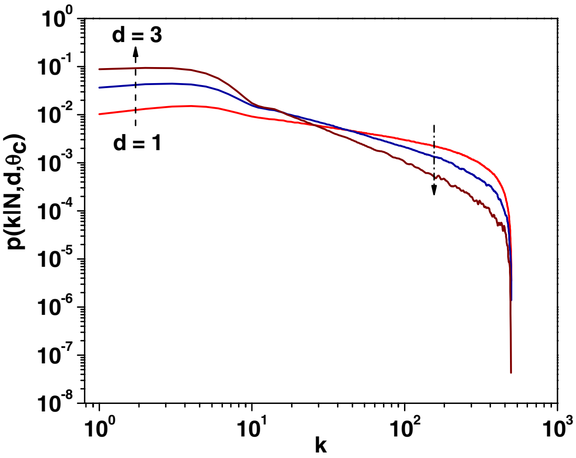

Figure 10. Distribution of random jump sizes k (measured in bps in the context of site specific DNA–protein interactions) in various dimensions d. Here a polymer chain of size $N = 500$ units in length is embedded in one-, two- and three-dimensional lattice boxes such that the volume compression ratio $\theta = (V_B/V_N)$ is beyond the critical compression ratio $\theta \geqslant \theta_c$ where V_B is the volume of the lattice box and V_N is the volume of the polymer lattice. To compute the probability distribution function, self-intersection loop lengths of the embedded polymer were sampled from 10^5 polymeric trajectories which were all starting from the origin. From the earlier studies [22] we learn that when $\theta \geqslant \theta_c$ the average jump size k approaches the critical jump size limit as $k \to k_c(N, d) \sim 2^{2-d}N^{2/3}$ where d is the dimensionality of the lattice box under consideration. For example $\theta_c \sim 0.01$ for $d = 1$ and $\theta_c \sim 100$ for $d = 3$. Here the jump size k is the average of self-intersection loop lengths in various dimensions d. This is analogous to the average loop lengths associated with the ring closure events in site specific DNA–protein interactions. One should note that at lower values of k, the probability distribution is almost flat one. Particularly for $d = 3$, we have the critical jump size limit as k_c (500, 3) $\sim$ 32 bps and in this range the probability distribution function is almost a flat one. This means that one can assume an unbiased random jump condition with equal probabilities whenever $k \ll k_c$ $(N, 3)$.

that is associated with the dynamics of the TF molecule on DNA is such that $k \ll k_c(N, d)$, where $k_c(N, d) \sim 2^{2-d}N^{2/3}$ is the critical jump size associated with the dynamics of TF that is embedded in a d-dimensional lattice box, the probability distribution function associated with the jump size k is almost flat in shape as $w_k \propto (2k)^{-1}$ (figure 10). One should note that the critical jump size that is associated with the dynamics of the TF molecule on the genomic DNA is automatically attained [22] when the volume compression ratio θ of the DNA chain inside the 3D cellular lattice box is beyond certain critical values θ_c as $\theta \geqslant \theta_c$ (figure 8). The volume compression ratio of the genomic DNA inside the cellular lattice box is defined as $\theta = (V_B/V_N)$ where V_B is the volume of the cellular lattice box and V_N is the volume of the embedded DNA chain. It also seems that θ of the *E. coli* genomic DNA is already beyond [22] such critical limit. This means that the spatial organization of the genomic DNA of *E. coli* is designed such that the jump size k that is associated with the dynamics of the non-specifically bound TF molecule on the genomic DNA is equal to the critical jump size limit. From numerical simulations we find that [22] the critical volume compression ratios θ_c in various dimensions are $\theta_c \sim 0.01$ for $d = 1$ and $\theta_c \sim 100$ for $d = 3$. Here the jump size k can be thought as the average of the self-intersection loop lengths of DNA which is embedded in various d-dimensional lattice boxes. This is analogous to the average loop lengths associated with the ring-closure events on the DNA chain which lead to the inter-segmental transfers in the site-specific DNA–protein interactions. Particularly when $d = 3$, we have $k_c(500, 3)\sim$ 32 bps and whenever k is such that $k \ll k_c(500, 3)$ we find that the shape of the probability distribution function associated with the hopping distances is almost a flat one (figure 10). This means that the flat distribution assumption $w_k \propto (2k)^{-1}$ for the hopping distances is indeed valid under *in vivo* conditions.

