Theory on the looping mediated directional-dependent propulsion of transcription factors along DNA

R. Murugan

Department of Biotechnology, Indian Institute of Technology Madras, Chennai, India

Phone: +91 44 2257 4116. E-mail: rmurugan@gmail.com

ABSTRACT

We show that the looping mediated transcription activation by the combinatorial transcription factors (TFs) can be achieved via directional-dependent propulsion, tethered sliding and tethered binding-sliding-unbinding modes. In the propulsion mode, the first arrived TF at the *cis*-regulatory motifs (CRMs) further recruits other TFs via protein-protein interactions. Such TFs complex has two different types of DNA binding domains (DBDs) viz. DBD1 which forms tight site-specific complex with CRMs via hydrogen bonding network and the promoter specific DBD2s which form nonspecific interactions around CRMs. When the sum of these specific and cumulative nonspecific interactions is sufficient, then the flanking DNA of CRMs will be bent into a circle over the TFs complex. The number of TFs involved in the combinatorial regulation plays critical role here. When the site-specific interactions and the cumulative nonspecific interactions are strong enough to resist the dissociation, then the sliding of DBD2s well within the Onsager radius associated with the DBD2s-DNA interface towards the promoter is the only possible way to release the elastic stress of the bent DNA. The DBD2s form tight synaptosome complex upon finding the promoter via sliding. When the number of TFs is not enough to bend the DNA in to a circle, then tethered sliding or tethered binding-sliding-unbinding modes are the possibilities. In tethered sliding, the CRMs-TFs complex forms nonspecific contacts with DNA via dynamic loops and then slide along DNA towards promoter without dissociation. In tethered binding-sliding-unbinding, the CRMs-TFs performs several cycles of nonspecific binding-sliding-unbinding before finding the promoter. Elastic and entropic energy barriers associated with the looping of DNA shape up the distribution of distances between CRMs and promoters. The combinatorial regulation of TFs in eukaryotes has evolved to overcome the looping energy barrier.

KEY WORDS

DNA-protein interactions, combinatorial binding, transcription factor binding, DNA-looping, *cis*-regulatory modules

INTRODUCTION

Site-specific binding of transcription factors (**TFs**) at their *cis*-regulatory motifs (**CRMs**) on the genomic DNA in the presence of enormous amount of nonspecific binding sites is essential for the activation and regulation of several genes across prokaryotes to eukaryotes (1-3). Binding of TFs with their CRMs was initially thought as a single-step three-dimensional (**3D**) diffusion-controlled collision process. Kinetic experiments on *lac*-repressor-Operator system revealed a bimolecular rate in the order of ~10^9-10^{10} $M^{-1}s^{-1}$ that is ~10-10^2 times faster than the Smoluchowski type 3D diffusion-controlled rate limit (4). Berg et.al. (5, 6) successfully explained this inconsistency using a two-step mechanism by establishing the key concept that TFs first bind with DNA in a nonspecific manner via 3D diffusion and then search for their cognate sites via various one-dimensional (**1D**) facilitating processes such as sliding, hopping and intersegmental transfers. Here 1D diffusion with unit base-pair step-size of TFs is the sliding, few base-pairs (bp, 1 bp = l_d ~ 3.4 x 10^{-10} m) step-size is called hopping and few hundred to thousand bps step-size is called intersegmental-transfer. Intersegmental transfers occur whenever two distal segments of the same DNA polymer come in nearby over 3D space via ring closure events (7-9).

Specific binding of TFs with DNA is affected by several factors (9) viz. **a**) conformational state of DNA (9, 10) **b**) spatial organization of various functionally related combinatorial CRMs along the genomic DNA (11, 12), **c**) presence of similar or other dynamic roadblock proteins (13) and semi-stationary roadblocks such as nucleosomes in eukaryotes (14-17), **d**) naturally occurring sequence mediated kinetic traps on DNA (18, 19), **e**) conformational fluctuations in the DNA binding domains of TFs (20-22) and **f**) the nonspecific electrostatic attractive forces and the counteracting shielding effects of other solvent ions and water molecules acting at the DNA-protein interface (23). Several theoretical models (8, 9, 18, 21, 24), computational (25-28) and experimental studies (29) have been carried out to understand the effects of factors **a-f** on the kinetics of site-specific DNA-protein interactions.

