Physical Biology

PAPER

Theory on the mechanism of site-specific DNA–protein interactions in the presence of traps

G Niranjani and R Murugan
Department of Biotechnology, Indian Institute of Technology Madras, Chennai, India
E-mail: rmurugan@gmail.com

Keywords: sequence traps, site-specific binding, DNA–protein interactions

Abstract

The speed of site-specific binding of transcription factor (TFs) proteins with genomic DNA seems to be strongly retarded by the randomly occurring sequence traps. Traps are those DNA sequences sharing significant similarity with the original specific binding sites (SBSs). It is an intriguing question how the naturally occurring TFs and their SBSs are designed to manage the retarding effects of such randomly occurring traps. We develop a simple random walk model on the site-specific binding of TFs with genomic DNA in the presence of sequence traps. Our dynamical model predicts that (a) the retarding effects of traps will be minimum when the traps are arranged around the SBS such that there is a negative correlation between the binding strength of TFs with traps and the distance of traps from the SBS and (b) the retarding effects of sequence traps can be appeased by the condensed conformational state of DNA. Our computational analysis results on the distribution of sequence traps around the putative binding sites of various TFs in mouse and human genome clearly agree well with the theoretical predictions. We propose that the distribution of traps can be used as an additional metric to efficiently identify the SBSs of TFs on genomic DNA.

1. Introduction

Binding of transcription factor (TFs) proteins at specific cis-regulatory modules (CRMs) on the genomic DNA in the presence of enormous amount of non-specific sites is critical for the expression and regulation of various genes inside the living cell [1–5]. In earlier studies, the site-specific binding of TFs was modeled as one-step Smolochowski type three-dimensional diffusion (3Dd) controlled collision process [6]. Later in vitro experiments on the site-specific binding of lac repressor protein with its operator site showed a bimolecular collision rate in the order of $10^{10}$–$10^{11}$ M$^{-1}$ s$^{-1}$ which was $10$–$10^2$ times higher than the Smolochowski type 3Dd controlled rate limit [6–8]. These experimental observations clearly ruled out the possibility of 3Dd only model and suggested a two-step 3D1D diffusion model. According to this, TFs first non-specifically bind with DNA via 3Dd and then search for their specific sites via one-dimensional diffusion (1Dd) [7, 8].

The nonspecific binding of TFs is mainly driven by the electrostatic attractive forces present in between the positively charged DNA binding domains (DBDs) of DNA binding proteins (DBPs) and negatively charged phosphate backbone of DNA. The site-specific binding of TFs via a combination of 3Dd and 1Dd seems to be facilitated by sliding, hopping and inter-segmental transfer type dynamics [7–12]. The non-specifically bound TFs diffuse along DNA with a step size of unit base-pair (bps) in sliding, few bps in hopping and few hundred to thousand bps in intersegmental transfers. Intersegmental transfers occur when two distal segments of the same DNA polymer come into contact through ring-closure events over 3D space (figure 1(A)). The conformational state of DNA seems to play critical role in modulating various 1D facilitating processes. Sliding and hopping will be the dominating modes of dynamics on the relaxed conformational state of DNA. Intersegmental transfer dynamics predominantly occurs on the condensed conformational state of DNA. Various symbols and abbreviations used in this paper are summarized in table 1.

In general, both the nonspecific and site-specific binding of TFs are influenced by several factors viz. conformational state of DNA [13–16], electrostatic...
attractive forces acting at the DNA–protein interface [17–20] and the counteracting shielding effects of solvent ions [10], presence of semi-stationary roadblocks [21] such as nucleosomes [22–25] and dynamic roadblocks on DNA [26–28], conformational fluctuations in the DBDs of TFs [29, 30] (figure 1(B)), and randomly occurring kinetic traps along the DNA sequence [31–33]. Apart from these factors, the spatial organization of the genome structure also play important roles in accelerating the search process of TFs for their cognate sites on DNA [34, 35]. The electrostatic interactions along with the counteracting shielding effects of solvent ions creates a fluidic type environment for the 1D of DBDs of TFs at the DNA–protein interface [16].

Presence of roadblocks increases the dissociation of TFs and subsequently drive the mode of dynamics towards 3D mediated excitations [32, 33]. Switching between 1D and 3D will be enhanced by the conformational fluctuations at the DBDs of TFs across stationary and mobile states [29]. Here the stationary state is more sensitive to the sequence information than mobile one but moves slowly along DNA. Whereas mobile state is less sensitive to the sequence information but moves quickly along DNA. The conformational fluctuations in the DBDs of DBPs seem to be purely thermal driven. The free energy barrier that separates the stationary and mobile states seems to be close to the thermal energy which resembles the dynamics of downhill folding proteins at their mid-point denaturation temperatures [36]. Detailed calculations showed that the extent of thermodynamic coupling between the search dynamics and the rate of conformational fluctuations in the DBDs of TFs will be an optimum at the barrier of \( \sim k_B T \ln 2 \) [36].

Presence of sequence mediated kinetic traps retards the rate of site-specific binding of TFs in several ways. First of all there is a nonzero probability of occurrence of trap sites similar to the specific binding site (SBS) of TFs for sufficiently large genomes. Sequence traps slow down the 1D dynamics of TFs and also increase the overall dissociation time compared to that of other nonspecific binding sites (nSBS). In other words, kinetic traps increase the overall ruggedness or frustration of the binding energy landscape of the DNA sequence. It is still not clear how exactly the effects of kinetics traps are handled by various TFs under in vivo conditions.

Recent experimental studies suggested that the presence of similar binding site adjacent to the SBS significantly influences the site-specific association rate [14, 31, 37]. Especially the extent of such influences seems to be directly proportional to the distance between the SBSs and the kinetic traps. In this context it is still not clear about (a) how the relative position of traps with respect to the position of initial nonspecific contact of TFs with DNA influences the site-specific association rate and (b) how the distribution of the distances between traps and SBSs in the real genomes (whether it is randomly distributed or correlated)
affects the overall site-specific binding of TFs. Using a combination of theoretical and computational tools we will address these issues in detail.

2. Theory

Presence of sequence traps influences the binding of TFs with DNA in two different ways viz. (a) they compete with other nSBSs for the available pool of TFs and TFs bound at traps will be temporarily immobilized and (b) they retard the 1Dd dynamics of TFs on DNA. One can approximately ignore the competing effects of traps since the number of traps in a genome will be much lesser than the number of nSBSs. The overall search time or mean first passage time (MFPT) $\tau_{S,U}$ associated with TFs to find their SBSs on DNA via two-step 3D1D mechanism can be written as follows [8, 16]

$$\tau_{S,U} \simeq \left( \frac{P_{BTF}}{U} \right) \left( k_{fa} + k_{fl}(1 + k_{d} \eta_{U}) \right)^{-1};$$

(1)

In this equation $P_{BTF}$ (M, mol $^{-1}$ s $^{-1}$) is the concentration of TFs of interest in cytoplasm, $k_{fa}$ (M $^{-1}$ s $^{-1}$) is the bimolecular rate constant associated with the

Table 1. Various symbols and parameters used in the main text.