5. Conclusions

In this paper, we have derived a functional relationship between the mean first passage time (MFPT) associated with the binding of multiple transcription factors (TFs) at their combinatorial CRM binding sites which are all located on the same DNA chain and the number of TFs n involved in the combinatorial regulation of the initiation of transcription of the gene of our interest. Our results suggested that the overall search time τ_s that is required by n such combinatorial TFs to simultaneously assemble at their sequentially located binding sites via 1D diffusion dynamics along the DNA chain scales with n as $\tau_s \propto n^\alpha$ where the value of the exponent is $\alpha \sim (2/5)$. When the jump size k that was associated with the dynamics of TFs along the DNA chain was higher than that of the critical jump size k_c that scales with the size of DNA as $k_c \sim N^{2/3}$, we observed similar power law scaling relationship and the exponent α. When the jump size k was less than that of the critical jump size k_c, the exponent α showed a strong dependence on both k and n. Apparently there was a critical number of combinatorial TFs $n_c \sim 20$ that is required to efficiently regulate the transcription of a single gene of interest below which the exponent α was such that $(2/5) < \alpha < 1$ and beyond which the exponent α was such that $\alpha > 1$. These results seem to be independent of the initial distance between the TFs and their *cis*-acting binding sites which are present on the same DNA chain and also suggest that the maximum number of the TF protein molecule involved in a given combinatorial regulation of the initiation of transcription of a gene of interest seems to be strongly restricted by the degree of condensation of the genomic DNA. Our further results suggest that the optimum number of roadblock protein molecules per genome at which the search time associated with these n TFs to locate their binding sites is minimum seems to scales as $m_{\text{opt}} \propto Ln^{\alpha/2}$ where L is the sliding length of TFs on DNA whose maximum seems to be such that $L \leqslant 10^4$ bps for the *E. coli* genome. Since the number of roadblock protein molecules during the log-phase of the growth kinetics of *E. coli* is in the order of $m \sim 10^4$, our results suggest that the number of TFs in the combinatorial regulation of transcription of a gene of interest inside the *E. coli* cell volume should be restricted to $n \sim 1$.

References

[1] Levin B 2003 *Genes VIII* (Englewood Cliffs, NJ: Prentice Hall)
Alberts B, Bray D, Lewis J, Roberts K and Watson J D 1994 *Molecular Biology of the Cell* (New York: Garland)
Ptashne M 2004 *A Genetic Switch* (New York: Cold Spring Harbor Laboratory)

[2] Ptashne M and Gann A 2001 *Genes and Signals* (New York: Cold Spring Harbor Laboratory)
Murugan R 2005 *Biophys. Chem.* **116** 105

[3] Murugan R 2007 *J. Theor. Biol.* **248** 696

[4] Blackwood E M and Kadonaga J T 1998 *Science* **281** 60

[5] Adam G and Delbruck M 1968 *Structural Chemistry in Molecular Biology* (San Francisco: Freeman)
Riggs A D, Bourgeois S and Cohn M 1970 *J. Mol. Biol.* **53** 401

[6] Berg O G, Winter R B and von Hippel P H 1981 *Biochemistry* **20** 6929
Winter R B, Berg O G and von Hippel P H 1981 *Biochemistry* **20** 6961

[7] Murugan R 2007 *Phys. Rev.* E **76** 011901
Murugan R 2004 *Phys. Rev.* E **69** 011911

[8] Gardiner C W 2004 *Handbook of Stochastic Methods* (Berlin: Springer)

Risken H 1996 *Fokker Plank Equation* (Berlin: Springer)
van Kampen N G 2004 *Stochastic Processes in Physics and Chemistry* (Amsterdam: North-Holland)
Redner S 2001 *A Guide to First Passage Processes* (London: Cambridge University Press)
[9] Murugan R 2009 *Phys. Rev.* E **79** 041913
[10] Murugan R 2006 *J. Phys. A: Math. Gen.* **39** 1575
Murugan R 2006 *J. Phys. A: Math. Gen.* **39** L199
[11] Lizana L and Ambjornsson T 2008 *Phys. Rev. Lett.* **100** 200601
Lizana L and Ambjornsson T 2009 *Phys. Rev.* E **80** 051103
[12] Li G W, Berg O G and Elf J 2009 *Nature Phys.* **5** 294
[13] Ali Azam T, Iwata A, Nishimura A, Ueda S and Ishihama A 1999 *J. Bacteriol.* **181** 6361
Johnson R C, Johnson L M, Schmidt J W and Gardner J F 2005 *The Bacterial Chromosome* ed N P Higgins (Washington, DC: ASM)
[14] Sokolov I M, Metzler R, Pant K and Williams M C 2005 *Biophys. J.* **89** 895
[15] Goodsell D S 1991 *Trends Biochem. Sci.* **16** 203
Albe K R, Butler M H and Wright B E 1990 *J. Theor. Biol.* **143** 163
Elowitz M B *et al* 1999 *J. Bacteriol.* **181** 197
[16] Sutcliffe J G 1979 *Cold Spring Harb. Symp. Quant. Biol.* **43** 77
[17] Elf J, Li G W and Xie X S 2006 *Science* **316** 119
[18] Wang Y M, Austin R H and Cox E C 2006 *Phys. Rev. Lett.* **97** 048302
[19] Loverdo C *et al* 2009 *Phys. Rev. Lett.* **102** 188101
[20] van den Broek B *et al* 2008 *Proc. Natl Acad. Sci. USA* **105** 15738
Lomholt M A *et al* 2009 *Proc. Natl Acad. Sci. USA* **106** 8204
Lomholt M A *et al* 2005 *Phys. Rev. Lett.* **95** 260603
Oshanin G *et al* 2007 *J. Phys.: Condens. Matter.* **19** 065142
[21] Bonnet I *et al* 2008 *Nucleic Acids Res.* **36** 4118
[22] Murugan R 2009 *Phys. Rev.* E **79** 061920