In general, the searching efficiency of TFs depends on the relative amount of times spent by them on the 3D and 1D diffusions (11, 21). Clearly, neither pure 1D nor 3D diffusion is an efficient mode of searching (9, 11). Under ideal situation, maximum searching efficiency can be achieved only when TFs spend equal amount of times in both 1D and 3D diffusions (8, 11). This trade off balance between the times spent on different modes of diffusions will be modulated by the factors **a-f**. For example, presence of nucleosome roadblocks warrants more dissociations and 3D excursions of TFs rather than 1D sliding (14). Sequence specific fast conformational switching of DNA binding domains between stationary and mobile states helps TFs to overcome the sequence traps (14). Relaxed conformational state of DNA enhances more sliding rather than hopping and intersegmental transfers and so on (9). Conformational dynamics of DNA also modulates the speed of gene activation and regulation. In this context, looping of DNA is critical for the activation and expression of various genes across prokaryotes to eukaryotes (3, 30-34). Combinatorial binding of TFs with their specific CRMs on the genomic DNA activates the downstream promoters of genes via looping of the intervening DNA segment to form a synaptosome type complex (1, 35). In most of the molecular biological processes, DNA-loops are warranted for the precise protein-protein interactions which are the prerequisites for transcription and recombination (36).

The statistical mechanics of looping and cyclization of linear DNA has been studied extensively in the literature (33, 37, 38). However, it is still not clear why DNA-loops have evolved as an

integral part of the activation and repression of transcription and recombination although such underlying site-specific protein-protein and protein-DNA interactions can also be catered straightforwardly via a combination of 1D and 3D diffusions of TFs (5, 6, 21, 39). That is to say, upon arrival at the CRMs, TFs can directly slide or hop along the DNA polymer to reach the promoters. Schleif (31) had argued that the looping of DNA can simplify the evolution of the genomic architecture of eukaryotes by not imposing strict conditions on the spacing between the TF binding sites and the promoters. This is logical since a given set of TFs need to regulate several different genes across the genome. Therefore, placement of TF binding sites near a specific gene can be a disadvantage for other genes along the genomic evolution. Similarly, placement of TF binding sites near every gene is not an efficient genome design. The DNA loops also play critical roles in the transcription bursting (40) and memory (41). It is still not clear about **a**) the purpose of DNA loops in the transcription activation and **b**) the exact mechanism by which the DNA-loop is formed between the CRMs and promoters via TFs though Rippe et.al., (32) had already taken several snapshots of the looping intermediates. In this paper, we will show that the DNA-loops can propel TFs towards the promoters. Using computational tools, we further demonstrate that the looping mediated propulsion of TFs along DNA can actually help in finding the direction of the promoter region and also shape up the genomic architecture.

THEORETICAL METHODS

DNA loops mediated transcription activation in eukaryotes

Let us first list out the basic facts on the mechanism of distal action of CRMs-TFs system on the downstream promoters in the process of transcription activation.

a) Both theoretical investigations (5, 6, 8, 9, 21) and experimental observations (42, 43) suggested that TFs recognize their CRMs via a combination of 1D and 3D diffusions. The key idea here is that TFs scan a random piece of DNA via 1D diffusion after each of the 3D diffusion mediated nonspecific collisions (5, 9, 21). In contrast, the reacting molecules dissociate immediately upon each of their unfruitful collisions in the standard Smoluchowski model. When the dynamics of TFs is confined well within the *Onsager radius* of the DNA-protein interface, then it is categorized as the 1D diffusion. When TFs escape out of the Onsager radius and perform free 3D excursions, then we classify it as the 3D diffusion (9). The ***Onsager radius*** connected with the DNA-protein interface is defined (9) as the distance between the positively charged DNA binding domains of TFs and the negatively charged phosphate backbone of DNA at which the overall electrostatic interaction energy is same as that of the background thermal energy (equals to ~1 k_BT) (**Section 1**, Supporting Materials). Further, one can define the dissociation of site-specific CRMs-TFs complex as the thermally induced separation of TFs from CRMs out of the Onsager radius associated with their interface.

b) The first arrived one of the combinatorial TFs further recruits other TFs (3, 44) around CRMs mainly via the cooperative protein-protein interaction among TFs and nonspecific electrostatic interactions between the DBDs of other TFs and DNA. This results in the formation of the CRMs-TFs complex. This is the well-established recruitment model (3) on the regulatory mechanism of the combinatorial TFs networks (45).

c) Transcription activation is achieved upon the distal communication between the CRMs-TFs complex with the corresponding promoter (1-3).

d) Binding of these combinatorial TFs at CRMs locally bends the DNA. At the end of these processes, synaptosome type DNA-loops connecting CRMs-TFs complex with the promoters are observed in most of the transcriptionally active genes of eukaryotes (3, 44).