| Symbol | Definition | Remarks |
|--------|------------|---------|
| SBS    | Specific binding site | |
| nSBS   | Nonspecific binding site | |
| DBD    | DNA binding domain | |
| DBP    | DNA binding protein | |
| TBS    | Transcription factor binding site | |
| TSS    | Transcription start site | |
| TF     | Transcription factor | |
| nDd    | n-dimensional diffusion | |
| CRM    | cis-regulatory module | |
| $x_l$  | Left side boundary of linear lattice for the 1D random walker | bps |
| $x_R$  | Right side boundary of linear lattice for the 1D random walker | bps |
| $x_A$  | Absorbing boundary for the 1D random walker | bps |
| $x_Z$  | Initial position of the 1D random walker on linear lattice | bps |
| $x_R$  | Position of $r$th trap on linear lattice | bps |
| $l_d$  | base-pair, bps (1 bps = $3.4 \times 10^{-6}$ m) | Dimensionless |
| $X_L$  | $x_l/l_d$ | Dimensionless |
| $X_R$  | $x_R/l_d$ | Dimensionless |
| $X_A$  | $x_A/l_d$ | Dimensionless |
| $X_Z$  | $x_Z/l_d$ | Dimensionless |
| $X_r$  | $x_r/l_d$ | Dimensionless |
| $\phi$ | Average value of microscopic transition rates (w) over various hop sizes. | s $^{-1}$ |
| $k_d$  | Dissociation rate constant associated with the 1D random walker that got stuck at $r$th trap. | s $^{-1}$ |
| $p_u$  | Rate constant associated with nonspecific contact formation between DNA and TFs | M $^{-1}$ s $^{-1}$ |
| $k_d$  | Dissociation rate constant associated with the nonspecific DNA–protein complex. | s $^{-1}$ |
| $\kappa$ | Onsager radius which is defined as the distance between the DBDs of TFs and the phosphate backbone of DNA at which the overall electrostatic attractive forces will be comparable with that of the background thermal energy ($\sim k_B T$). | bps |
| $k$ | Hop size associated with the 1D random walker or the DBPs which are performing 1Dd along DNA lattice. | Dimensionless |
| $d_d$ | 1Dd coefficient associated with the DBPs. In case of the 1D random walker we can define it as | bps $^2$ s $^{-1}$ |
| $D_o$ | Average 1Dd length associated with the dynamics of TFs on DNA. | bps |
| $p_c$ | Probability associated with the entry of TFs from left side of the SBS. | |
| $p_r$ | Probability associated with the entry of TFs from right side of the SBS. | |
| $u$ | Distance between the adjacent located CRMs. | bps |
| $p(u)$ | Probability associated with the TFs which landed at the interconnecting region between two adjacentaly located CRMs to reach any of them without dissociation. | |
| $q$ | Length of genomic DNA spanned upon binding of TFs of interest. | bps |
| $q_c$ | (Solution of $(N-q)(1/4)^N = 1$ for $q$) critical length of binding stretch of TFs at which only one of the SBS can be found on the genome of size $N$ bps. | bps |
| $p_{LZ}$ | Probability associated with the 1D random walker to reach $X_L = 0$ starting from $X_Z$ when $X_A = X_R$. | |
| $p_{LR}$ | Probability associated with the 1D random walker to reach $X_R = 0$ starting from $X_L$ when $X_A = X_R$. | |
direct site-specific binding of TFs via 3Dd only mode, \( k_{\text{on}} \) (M\(^{-1}\) s\(^{-1}\)) is the overall bimolecular non-specific binding rate that depends on the conformational state of DNA and \( k_{\text{d}} \) (s\(^{-1}\)) is the dissociation rate of nonspecifically bound TFs. Further \( \lambda \) is the number of association-scan-dissociation cycles required by TFs to scan the entire DNA and \( \eta_{\text{on}} \) is the overall average time that is required by TFs to scan \( U \) bps of DNA before dissociation. Here \( U \) is a random variable that will take different values in each cycle of association-scan-dissociation. The probability density function associated with the 1Dd lengths \( U \) can be written as follows [16]

\[
p_{U}(U) \approx 2Ue^{-U/\Pi d_{U}}/\Pi \Lambda_{d} \; ; \; \Pi \Lambda_{d} = \sqrt{12d_{U}/k_{d}} . \tag{2}
\]

Here \( \Pi \Lambda_{d} \) is the maximum possible 1Dd length of the nonspecifically bound TFs on DNA that is measured in bps where 1 bps \( \sim 3.4 \times 10^{-10} \) m and \( d_{U} \) (bps\(^{-1}\)) is the diffusion coefficient associated with the 1Dd of TFs on DNA. Clearly presence of sequence traps increases the 1Dd time \( \eta_{\text{on}} \) which in turn increases the overall search time. When TFs of interest move with unit bps step size then the 1Dd coefficient can be defined as \( d_{U} = \sum_{i=1}^{\infty} \lambda_{\text{on}} \). Here \( \lambda_{\text{on}} \) are the microscopic transition rates associated with the forward and reverse movements of TF on DNA and \( \lambda_{\text{on}} \) are the microscopic transition probabilities [10, 38]. Here the step length \( d_{U} \) is measured in bps. We have defined \( \Pi_{d} = 1 \) bps. Since the dynamics at the DNA–protein interface involves segmental motion of DBDs of TFs one can assume protein folding rate limit [39] for the transition rates as \( w_{\text{on}} \sim 10^{6} \) s\(^{-1}\). Noting that \( p_{\text{on}} \sim 1/2 \) for an unbiased 1Dd random walker one finds that \( d_{U} \sim 10^{5} \) bps\(^{-1}\). Approximately this is the value of 1Dd coefficient associated with the sliding dynamics of TFs on DNA [16, 40]. In the following section we will try to understand the effects of sequence mediated kinetic traps on 1Dd dynamics of TFs within the framework of random walks with random hop size.

2.1. Random walks with random hop size and traps

Let us consider the DNA chain as a linear lattice confined within \( (x_{L}, x_{R}) \). Inside this lattice we consider DBP as an unbiased 1Dd random walker that is searching for the absorbing point at \( x = x_{A} \) and \( (x_{L}, x_{R}) \) are the reflecting boundaries (figure 2(A)). Here the absorbing point is the CRMs (SBSs) associated with the TFs which means that whenever TFs hit these points then transcription initiation starts approximately with a probability of one. The Langevin type stochastic differential equation that describes the dynamics of such 1Dd random walker can be written as follows [38, 41–43]

\[
dx/dt = \sqrt{\Lambda_{d}}\, \Gamma_{t}; \; x = x_{A} ; \; \langle \Gamma_{t} \rangle = 0 ; \; \langle \Gamma_{t} \rangle_{\pm} = \delta(t - t') . \tag{3}
\]

Here \( x \) is the position of 1Dd random walker at time \( t \) with the condition that it was at \( x = x_{A} \) at \( t = t_{0} \) and \( \Gamma_{t} \) is the Gaussian white noise whose mean and covariance properties are defined as in equation (3). From equation (3) one can conclude about the mean and variance of the position of an unbiased 1Dd random walker performing normal diffusion as follows

\[
x = x_{A} + \sqrt{\Lambda_{d}} \int_{t_{0}}^{t} \Gamma_{s} ds; \quad \langle x \rangle = x_{A}; \quad \langle x^{2} \rangle = \langle \Gamma^{2} \rangle = d_{U}t . \tag{4}
\]

Here \( d_{U} \) is the phenomenological 1Dd coefficient. The probability density function associated with the dynamics of the 1Dd random walker on a linear lattice obeys the Fokker–Planck equation (FPE) which can be written along with the boundary conditions as follows

\[
\partial_{x}P(x, t|x_{2}, t_{0}) = (d_{U}/2)\partial_{x}^{2}P(x, t|x_{2}, t_{0}); \quad \partial_{x}P(x, t|x_{2}, t_{0}) \big|_{x = x_{L}} = 0 ; \quad P(x_{A}, t|x_{2}, t_{0}) = 0 . \tag{5}
\]

Here \( P(x, t|x_{2}, t_{0}) \) is the probability of observing the 1Dd random walker at position \( x \) at time \( t \) with the condition that it was at \( x = x_{A} \) at \( t = t_{0} \). Apart from the boundary conditions given in equation (5) we also set the initial condition as \( P(x, t_{0}|x_{2}, t_{0}) = \delta(x - x_{A}) \). To simplify our analysis and other computations we use the following scaling transformations so that the dynamical variables in equation (5) become dimensionless

\[
\left\langle w_{\text{on}} \right\rangle = \phi ; \quad \tau = \phi t ; \quad x = x/L ; \quad d_{U} = d_{U}/\phi L^{2} . \tag{6}
\]