Clearly, TFs activate transcription via two sequential steps viz. they bind their CRMs in the first step and then distally communicate with the promoter in the second step. To understand the role of DNA-loops, we consider two possible scenarios viz. looping mediated versus a hypothetical pure 3D1D diffusion mediated distal communication between CRMs-TFs and promoters. In both these scenarios, TFs locate their respective CRMs via a combination of 1D and 3D diffusion in the first step. They differ only in the second step where TFs dissociate from their CRMs and communicate with the promoters via a combination of 1D and 3D diffusions in the second case and, the distal communication will be through DNA-loops in the first case. We denote the search time required by TFs to locate their CRMs in the first step of transcription activation as τ_S. Clearly, those factors **a-f** listed in the introduction section significantly modulate this quantity. We will not recalculate this here since enormous amount of literature already exists (**Section 1** of the Supporting Materials) on the derivation of this quantity under various conditions (9, 18, 21, 46-48). In the following sections, we will compute the mean time required by the CRMs-TFs complex to communicate with the promoter via DNA-loops.

Preliminary assumptions

Upon observing the open synaptic complexes of the transcriptionally active genes of eukaryotes with DNA loops, one can conclude that the combinatorial TFs complex which activates the transcription via DNA-loop has at least two different types DNA binding domains (DBDs) viz. DBD1 that corresponds to CRMs (**Fig. 1**) and DBD2 that corresponds to the promoter region. These two different DBDs may belong to two different TFs of the combinatorial TF complex. For example, in **Fig 1**, the DBD of TF-1 (DBD1 type) specifically binds CRMs and DBD of TF-4 (DBD2 type) specifically binds the promoter. Clearly, DBD1 type domains specifically interact with CRMs via hydrogen bonding network. Instead, the DBD2s of the other combinatorial TFs bind near CRMs region via non-specific electrostatic interactions and they are further stabilized by the protein-protein interaction network. For example, the tetrameric Lac I complex binds two different Operator regions that induces looping of DNA (2, 3). Further, the tetramers of repressor molecules bound at these two different binding sites communicate via protein-protein interactions among them. The DNA-loop is stabilized by an octamer form of the Lac I repressor protein. Such mechanisms are common in case of multiprotein mediated DNA-looping and transcription activation in eukaryotes. We further assume that TFs reach their specific binding site in the first step via a combination of 3D and 1D diffusions (5, 6, 8, 21, 49) in line with two-step site-specific DNA-protein interaction models and subsequently bends the DNA upon binding their specific sites located at the corresponding CRMs (32, 33).

Energetics of the site-specific binding of TFs and bending of DNA

Let us assume that the radius of gyration of the fully assembled combinatorial TFs complex is r_P bp (**Fig. 1B**) and CRMs are located at S1 of DNA. We further assume that there are two different type of DBDs of TFs viz. DBD1 and DBD2. Here DBD1 site-specifically binds CRMs via hydrogen bonding and DBD2s bind near CRMs via nonspecific electrostatic interactions. We

denote the distance between CRMs and the promoter by the variable X (measured in bp). Upon binding the cognate stretch of DNA with size of X_0 bp, the TFs complex bends the DNA segment into a circle around its overall spherical solvent shell surface such that the radius of curvature of the bent DNA segment is same as that of the radius of gyration of the combinatorial TFs complex. In this situation $X_0 \simeq 2\pi r_P$ as shown in **Fig. 1B**. In general, one can write as $X_0 \simeq \delta r_P$ where $0 < \delta \leq 2\pi$. We set $X = 0$ at S1 and $X = X_0$ at the end of the DBD2s of TFs on DNA. Here S1 is the specific site for DBD1 and P is the specific site for DBD2s by definition. Here segment of DNA under consideration spans over the range $(0, L)$ as in **Fig. 1A** and X is the current loop-length. The total energy required to bend a linear piece of DNA will be the sum $E_{\text{bend}} = E_{\text{elastic}} + E_{\text{entropy}}$ and in general $E_{\text{bend}} \geq 0$. For the radius of curvature r_P, one finds that $E_{\text{elastic}} \simeq aX/2r_P^2$ (k_BT units) where a is the persistence length of DNA (37, 50). Clearly, E_{elastic} that is required to bend a DNA segment of length X into a circle will be $E_{\text{elastic}} \simeq 2\pi^2 a/X$. Here one can write $E_{\text{elastic}} \simeq \delta^2 a/2X$ where $0 < \delta \leq 2\pi$. Clearly, E_{elastic} will be at maximum when $\delta = 2\pi$. This energy has to be derived either solely from the binding of TFs on DNA or via an external energy input in the form of ATP hydrolysis (51). Noting that $E_{\text{entropy}} \simeq (3/2)\ln(\pi X/6b)$ where b = 1 bp is the distance between two consecutive nucleotide monomers of DNA, one finally arrives at the following expression for the overall bending energy (**Eq. A1** of **Appendix A**).

$$E_{\text{bend}} \simeq \delta^2 a/2X + (3/2)\ln(\pi X/6b) \tag{1}$$

Clearly, E_{bend} attains a minimum value as $E_{\text{bend}}^{\min} = (3/2)\left[1 + \ln\left(\pi\delta^2 a/18b\right)\right]$ at $X_C = \delta^2 a/3$. In the later sections, we will show that this non-monotonic behavior of the bending energy profile will restrict the possible distances between CRMs and their corresponding promoters.

Looping mediated communication between CRMs-TFs and promoters

Let us assume that the first arrived TF at CRMs recruits other combinatorial TFs via protein-protein interactions. When all the TFs bind their CRMs in the first step of transcription activation and bend the DNA over their spherical solvent shell surface, then the overall *binding energy* (E_{bind}) can be written as $E_{\text{bind}} = E_{\text{bond}} + E_{\text{elastic}} + E_{\text{entropy}}$. In general, one should note that $E_{\text{bond}} \leq 0$ and $E_{\text{elastic}} \geq 0$. Apparently, a stable complex of CRMs-TFs can be formed only when $E_{\text{bind}} < 0$. The bonding energy released at the CRMs-TFs interface is utilized to compensate both the elastic energy required to bend the DNA (E_{elastic}) and the chain entropy loss (E_{entropy}) at the specific binding site. One should note that $E \simeq |E_{\text{bond}}| + E_{\text{elastic}}$ is the overall potential energy barrier which acts against any kind of distortion or dissociation of the CRMs-TFs complex. Here E_{bond} comprises of energies associated with the formation of site-specific hydrogen bonding network (DBD1-S1 interactions), protein-protein interactions among the combinatorial TFs and the overall cumulative non-specific interactions between the DBD2s of TFs and DNA i.e. $E_{\text{bond}} = E_{\text{hydrogen}} + E_{\text{protein-protein}} + E_{\text{electrostatic}}$. Among these three, only the electrostatic interactions can be easily perturbed by the thermally induced fluctuations. Conversely, $E \simeq E_{\text{elastic}} + E_{\text{entropy}}$ is the potential energy barrier that resists the formation of loops out of a linear piece of DNA. With this background, the elastic stress involved in the bent DNA of CRMs-TFs complex can be released via the following four different modes

(types I, II, III and IV) of dynamics depending on the number of TFs (n) involved in the combinatorial regulation. For the comparison purpose, we denote the number of TFs involved in these types of dynamics as $n = (n_I, n_{II}, n_{III}$, and $n_{IV})$.

I. Thermally induced physical dissociation of the entire CRMs-TFs complex along with increase in the overall chain entropy. This mechanism will be frequent especially when $n_I = 1$. However, when the number of TFs involved in the nonspecific electrostatic interactions is high, then their cumulative electrostatic interactions will be strong enough to resist the thermally induced dissociation of DBD2s of the CRMs-TFs complex from DNA. Clearly, when n is sufficiently large then the thermally induced spontaneous dissociation of the CRMs-TFs complex is not possible in the physiologically relevant timescales.