When \( w_{\text{on}} = \phi \) and \( p_{\text{on}} \sim 1/2 \) then \( D_{d} = 1 \). Upon rescaling the variables in equation (5) as in (6) we obtain the following FPE in dimensionless form

\[
\partial_{x}P(x, \tau|x_{2}, \tau_{0}) = (D_{d}/2)\partial_{x}^{2}P(x, \tau|x_{2}, \tau_{0}); \quad \partial_{x}P(x, \tau|x_{2}, \tau_{0}) \big|_{x = x_{L}} = 0 ; \quad P(x_{A}, \tau|x_{2}, \tau_{0}) = 0 . \tag{7}
\]

The corresponding initial and boundary conditions are as follows

\[
P(x, \tau_{0}|x_{2}, \tau_{0}) = \delta(x - x_{A}) ; \quad \partial_{x}P(x, \tau|x_{2}, \tau_{0}) \big|_{x = x_{L}} = 0 ; \quad P(x_{A}, \tau|x_{2}, \tau_{0}) = 0 . \tag{8}
\]

Here \( x_{Q} = x_{A}/L \) where the subscript can be \( Q = L, R, Z, A \). The MFPT that is measured in terms of dimensionless number of steps associated with the escape of the 1Dd random walker from the interval \( (x_{L}, x_{R}) \) through the absorbing point \( x_{A} \) starting from \( x_{0} \) will obey the following backward type FPE with appropriate boundary conditions

\[
D_{d}d_{U}^{2}\Pi_{S} = \left\{ \Phi_{x} - 2 \right\}d_{U}; \quad \Pi_{S}d_{U}x_{L} = \left\{ \Phi_{x} - 2 \right\}x_{L} = 0 ; \quad \Pi_{S} = (x_{Q}^{2} - x_{Q}^{2} - 2x_{L}(x_{R} - x_{Q}))/D_{d} . \tag{9}
\]

The results presented in equations (3)–(9) are standard and well known [9, 38, 44, 45]. Now we introduce traps at random locations inside the interval \( (x_{L}, x_{R}) \). Traps are different from the absorbing boundaries in the sense that the free energy associated with the binding of TFs at traps will be much lower than SBSs but much higher than nSBSs. When traps act as transient sinks (absorbing boundaries) for the inflowing TFs then the overall probability \( (p_{\text{sp}}) \) associated with the
Figure 2. (A) Random walk model with random hop size and traps. We consider a linear lattice confined in \((X_L, X_R)\) which are reflecting boundaries i.e. the random walker cannot escape over these boundaries. Inside this interval there is a 1D random walker (TFs) searching for the absorbing point \(X_A\) starting from \(X_Z\). When the 1D random walker hits \(X_A\) (SBS) then it will be removed from the system. With this setting, we introduce a trap at \(X_r\). When the 1D random walker hits this trap then it will get stuck there for an average dwell time of \(r\) and then escapes back into the original lattice. In the context of site-specific DNA–protein interactions the 1D random walker will resemble TFs and absorbing point is the corresponding TF binding site (TBS). Here \(\kappa\) is the Onsager radius associated with the DNA–protein interface [16]. (B) Here the settings are \(X_L = 0, X_R = X_A = 25, \rho_r = 15, X_Z = 13\) and \(X_r\) was iterated from 1 to 24. MFPT was computed over \(10^5\) trajectories (red filled circles) and blue solid line is the prediction by equations (10) and (11) in all the simulations. (C) Settings are \(X_L = 0, X_R = X_A = 25\) and traps were arbitrarily kept at \(X_r = [3, 5, 8, 11, 15, 18, 21, 24]\) and the corresponding \(\rho_r = [5, 15, 25, 35, 45, 55, 65, 75]\) and \(X_Z\) was iterated from 1 to 24. (D) Settings are \(X_L = 0, X_R = X_A = 25\) and traps were arbitrarily kept at \(X_r = [3, 5, 8, 11, 15, 18, 21, 24]\) and the corresponding \(\rho_r = [5, 15, 25, 35, 45, 55, 65, 75]\) and \(X_Z\) was iterated from 1 to 24. (E) Settings are \(X_L = 0, X_R = X_A = 25\) and traps were arbitrarily kept at \(X_r = [3, 5, 8, 11, 15, 18, 21, 24]\) and the corresponding \(\rho_r = [5, 15, 25, 35, 45, 55, 65, 75]\) and \(X_Z\) was iterated from 1 to 24. (F) Settings are \(X_L = 0, X_R = X_A = 25\) and traps were arbitrarily kept at \(X_r = [3, 5, 8, 11, 15, 18, 21, 24]\) and the corresponding \(\rho_r = [5, 15, 25, 35, 45, 55, 65, 75]\) and the hop size \(k\) was iterated from 1 to 15 and \(X_Z\) was iterated in \([25, 35, 45, 55]\) along the dotted arrow. (G) Settings are \(X_L = 0, X_R = X_A = 25\) and \(X_Z\) was iterated in \([14, 19, 23]\) along the dotted arrow and traps were at \(X_r = [3, 5, 8, 11, 15, 18, 21, 24]\) with corresponding \(\rho_r = [5, 15, 25, 35, 45, 55, 65, 75]\) and the hop size \(k\) was iterated from 1 to 15.
nonspecifically bound TFs to specifically bind with their targets will be \( p_{fi} < 1 \) [31, 37, 40] which is mainly due to the partitioning of trajectories of inflowing TFs between SBSs and traps. In our realistic model, binding of TFs at traps cannot initiate transcription i.e. traps are neither sources nor sinks for the probability influx associated with TFs towards their SBSs.

However traps can significantly slow down the site specific binding rate and hence the overall transcription rate will be reduced. In this background we assume that when the 1D random walker visits a trap, it will get stuck there for a fixed average amount of time (dwell time) and then escapes back to the original lattice interval. We denote the position of \( r \)th trap on linear lattice as \( X_r \), where the subscript \( r \) ranges from 1 to \( m \) and the corresponding average dwell times of 1D random walkers at these traps are denoted as \( \rho_r \). We denote the microscopic rate constant associated with the dissociation of TFs from a trap at position \( X_r \) in a dimensionless form as \( k_r/\phi \). The probability density function associated with the distribution of dwell times of the 1D random walker at the particular trap will be \( \rho_r(\rho) \approx (k_r/\phi) \exp(-k_r/\phi) \) [46, 47]. So that the mean dwell time associated with the random walker that was stuck at the respective trap that is present at the lattice position \( X_r \) can be defined as \( \rho_r = \int_0^{\infty} \rho(\rho) \mathrm{d}\rho \). This can be explicitly written as \( \rho_r = \phi/k_r \). Here \( k_r \) and \( \phi \) are measured in s\(^{-1} \) so that \( \rho_r \) will be measured in dimensionless number of simulation steps since the average time required for each forward or reverse movement of the 1D random walker will be \( 1/\phi \).

There are \( n \) number of traps in \((X_L, X_Z)\) and \( m-n \) number of traps in \((X_Z, X_R)\) and there are totally \( m \) number of traps inside the entire interval \((X_L, X_R)\). Here trap positions act a partial absorbing boundaries for the backward FPE defined in equation (9). The differential equation given in equation (9) cannot be solved analytically with so many number of boundary conditions. However one can derive an approximate formula for the overall MFPT in the presence of traps using the following arguments. First of all one should note that the MFPT in the presence of traps will be always higher than the MFPT in the absence of traps i.e. traps always retard the 1D diffusion dynamics of the random walker. When \( X_L = 0 \) and \( X_Z = 0 \), then all the traps will be located inside \((X_Z, X_R)\) and from equation (9) one finds that the random walker visits each and every site of the interval on an average \( X_R \) number of times under such conditions. This means that the total dwell time of the 1D random walker at those \( m \) number of traps will be added up to the original MFPT associated with the case of no traps as in equation (9) and it will be directly proportional to \( \rho_r \approx \sum_{r=1}^{m} X_r \rho_r \). This result follows from the fact that when \( X_Z = 0 \), \( X_L = 0 \) and \( k = 1 \), then \( D_o = 1 \) and one obtains \( \Pi_\infty = X_R^2 \) from which one can conclude that the 1D random walker visits each site of the interval \((X_L = 0, X_R)\) on an average \( X_R \) number of times before reaching the absorbing point \( X_R \).

Now let us assume that \( X_Z \) is located well inside \((X_L, X_R)\) such that \( X_L < X_Z < X_R \) and \( X_R = 0 \). Depending on the relative position of the trap, the corresponding dwell time will vary. When there are \( n \) number of traps with positions \( X_r \), where \( r = 1 \) to \( n \) located inside \((X_L, X_Z)\) such that \( X_L < X_s < X_Z < X_R \) then the total dwell time \( \rho_T \) will be directly proportional to the splitting probability associated with the 1D random walker to reach \( X_s \) starting from \( X_s \). Here \( \rho_T \) is defined as \( \rho_T \approx \sum_{r=1}^{n} X_r \rho_r \). This is because there is a nonzero probability for the 1D random walker to reach the absorbing point \( X_s = X_R \) that is located right side of \( X_Z \) without visiting the traps located in the left side of \( X_Z \). On the other hand when there are \( m-n \) number of traps located inside \((X_Z, X_R)\) such that \( X_L < X_s < X_s < X_R \), then the total dwell time will be directly proportional to the splitting probabilities associated with the 1D random walker to reach the respective trap positions \( X_r \) as \( \rho_T \approx \sum_{r=m-n+1}^{m} X_r \rho_r \) which decides the number of times the random walker revisits the corresponding traps located in \((X_Z, X_R)\). Here the splitting probability associated with the random walker to reach the respective trap \( X_r \) starting from \( X_r \) (assuming that \( X_r \) is a probability sink) can be defined as \( p_r = (1 - X/Z) \). When \( X_r = 0 \), \( X_s = X_R \) and noting the fact that \( D_o = 1 \) for the hop size of \( k = 1 \) then one can summarize these results using the following approximate formula

\[
\Pi_\infty \approx \Pi_\infty + 2X_R \times \left\{ \sum_{r=1}^{n} \rho_r \frac{1}{X_R} |X_r < X_s \leq X_Z| + \sum_{r=m-n+1}^{m} \rho_r \frac{1}{X_R} |X_Z < X_s < X_R| \right\}.
\]

Here the factor 2 appears in the expression for dwell time since the flux of those trajectories flowing towards \( X_s \) will be again reflected back towards the absorbing boundary \( X_R \). Now we drop the idea of unit step size movement of the 1D random walker and assume that the random walker can hop for \( k \) steps at a time. Suppose the current position of the 1D random walker is \( X \). Then in the next step it can move anywhere in \([X-k, X+k]\) with equal probabilities i.e. \( 1/2k \) and one finds that \( d_o = k \sum_{i=-k}^{k} \rho_r \). Though we have \( d_o \propto k \), 1D coefficient cannot be higher than 3D coefficient \( d_i \) since there is a strong negative correlation between \( p_{li} \) and \( i \). In general \( d_o \leq d_i \) [16]. In the context of site-specific DNA–protein interactions, the hop size is directly proportional to the degree of condensation or supercoiling of the DNA polymer.
This means that condensed conformational state of DNA favors higher hop sizes than relaxed conformational state of DNA. Following the detailed theory of random walks with random step size \([44, 45]\) one can derive the following approximate expression for the MFPT for the 1D random walker with random hop size of \(k\)

\[
\hat{\Pi}_S \approx \Pi_s + \frac{2X_R}{D_0} \left\{ \sum_{r=1}^{n} \rho_r p_{LZ} |X_L < X_r \leq X_Z \right\} + 2 \left( X_R - X_Z + \sum_{r=1}^{m} \rho_r \right) \left( 1 - 1/k \right).
\]

(11)

For \(X_L < X_r < X_A = X_R\), when transition rates are \(w_{ij} = 0\) and the transition probabilities are \(p_{LZ} \simeq 1/2k\) then one finds that \(D_0 \simeq (k + 1)(2k + 1)/6\). Clearly for the hop size of \(k = 1\), we find that \(D_0 = 1\) and equation (11) reduces to equation (10). When \(k > 1\); \(X_0 = 0\; \text{and} \; X_A = X_R\) and \(X_A < X_r \leq X_Z\) then one can derive the following explicit expression for the overall MFPT associated with the escape of the 1D random walker through \(X_R\) starting from \(X_Z\) in the presence of \(m\) traps

\[
\hat{\Pi}_S \approx \left( X_R^2 - X_Z^2 \right)/D_0 + \frac{2X_R}{D_0} \left\{ \sum_{r=1}^{n} \rho_r p_{LZ} |X_L < X_r \leq X_Z \right\} \times \left\{ \sum_{r=n+1}^{m} \rho_r p_{LZ} |X_L < X_r < X_R \right\} \left( 1 - 1/k \right).
\]

(12)

Here one can define the initial position as well as trap position averaged MFPT associated with the 1D random walker to reach the absorbing point \(X_R\) starting from \(X_Z\) as follows i.e.

\[
\hat{\Pi}_S = \left( \int_{X_0}^{X_R} \int_{X_0}^{X_R} \hat{\Pi}_dX_dX_r + \int_{X_0}^{X_R} \int_{X_0}^{X_R} \hat{\Pi}_dX_dX_r \right)/X_R,
\]

\[
\hat{\Pi}_S = 6\left( X_R^3/3 + X_R \sum_{r=1}^{m} \rho_r \right)/(k + 1)(2k + 1) + 2 \left( X_R + \sum_{r=1}^{m} \rho_r \right) \left( 1 - 1/k \right).
\]

(13)

For \(k = 1\) and sufficiently large values of hop size \(k\), equation (13) suggests the following limiting conditions

\[
\lim_{k \rightarrow 1} \hat{\Pi}_S = 2X_R^3/3 + X_R \sum_{r=1}^{m} \rho_r
\]

\[
\lim_{k \rightarrow \infty} \hat{\Pi}_S = 2 \left( X_R + \sum_{r=1}^{m} \rho_r \right).
\]

(14)

Equation (14) suggests that at sufficiently large values of hop sizes the 1D random walker visits each sites of the linear lattice for only once. One needs to multiply \(\hat{\Pi}_S\) by \(\phi\) to transform it back into time units in seconds where \(X_R\) will be still a dimensionless count as in equation (6). Equations (10)–(14) are the central results of this paper which clearly suggest that in the presence of sequence traps the 1Dd time \(\eta_{U}\) will transform as \(\eta_{U} \rightarrow \eta_{U} + \eta_{F}\) where \(\eta_{F}\) is the sum of contributions from the dwell times of traps.

3. Results

From equation (10) one can conclude that the overall delay in the search time \(\eta_{U}\) due to the presence of traps will be (a) directly proportional to the distance of the initial position of TFs from the absorbing boundary located at SBSs \((X_R - X_Z)\) when the traps are not located in between \((X_Z, X_R)\), (b) directly proportional to the distances of the traps from the SBSs \((X_R - X_Z)\) when the traps are located in between \(X_Z\) and \(X_R\). Further (c) relatively less delay in the search time due to the traps will be possible only when there is a negative correlation between the dwell times \(\rho_i\) and the corresponding distances of traps \((X_R - X_i)\) from the SBSs which ensure that the second sum in equation (10) will be at minimum. Here observation (c) is an important one which tell us about the principles associated with the design of CRMs. That is to say the efficiency of CRMs in the gene regulation can be fine-tuned by suitably arranging a set of traps around them. In other words a set of identical CRMs can be well differentiated by appropriately arranging the kinetic traps around them. In the following sections we will first check the validity of the approximate expressions given by equations (10) and (12) by detailed random walk simulations. Then we will analyze the distribution of naturally occurring SBSs of various TFs and the associated sequence traps to check the validity of our propositions (a)–(c).