II. Physical dissociation of only DBD2s of TFs complex from DNA and their re-association somewhere else via looping over 3D space (which is resisted by the loop-length dependent potential energy barrier $E \simeq E_{\text{elastic}} + E_{\text{entropy}}$) while the specific interactions at S1-DBD1 of main TF is still intact. This is the *tethered binding-unbinding model* of Shvets and Kolomeisky developed in Ref. (52). In this model, there is no restriction on the initial value of X i.e. the DBD2s of TFs can land anywhere within (0, L). Additionally, the sliding of DBD2s on DNA is not allowed in this model. This model will work only when the nonspecific electrostatic interactions associated with the DBD2s of TFs are weak enough to dissociate frequently. This in turn necessitates a small number of combinatorial TFs.

III. When the number of TFs involved in the nonspecific electrostatic interactions is high enough, then the thermally induced physical dissociation of DBD2s of TFs will be less probable. However, thermally induced sliding of DBD2s is still possible well within the *Onsager radius* associated with the nonspecifically bound regions of DBD2s-DNA interface. Since there is a reflecting boundary condition imposed by the strong DBD1-S1 interactions at CRMs, there is a possibility for the directional-dependent propulsion of the CRMs-TFs complex on DNA via sliding of DBD2s of CRMs-TFs complex towards the promoter. This is achieved through gradual increase in the value of X from X_0 towards L while the specific interactions at S1-DBD1 of main TF is still intact. This is the *directional-dependent propulsion model*.

IV. Tethered sliding of DBD2s of TFs with dynamic loop structure of DNA and intact DBD1-S1 interactions at CRMs. This is similar to type II where sliding of DBD2s of TFs on DNA is allowed before unbinding. This further necessitates a small number of combinatorial TFs (lesser than type III but more than II). Depending on the number of combinatorial TFs involved in the regulation one can consider two different possibilities here viz. a *pure tethered sliding* and *tethered binding-sliding-unbinding*. In pure tethered sliding, the DBD2s of the CRMs-TFs complex forms nonspecific contacts at an arbitrary location on DNA via a dynamic loop structure and then searches for the promoter region via sliding dynamics without dissociation. Here the cumulative electrostatic interactions between the DBD2s and DNA are strong enough to retain the dynamic loop structure until DBD2s reach the promoter. In the tethered binding-sliding-unbinding mode, the DBD2s of CRMs-TFs performs several cycles of binding-sliding-unbinding before finding the promoter region.

One can arrange the number combinatorial TFs involved in II, III and IV types as $n_{II} < n_{IV} < n_{III}$. In the propulsion mechanism, mainly a gradual release of the elastic stress increases the radius of curvature of the bent DNA. This in turn causes bulging of the DNA-loop around the TFs complex as described in **Fig. 1C**. The chain entropy does not increase much here since the intervening DNA is still under loop conformation. This is similar to the sliding of nucleosomes via bulge induced reptation dynamics of DNA (14-16). In general, dissociation of CRMs-TFs complex will be an endothermic process since $|E_{\text{bond}}| > |E_{\text{elastic}}|$. *Clearly*, *when the number of combinatorial TFs is large enough*, *spontaneous physical dissociation of the CRMs-TFs complex will not be the most probable route for the release of elastic stress of the bent DNA polymer*. With this background, the DBD2s of TFs complex need to distally interact with the promoter in the second step and activate the transcription via looping of the intervening DNA segment that connects CRMs and the promoter. In this context, only II, III and IV types of dynamics can be considered here and we rule out the possibility of the type-I dynamics. The repeated binding-unbinding model that is characterized by the type II dynamics of CRMs-TFs complex has been studied in detail by Shvets and Kolomeisky (52). Actually, type II is a special case of type IV dynamics. In the following sections we will consider the possibility of types III and IV in the looping mediated transcription activation in detail i.e. directional-dependent propulsion and tethered sliding models. All the symbols used in this paper are listed in **Table S1** of the Supporting Material.