3.1. Stochastic simulations on random walks with random hop size and traps

To check the validity of equations (11) and (12) we performed stochastic random walk simulations. We considered an unbiased the 1D random walker on a linear lattice of length with \(X_R = 25\) and \(X_Z = 0\). We first simulated with hop size of \(k = 1\). Here \(X_Z\) is a reflecting boundary and \(X_A = X_R\) is the absorbing boundary. This means that when the 1D random walker whose current position is \(X = 0\) tries to visit \(X = -1\) then it will be put back to \(X = 1\). With these settings when \(X_Z = 0\) then from equation (10) one obtains \(\Pi_Z = 625\) since \(D_0 = 1\). When \(X_Z > 0\), then the MFPT associated with the escape of the 1D random walker into \(X_R\) starting from \(X_Z\) will be given by \(\Pi_\Sigma = 625 - X_Z^2\) which is a well-known result. We measure the MFPT in terms of dimensionless number of simulation steps taken by the 1D random walker to reach \(X_A\) starting from \(X_Z\).
Now we introduce a single trap at \( X_c \) and fix the initial position arbitrarily at \( X_2 = 13 \). The trap at \( X_c \) is characterized by a dwell time of \( \rho_c = 25 \) simulation steps. This means that whenever the 1D random walker visits \( X = X_c \), then it will get stuck there for 25 simulation steps. With this settings we iterated \( X_2 \) from 1 to 24 and the results are shown in figure 2(B). Clearly when \( X_c < X_2 \) = 13 then using equation (11) one can compute the MFPT as \( \Pi_2 = 456 + 2 \rho_c (X_2 - X_c) \) = 1056. When \( X_c > X_2 \) then one finds that \( \Pi_2 = 456 + 50 (25 - X_c) \). With these settings we increase the hop size from 1 to 2, 3 and 4. Under such conditions one finds the 1D diffusion coefficient as \( D_1 = (k + 1) (2k + 1) / 6 \). When \( X_2 < 13 \) then using equation (12) one can compute the MFPT for \( k = 2 \) as \( \Pi_5 \approx 472 \). For \( k = 3 \) one finds that \( \Pi_6 \approx 293 \) and for \( k = 4 \) we find that \( \Pi_6 \approx 216 \). The simulation results depicted in figure 2(C) are consistent with these theoretical predictions by equation (10) for \( m = 1 \).

To check the validity of equation (10) for many number of traps i.e. \( m > 1 \) we introduced traps at arbitrary locations \( X_c = \{3, 5, 8, 11, 15, 18, 21, 24\} \) i.e. \( m = 8 \). The corresponding dwell times for these traps were set at \( \rho_c = \{5, 15, 25, 35, 45, 55, 65, 75\} \) and iterated \( X_2 \) from 1 to 24. This means that \( n \) will vary depending on \( X_2 \). For \( X_2 = 1 \) one finds that \( n = 0 \), for \( X_2 = 12 \) one finds that \( n = 4 \) and so on. The simulation results for \( X_c = 0, X_c = X_2 = 25 \) are shown in figure 2(C) which agree well with the theoretical predictions by equation (10). The effects of various types of correlation between the variables \( (X_R - X_m, \rho_c) \) are shown in figures 2(C)–(E). Here minimum value of MFPT can occur only when there is a negative correlation between these variables as shown in figure 2(C). The MFPT will be a maximum when there is a positive correlation between these variables (figure 2(D)). The MFPT corresponding to the case of random arrangement of dwell times will be somewhere in between these two extreme cases (figure 2(E)). Figures 2(F) and (G) shows the effects of variation in \( X_R \) and \( X_2 \) over the MFPT associated with the 1D random walker to escape into \( X_R \) starting from \( X_2 \) in the presence of \( m \) number of traps and various hop sizes \( k > 1 \). These stochastic random walk simulation results are consistent with the theoretical predictions by equations (10) and (11) which suggest that the overall MFPT will be insensitive to the variations of \( X_2 \) at higher hop sizes.

### 3.2. Naturally occurring TF binding sites and their traps

Let us consider a genomic DNA of size \( N \) bps containing a SBS (CRM) on it for a TF protein of interest. We denote the binding stretch of TFs as \( q \) bps and assume that the probability associated with the occurrence of each base \( A, T, G \) and \( C \) in the genome is \( \frac{1}{4} \). With this setting one can calculate the minimum required value of \( q = q_C \) to ensure that there is only one copy of the SBS in the entire genome by chance as follows

\[
s = (N - q)(1/4)^q; \quad s - 1 = 0;
\]

\[
q_C = (2N - \text{LambertW}(2^{N+1} \ln 2)) / 2 \ln 2.
\]  

In this equation \( s \) is the number of similar binding sites with size of \( q \) bps which can occur by chance and \( y = \text{LambertW}(x) \) is the solution of \( y \exp(y) = x \). When \( q < q_C \) for a TF binding site then there is a definite chance of random occurrence of similar binding sites on the genome. One should also note the asymptotic result for a sufficiently large value of genome size \( N \) as \( q_C \sim N / 2 \ln 2 \). Size of haploid human and mouse genomes seem to be \( \sim 3.3 \times 10^9 \) and \( \sim 2.8 \times 10^9 \) bps respectively \([2, 48, 49]\). The corresponding critical length of binding stretch \( q_C \) calculated from equation (16) for human and mouse genomes will be \( \sim 15.8 \) and \( \sim 15.7 \) bps respectively (figures 3(A) and (B)). Statistical analysis of the binding stretches of TFs in human and mouse obtained from JASPAR database \([50]\) suggested the mean values of \( q \) as \( \sim 13 \pm 0.43 \) and \( \sim 12.77 \pm 0.43 \) bps respectively at a confidence level of 0.95. The median of the length of binding stretch of TFs in human seems to slightly higher (\( \sim 13 \) bps) than the case of TFs in mouse (\( \sim 12 \) bps). These results clearly suggest that \( q < q_C \) for the genomes of human and mouse. This means that there is a nonzero probability of random occurrence of binding sites similar to that of SBSs corresponding to TFs on the same genomic sequence.

Using equation (11) we have already shown that the delay in the overall search time due to the presence of traps will be minimum when there is a negative correlation between the dwell times \( \rho_e \) and the corresponding distances of traps \( (X_R - X_m) \) from SBSs. This in turn predicts that an efficient configuration of the SBSs with traps should be such that those traps with high affinity for TFs should be closer to the original SBS than those traps with low affinity towards TFs. To check the validity of this prediction in natural system we computed the distribution of distances of traps with various binding affinities from the putative SBSs for various TFs in the upstream sequences of various genes of mouse and human.

The upstream 5000 bps sequences of various genes of human and mouse genomes were obtained from UCSC genome database (February 2009 assembly, hg19 version for human genome and December 2011 assembly, mm10 version of mouse genome) and position weight matrices (PWMs) \([51, 52]\) of various TFs of mouse and human were obtained from publicly available JASPAR database \([50, 53]\). There were 28 365 sequences from mouse genome and 28 824 sequences from human genome. Using the PWMs of various TFs we generated the score table for various upstream sequences based on the following equation \([51]\)
In this equation $S_{v,i}$ is the score value of PWM at $i$th position on $v$th sequence, $q$ is the length of binding stretch of the corresponding TF, $f_b$ is the background probability of observing base $b$ in the corresponding genome, and $f_{b,w}$ is the probability of observing base $b$ at position $w$ of the TF binding site. Here $f_b$ can be calculated from the random sequences corresponding to the given genome. There is a strong positive correlation between the score value and the binding free energy of TFs [51]. In parallel we also generated score table for random sequences using the same PWM from which we obtained the score distribution and the cutoff score value for the given matrix corresponding to a given $p$-value. In our calculations we have set the $p$-value $< 10^{-6}$ for putative SBSs. Sample score table associated with the upstream sequence of mouse Fibin gene is depicted in figure 3(C) and the corresponding score distribution is shown in figure 3(D). Here we have used the PWM corresponding to the mouse TF protein POU2F1a. We defined those lattice positions with score values higher than the cutoff score value with respect to $p$-value $< 10^{-6}$ as putative SBSs of the corresponding TFs. We further defined those sites whose score values corresponding to the $p$-values ranging from $10^{-4}$ to $10^{-6}$ as kinetic traps associated with the SBSs. Based on these definitions we computed the probable distances of traps from the putative SBSs of various TFs. We combined these distances for various TFs and generated the overall histograms for the entire genome with a bin size of $\sim 200$ bps. We used the random sequences associated with each genome that is available at UCSC database to compute the probability of occurrence of putative binding sites by chance. We considered random sequences of size $5 \times 10^6$ bps and fragmented it into $10^6$ number of sequences with length of 5000 bps. Then we scanned each random sequence with same PWM and obtained the number

$$S_{v,i} = -\sum_{b=1}^{q} \left( \sum_{w=1}^{T} f_{b,w} \log_e \left( \frac{f_{b,w}}{f_b} \right) \right);$$

$$b = \{A, C, G, T\}. $$
of putative SBSs (false positives). The probability of observing a SBS by chance will be calculated as
\[ p_{NF} = \frac{\text{number of false positives}}{1000}. \]

These distances were obtained by scanning the upstream 5000 bps sequences of various genes of human and mouse with PWMs available for various human and mouse TFs in JASPAR database. For each combination of PWM and upstream sequence a score table was constructed and the cutoff scores for specific and trap sites were computed. In all these cases the specific binding site was defined by the \( p \)-value < 10\(^{-6} \). In (A) and (B), the traps were defined by \( p \)-value < 10\(^{-3} \). In (C) and (D), the traps were defined by \( p \)-value < 10\(^{-4} \). In (E) and (F) the traps were defined by \( p \)-value < 10\(^{-5} \). These results clearly suggest that traps with strong affinity towards TFs are preferably located near the specific binding site which is in line with the prediction of equations (10) and (11).

4. Discussion

Binding of TFs at their CRMs is essential for tight regulation of the intracellular levels of several proteins.
The rate of site-specific binding of TFs seems to be influenced by the conformational state of DNA, fluctuations in the conformational states of DBDs of TFs, overall electrostatic interactions at the DNA-protein interface, presence of other roadblock proteins on the same DNA and sequence mediated kinetic traps. The nonspecific binding of TFs on DNA is mainly driven by the electrostatic forces present in between the positively charged amino acid side chains of DBDs of TFs and negatively charged phosphate backbone of DNA. It seems that positively charged protein residues are much more abundant especially near the protein DNA interface. The positions of individual DNA phosphates correlates well with the positive protein charges in proximity of DNA [19].

Traps are those sequences which share significant amount of similarity with SBS. Presence of sequence traps always slow down the 1Dd mediated search process of TFs towards their cognate sites on DNA in a position dependent manner. When the traps are located in between the initial position of TFs on DNA and SBS, then the degree of influence of such traps on the site-specific binding rate will be directly proportional to the distance of traps from SBS. When the traps are not located in between the initial position of TFs on DNA and SBS, then the degree of influence of such traps on the site-specific binding rate will be directly proportional to the distance of initial position of TFs from SBS. This is reasonable since there is always a nonzero probability associated with TFs to reach SBS without visiting the traps in the latter case. However such probabilities approach towards zero as the distance between the initial position of TFs on DNA and SBS increases.

Equations (10) and (11) suggest that the degree of influence of sequence traps on the site-specific association rate of TFs will be minimum when the traps are arranged around SBS such that there is a negative correlation between the distance of traps from SBS and their thermodynamic binding affinities towards TFs. Such arrangement of sequence traps will in turn generate a concentration profile of TFs with peaks near SBS. This may play critical roles in manipulating the nucleosome occupancy pattern near SBS. Remarkably the distribution of sequence traps around the putative binding sites of various TFs of human and mouse genome showed a negative correlation between the distance of traps and their binding affinities for most of the TFs. This result substantiates our theoretical predictions (figure 4).

From equation (15) one can conclude that the retarding effects of randomly occurring sequence traps can be minimized when the length of recognition stretch $q$ increased beyond the critical value $q_c$. However this will not be an efficient one since there is an anti-correlation between the speed and fidelity associated with the recognition of SBS by TFs. In this context one should note that the fidelity of RNA polymerase II (RNAPII) can be enhanced by the combinatorial binding of TFs at the corresponding cis-regulatory motifs [28] and subsequent distal action on the RNAPII-promoter complex via looping or tracking mechanism [54, 55].

4.1. Effects of traps which are identical to SBSs

In this section we consider a situation where there are two identical SBSs viz. CRM1 and CRM2 corresponding to the same TF (1D random walker) present on the DNA polymer with a distance of $u$ bps from each other as shown in figure 5 [37, 40]. When the mode of site-specific binding of corresponding TFs is via pure 3Dd route then the overall specific binding rate will be independent of the distance between those SBSs. Experimental results clearly corroborated the strong dependency of the overall site-specific binding rate on $u$ [37]. Since both are SBSs for the same TFs, the dwell
time of TFs at these binding sites will be significantly high. Therefore each site can also generate a potential roadblock for the 1Dd movement of TFs towards the other side rather than merely acting as kinetic traps. This will clearly decreases the degrees of freedom associated with the approach of TFs towards the SBSs from left and right sides of CRMs via 1Dd from 2 to 1. As a consequence the influx of TFs especially from the right side of CRM1 as well as left side of CRM2 will be blocked by the TFs which are already bound on the respective CRMs. This will in turn decreases the effective number of SBSs from 2 to 1 especially when $u = 0$. This is because the probability of TFs entering CRM1 from left side ($p_L$) will be $\frac{1}{2}$ and the probability of TFs entering into CRM2 from the right side ($p_R$) will be $\frac{1}{2}$ so that the total probability flux will be 1 since the probability flux through the interconnecting region is zero i.e. $P(0) = 0$.

Now we consider the influx of TFs through the interconnecting region between CRM1 and CRM2. Let $u$ denote the average 1D length of TFs on DNA as $u_0$. One can compute the splitting probabilities associated with the TFs landing at the interconnecting region to reach any of the CRMs only when $u$ is comparable or less than than that of $u_0$. When $u$ is much higher than $u_0$ then there is always a nonzero probability of TFs to dissociate from interconnecting region without meeting any of the CRMs. Alternatively the extent of roadblock will decrease as $u$ increases towards infinity which clearly uncouples these SBSs. In other words the incremental change in the probability of a TF which landed at the interconnecting region $p(u)$ to reach any one of the SBSs with respect to $u$ will be (a) inversely proportional to $u$ and (b) directly proportional to the product of probabilities $p(u)(1 - p(u))$. This is because when $u$ increases, then the extent of dynamic coupling of these CRMs via 1Dd of TFs decreases. When $u = u_0$ then the average (over the landing or initial position) splitting probabilities associated with the TFs landing at the interconnecting region to reach any one of the SBSs will be $P(u_0) = \frac{1}{2}$. From these observations one can derive the following result

$$dp(u)/du = p(u)(1 - p(u))/u;$$

$$p(u_0) = 1/2; \cdot \cdot \cdot p(u) = u/(u_0 + u).$$

From equation (17) one can conclude that the total number of effective binding sites as seen by TFs from bulk will be $b(u) = p_L + p(u) + p_R$. Clearly $b(0) = 1$, $b(u_0) = 3/2$ and $b(\infty) = 2$. These results are consistent with the experimental observations [37, 40]. These results suggest that the presence of similar binding sites adjacent to SBSs significantly retards the 1Dd dynamics of TFs mainly via introducing stationary roadblocks rather than transiently trapping TFs.