Directional-dependent propulsion model

The main idea of this model is as follows. Let us assume that the binding energy profile of the combinatorial TFs is such that the specific binding near DBD1-S1 as well as the cumulative nonspecific electrostatic interactions between DBD2s and DNA are strong enough to resist the abrupt dissociation of TFs from DNA for prolonged timescales. In this situation, since the DBD1-S1 is a strong specific binding (reflecting boundary), the elastic stress of the bent DNA of the CRMs-TFs complex can be released only via the thermally driven sliding of DBD2s well within the Onsager radius associated with the DBD2s-DNA interface. When the DBD2s of CRMs-TFs complex slide along DNA, then the elastic stress on DNA will be released in a gradual manner via bulging of the DNA-loop around the CRMs-TFs complex. This in turn propels the CRMs-TFs complex towards the promoter located at L as shown in **Figs. 1B-D**. A schematic representation of the propulsion model is given in **Fig. S1** of the Supporting Materials. Here the water molecules present at the interface of the positively charged DBD2s of TFs and the negatively charged backbone of DNA provide a fluidic type environment for the smooth sliding dynamics of CRMs-TFs complex (8, 21). Upon finding the promoter, CRMs-TFs forms tight synaptosome complex that is required for the transcription activation. Clearly, the following two conditions are essential for the feasibility of propulsion model.

a) The binding energies associated with the specific and nonspecific interactions should be strong enough to bend the CRMs-DNA segment around the TFs complex.
b) The overall cumulative nonspecific binding energies associated with the DBD2s-DNA interactions should be strong enough to sustain the sliding of DBD2s with intact dynamic loop structure of DNA until finding the promoter.

Using detailed calculations, we will show in the later sections that both these conditions can be fulfilled by raising the number of TFs involved in the combinatorial regulation. Although there is

no straightforward experimental evidence for this model, one can construe this idea indirectly from various other experimental studies. Particularly, Rippe et.al (32) have studied NtrC (Nitrogen regulatory protein C) system using the scanning force microscopy. In this study, they had taken snapshots of various intermediary states along the process of transcription activation from the closed to the open promoter complex. In their model system, binding of NtrC at its specific site (CRM) activates the downstream closed complex of *glnA* promoter-RNAP-σ^{54} via looping out of the intervening DNA segment. They have shown that the transition from the inactive-closed form to an active-open promoter complex involved a gradual increase in the bending angle of the intervening DNA. This in turn is positively correlated with an increase in the radius of curvature of the intervening DNA segment which is represented as bulging of the DNA-loop in our propulsion model. Therefore, our assumption that the propulsion of TFs via increase in the radius of curvature of the bent DNA is a logical one. Based on these, the dynamical position X of TF on DNA obeys the following Langevin type stochastic differential equation (53-55).

$$dX/dt = D_C F(X) + \sqrt{2D_C}\Gamma_t;\ \langle \Gamma_t \rangle = 0;\ \langle \Gamma_t \Gamma_{t'} \rangle = \delta(t - t')\ . \tag{2}$$

In **Eq. 2**, $F(X) = -dE/dX = 2\pi^2 a/X^2$ (bp^{-1}) is the force acting on DNA of the CRMs-TFs that is generated by the potential $E \sim E_{\text{elastic}} + |E_{\text{bond}}|$ upon bulging of the loop structure, Γ_t is the Δ-correlated Gaussian white noise and D_C (bp^2/s) is the 1D diffusion coefficient of the sliding of TFs. The energy involved in the bonding interactions will be a constant one so that it will not contribute to the force term. We ignore the energy dissipation via chain entropy of the bulging loop structure mainly because binding of TFs at CRMs attenuates the conformational fluctuations at the CRMs-TFs interface (8, 20, 21). The Fokker-Planck equation describing the probability of observing a given X at time t with the condition that $X = X_0$ at $t = t_0$ can be written as follows (53, 54).

$$\partial_t P(X,t \mid X_0,t_0) = -D_C\left(\partial_X\left[F(X)P(X,t \mid X_0,t_0)\right] - \partial_X^2 P(X,t \mid X_0,t_0)\right). \tag{3}$$

The form of $F(X)$ suggests that it can propel the DBD2s of CRMs-TFs complex only for short distances since $\lim_{X\to\infty} F(X) = 0$ although such limit will be meaningless for $X > 2\pi^2 a$ where E_{elastic} will be close to the background thermal energy. Further, one also should note the fact that the force $F(X)$ changes sign below some critical value of X (56). Initial condition for **Eq. 3** will be $P(X,t_0 \mid X_0,t_0) = \delta(X - X_0)$ where $X_0 = 2\pi r_P$ and the boundary conditions are given as follows.