Here one should note that equation (17) will be valid only for prokaryotic systems since we have not considered the presence of nucleosomes. Detailed observation of the experimental data [37, 40] on prokaryotic system suggest that the roadblock effects of adjacent SBSs like traps will be significant when the distance between them will be $< 150$ bps [16]. This means that the retarding effects of traps via roadblock mode cannot be ignored even in the presence of semi-stationary roadblocks such as nucleosomes in eukaryotes. This is because (a) the length of the linker DNA between two consecutive nucleosomes will be $\sim 100$ bps [56, 57] and (b) the dwell time of TFs at SBS like traps will be much higher than the lifetime of the nonspecifically bound nucleosome complexes.

4.2. Effects of nucleosomes and other roadblocks

The retarding effects of sequence traps on the site-specific binding of TFs will be significantly modulated by the nucleosome complexes present especially on the eukaryotic genomic DNA [21–25]. Nearly 147 bps of genomic DNA wrap around each nucleosome and subsequently become inaccessible for the binding of most of the TFs. Nucleosome complexes influences the overall dynamics associated with the site-specific binding of TFs at least in two different ways viz. they (a) dynamically control the accessibility of SBSs, sequence traps and significant fraction of nSBS towards the inflowing TFs from the bulk and (b) introduce semi-stationary roadblocks across the 1Dd dynamics of TFs along DNA [22, 23]. When SBSs are hidden by the nucleosome complex, then TFs need to wait until the dissociation of nucleosome complex. This dissociation time will be then added up to the overall search time associated with the TFs to locate their SBSs [16].

Depending on the relative timescales associated with the 1Dd of TFs or kinetics of nucleosome dissociation, the underlying process will be either diffusion-controlled or kinetic-controlled. The overall time required by TFs to locate their cognate sites will be approximately equal to the sum of the random search time and the time required for the dissociation of nucleosomes. This is similar to the case of coupled transcription and splicing [58] where the small nuclear ribonucleoproteins (snRNPs) which are performing 1Dd on the already emerged out segment of pre-mRNA need to wait until the emergence of splicing sites from the transcription assembly to initiate the splicing process.

Although the presence of nucleosomes reduces the accessibility of SBSs and sequence traps towards the inflowing TFs, the enhancement over 3Dd controlled rate via non-specific binding mechanism will also be decreased i.e. presence of nucleosomes increases the overall search time associated with TFs to find their targets. In this context one should note that around 81% of the yeast genome seems to be occupied by nucleosome complexes [57]. The length of linker DNA present in between two consecutive nucleosomes seems to vary from 10 to 90 bps [59, 60]. This means
that the average mean free path lengths associated with the 1Dd of TFs along the genomic DNA of eukaryotes will be in the range of ~10–100 bps and the chromatin remodeling factor such as ACF (ATP-dependent chromatin-remodeling factor) is required to increase the accessible DNA lengths.

Here one should also note that nucleosome-depleted regions of genomic DNA correlates well with the transcription start sites on genome level [57]. Presence of significant amount of site-specific protein–DNA interactions around the transcription start site could be one of the reasons for such nucleosome-depletion. Actually nucleosomes need to compete for the freely available DNA segments around the transcription start site with other DBPs involved in the transcription initiation. In this context one can conclude that the presence of traps around the SBSs of TFs can also cause nucleosome-depletion at those regions. This is mainly because the sequence traps are also weak binding sites for the same TFs. As a result one can expect a negative correlation between the distance from SBSs and the extent of nucleosome occupancy. In line with this prediction, detailed analysis on the nucleosome occupancy pattern around SBSs of various TFs of human genome shows a minimum around SBSs with well positioned nucleosomes around such minimum [56].

4.3. Ruggedness of sequence energy landscape
Here one should note from figure 3(C) that apart from the putative SBSs and traps there are background fluctuations in the score values. Since score values are positively correlated with the binding energy, these background fluctuations form an integral part of the sequence dependent binding energy landscape for the TFs and upstream sequence of interest. A strong frustration in the free energy profile of a sequence significantly influences the 1Dd dynamics of TFs since the microscopic transition rates \( w_{ij} \) in the definition of 1Dd are strongly dependent on the ruggedness of the binding energy landscape. Earlier studies suggested that \( D_0 \simeq D_0^c \exp(-\epsilon^2) \) [61–63] where \( D_0^c \) is the 1Dd coefficient corresponding to a smooth free energy landscape and \( \epsilon \) is the degree of frustration or ruggedness (measured in \( k_BT \)) in the free energy profile of the DNA sequence. In this context one should note that the presence of traps further increases the degree of frustration in the binding free energy profiles of DNA sequence. From equations (10) and (11) we found that the retarding effects of such rough sequence potentials can be well handled by increasing the hop size associated with the 1Dd dynamics of TFs that in turn requires a condensed conformational state of DNA [16].

5. Conclusions
In this paper we have investigated the effects of sequence traps on the site-specific binding of TFs at their respective CRMs. We have shown that the speed of site-specific binding of TFs with DNA seems to be strongly retarded by the randomly occurring sequence traps. We developed a simple random walk model on the site-specific binding of TFs with genomic DNA in the presence of sequence traps. Our dynamical model predicted that (a) the retarding effects of traps will be minimum when the traps are arranged around the SBS such that there is a negative correlation between the binding strength of TFs with traps and the distance of traps from the SBS and (b) the retarding effects of sequence traps can be soothed by the condensed conformational state of genomic DNA. Our computational analysis results on the distribution of sequence traps around the putative binding sites of various TFs in mouse and human genome clearly agree well with the theoretical predictions. We proposed that the distribution of traps can be used as an additional metric to efficiently identify the SBSs of TFs. It seems that the presence of similar binding sites adjacent to SBSs significantly retards the 1Dd dynamics of TFs mainly via introducing stationary roadblocks rather than transiently trapping TFs.

References
[1] Alberts B 2002 Molecular Biology of the Cell (New York: Garland Science)
[2] Levin B, Krebs J E, Kilpatrick S T, Goldstein E S and Lewin B 2011 Lewin’s Gene X (Sudbury, MA: Jones and Bartlett)
[3] Ptashne M 1986 A genetic switch: Gene Control and Phage [Lambda] (Cambridge, MA: Cell)
[4] Ptashne M and Gann A 2002 Genes & Signals (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory)
[5] Watson J D 1987 Molecular Biology of the Gene (Menlo Park, CA: Benjamin/Cummings)
[6] Atkins P W 1978 Physical Chemistry (San Francisco: Freeman)
[7] Berg O G, Winter R B and von Hippel P H 1981 Diffusion-driven mechanisms of protein translocation on nucleic acids: I. Models and theory Biochemistry 20 6929–48
[8] Winter R B, Berg O G and von Hippel P H 1981 Diffusion-driven mechanisms of protein translocation on nucleic acids: III. The Escherichia coli lac repressor—operator interaction: kinetic measurements and conclusions Biochemistry 20 6961–77
[9] Murugan R 2004 DNA–protein interactions under random jump conditions Phys. Rev. E 69 011911
[10] Murugan R 2007 Generalized theory of site-specific DNA–protein interactions Phys. Rev. E 76 011901
[11] Murugan R 2006 Critical jump sizes in DNA–protein interactions Biophys. Chem. 120 143–7
[12] Bonnet I, Biebricher A, Porte P L, Loverdo C, Benichou O, Voituriez R, Escude C, Wende W, Pingoud A and Deshioles P 2008 Sliding and jumping of single EcoK restriction enzymes on non-cognate DNA Nucleic Acids Res. 36 4118–27
[13] Koslover E F, Diaz de la Rosa M A and Spakowitz A J 2011 Theoretical and computational modeling of target–site search kinetics in vitro and in vivo Biophys. J. 101 836–65
[14] Bauer M and Metzler R 2013 In vivo facilitated diffusion model PLoS One 8 e53956
[15] van den Broek B, Lombolt M A, Kalisch S M, Metzler R and Wuite G J 2008 How DNA coil enhances target localization by proteins Proc. Natl. Acad. Sci. USA 105 15738–42
[16] Niranjan G and Murugan R 2016 Generalized theory on the mechanism of site-specific DNA–protein interactions J. Stat. Mech. 053501
[17] Chernysyvych A G, Kolomeisky A B and Korshnev A A 2008 Protein—DNA interactions: reaching and recognizing the targets J. Phys. Chem. B 112 4741–50
[18] Kolomeisky A B 2011 Physics of protein–DNA interactions: mechanisms of facilitated target search Phys. Chem. Chem. Phys. 13 2088–95
[19] Chernysyvych A G 2009 Positively charged residues in DNA-binding domains of structural proteins follow sequence-specific positions of DNA phosphates groups J. Phys. Chem. B 113 4242–7
[20] Bauer M, Rasmussen E S, Lombolt M A and Metzler R 2015 Real sequence effects on the search dynamics of transcription factors on DNA Sci. Rep. 5 10072
[21] Beshnova D A, Chernysyvych A G, Vainsh tein Y and Teif V B 2014 Regulation of the nucleosome repeat length in vivo by the DNA sequence, protein concentrations and long-range interactions PLoS Comput. Biol. 10 e1003695
[22] Schoplin R, Teif V B, Muller O, Weinberg C, Rippe K and Wedemann G 2013 Modeling nucleosome position distributions from experimental nucleosome positioning maps Bioinformatics 29 2380–6
[23] Teif V B and Rippe K 2011 Nucleosome mediated crosstalk between transcription factors at eukaryotic enhancers Phys. Biol. 8 044001
[24] Parmar J J, Das D and Padianhiaeeri R 2016 Theoretical estimates of exposure timescales of protein binding sites on DNA regulated by nucleosome kinetics Nucleic Acids Res. 44 1630–41
[25] Parmar J J, Marko J F and Padianhiaeeri R 2014 Nucleosome positioning and kinetics near transcription–start–site barriers are controlled by interplay between active remodeling and DNA sequence Nucleic Acids Res. 42 128–36
[26] Li G–W, Berg O G and Elj 2009 Effects of macromolecular crowding and DNA looping on gene regulation kinetics Nat. Phys. 5 294–7
[27] Murugan R 2009 Directional dependent dynamics of protein molecules on DNA Phys. Rev. E 79 041913
[28] Murugan R 2010 Theory of site-specific interactions of the combinatorial transcription factors with DNA J. Phys. A: Math. Theor. 43 195003
[29] Murugan R 2010 Theory of site-specific DNA–protein interactions in the presence of conformational fluctuations of DNA binding domains J. Phys. A: Math. Theor. 43 045002
[30] Kalodimos C G, Biriz N, Bonvin A M, Levadoski M M, Guennuegues M, Boelens R and Kaptein R 2004 Structure and dynamics of the DNA major groove J. Mol. Biol. 342 1441–52
[31] Lange M, Kochugaeva M and Kolomeisky A B 2015 Dynamics of the protein search for targets on DNA in the presence of DNA traps J. Phys. Chem. B 119 12410–6
[32] Marcovitz A and Levy Y 2011 Frustatjon in protein–DNA binding influences conformational switching and target search kinetics Proc. Natl. Acad. Sci. USA 108 17957–62
[33] Marcovitz A and Levy Y 2013 Weak frustatjon regulates sliding and binding kinetics on rugged protein–DNA landscapes J. Phys. Chem. B 117 13005–14
[34] Wunderlich Z and Mirny L A 2008 Spatial effects on the speed and reliability of protein–DNA search Nucleic Acids Res. 36 5279–8
[35] Murugan R 2009 Packageing effects on site-specific DNA–protein interactions Phys. Rev. E 79 061920
[36] Murugan R 2011 Theory on thermodynamic coupling of site-specific DNA–protein interactions with fluctuations in DNA-binding domains J. Phys. A: Math. Theor. 44 505002
[37] Mahmutovic A, Berg O G and Elj 2015 What matters for lac repressor search in vivo—sliding, hopping, intersegment transfer, crowding on DNA or recognition? Nucleic Acids Res. 43 3454–64
[38] Gardner CW 1985 Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences (Berlin: Springer)
[39] Yang W Y and Gruebele M 2003 Folding at the speed limit Nature 422 918–22
[40] Hammar P, Wallen M, Fange D, Persson F, Ballekin O, Ullman G, Leroy P and Elj 2014 Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation Nat. Genet. 46 405–8
[41] Risken H 1989 The Fokker–Planck Equation: Methods of Solution and Applications (Berlin: Springer)
[42] Kampen N G V 1981 Stochastic Processes in Physics and Chemistry (Amsterdam: Elsevier)
[43] Mazo R M 2002 Brownian Motion: Fluctuations, Dynamics, and Applications (Oxford: Oxford University Press)
[44] Murugan R 2006 Mean first passage time calculation for the random walk with random step size J. Phys. A: Math. Gen. 39 1575
[45] Murugan R 2006 Mean first passage time calculation for a one-dimensional random walk with random absorbing boundary J. Phys. A: Math. Gen. 39 199
[46] Gillespie D T 2007 Stochastic simulation of chemical kinetics Annu. Rev. Phys. Chem. 58 35–55
[47] Gillespie D T 1977 Exact stochastic simulation of coupled chemical reactions J. Phys. Chem. 81 2340–61
[48] Kimball J W 1965 Biology (Palo Alto, CA: Addison-Wesley)
[49] Campbell N A and Reece J B 2005 Biology (San Francisco, CA: Pearson)
[50] Sandelin A, Alkema W, Engstrom P, Wasserman W W and Lenhard B 2004 JASPAR: an open-access database for eukaryotic transcription factor binding profiles Nucleic Acids Res. 32 1991–4
[51] Stormo G D 2000 DNA binding sites: representation and discovery Bioinformatics 16 16–23
[52] Kreiman G 2004 Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes Nucleic Acids Res. 32 2889–900
[53] Mathelier A et al 2016 JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles Nucleic Acids Res. 44 D110–5
[54] Murugan R 2010 Theory on the mechanism of distal action of transcription factors: looping of DNA versus tracking along DNA J. Phys. A: Math. Theor. 43 415002
[55] Murugan R 2007 Action at distance along the DNA: a random jump model on enhancer action J. Theor. Biol. 248 696–705
[56] Wang J et al 2012 Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors Genome. Res. 22 1798–812
[57] Lee W, Tillo D, Bray N, Morse R H, Davis R W, Hughes T R and Nislow C 2007 A high-resolution atlas of nucleosome occupancy in yeast Nat. Genet. 39 1235–44
[58] Murugan R and Kreiman G 2012 Theory on the coupled stochastic dynamics of transcription and splice-site recognition PLoS Comput. Biol. 8 e1002747
[59] Szerlong H I and Hansen J C 2011 Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure Biochem. Cell Biol. 89 24–34
[60] Van Holde K E 1989 Physical Biochemistry (New York: Springer)
[61] Zwanenburg R 1999 Protein complexes with dynamical disorder Acc. Chem. Res. 32 148–52
[62] Zwanenburg R 1988 Diffusion in a rough potential Proc. Natl. Acad. Sci. USA 85 2029–30
[63] Sltusky M and Mirny L A 2004 Kinetics of protein–DNA interaction: facilitated target location in sequence-dependent potential Biophys. J. 87 4021–35