$$P(L,t \mid X_0,t_0) = 0;\ \left[F(X)P(X,t \mid X_0,t_0) - \partial_X P(X,t \mid X_0,t_0)\right]_{X=X_0} = 0\,. \tag{4}$$

Here X_0 acts as a reflecting boundary for a given size of TFs complex and L is the absorbing boundary where promoter is located. The strong site-specific interactions at the DBD1-S1 of TFs complex act as a reflecting boundary condition. The mean first passage time (MFPT) $T_B(X)$ associated with the DBD2s of CRMs-TFs to reach the promoter location L starting from an arbitrary $X \in (X_0, L)$ obeys the following backward type Fokker-Planck equation along with the appropriate boundary conditions (7, 8).

$$D_C\left[F\left(X\right)d_X T_B\left(X\right)+d_X^2 T_B\left(X\right)\right]=-1;\ T_B\left(L\right)=0;\ d_X T_B\left(X_0\right)=0. \tag{5}$$

The integral solution of **Eqs. 5** can be expressed as follows.

$$T_B\left(X\right)=\left(2\pi^2 a/D_C\right)\int_X^L\left\{\begin{array}{l} Z/2\pi^2 a+\exp\left(2\pi^2 a/Z\right)\left[\mathrm{Ei}_1\left(2\pi^2 a/X_0\right)-\mathrm{Ei}_1\left(2\pi^2 a/Z\right)\right] \\ -\left(X_0/2\pi^2 a\right)\exp\left(2\pi^2 a\left(X_0-Z\right)/X_0 Z\right)\end{array}\right\}dZ. \tag{6}$$

Here $\mathrm{Ei}_1\left(Y\right)=\int_1^\infty\left[\exp\left(-uY\right)/u\right]du$ (57) and interestingly $\lim_{L\to\infty}T_B\left(X\right)=T_N\left(X\right)$. Here $T_N\left(X\right)$ is the mean first passage time required by the DBD2s of TFs to reach L via pure 1D sliding in the absence of dynamic loop structures of DNA which is a solution of the following differential equation (7, 8, 21).

$$D_C d_X^2 T_N\left(X\right)=-1;\ T_N\left(L\right)=0;\ d_X T_N\left(X_0\right)=0;\ T_N\left(X\right)=\left(L^2-X^2\right)/2D_C-X_0\left(L-X\right)/D_C. \tag{7}$$

To obtain the target finding time, one needs to set $X = X_0$ in **Eqs. 6** and **7**. One can define the number of times the target finding rate of TFs can be accelerated by the looping mediated propulsion of TFs over 1D sliding as $\eta_P=\left[T_N\left(X_0\right)/T_B\left(X_0\right)\right]$ (here the subscript 'P' denotes the propulsion model) which is clearly independent of D_C of TFs complex and solely depends on (L, a, and X_0). Explicitly one can write it as,

$$\eta_P=\left(L-X_0\right)^2\left(\left(L^2-X_0^2\right)+4\pi^2 a\int_{X_0}^L\left\{\begin{array}{l}\exp\left(2\pi^2 a/Z\right)\left[\mathrm{Ei}_1\left(2\pi^2 a/X_0\right)-\mathrm{Ei}_1\left(2\pi^2 a/Z\right)\right] \\ -\left(X_0/2\pi^2 a\right)\exp\left(2\pi^2 a\left(X_0-Z\right)/X_0 Z\right)\end{array}\right\}dZ\right)^{-1}. \tag{8}$$

Detailed numerical analysis (**Section 2** of the **Supporting Material**) suggests that there exists a maximum of η_P at which $\partial\eta_P/\partial L=0$ with $L = L_{\mathrm{opt}}$ and clearly, we have $\lim_{L\to\infty}\eta_P=1$ (**Figs. 2A** and **B**). This is logical since when $L > L_{\mathrm{opt}}$ then $\eta_P\to 1$ and when $L < L_{\mathrm{opt}}$ then the stored energy is not completely utilized to propel the DBD2s of CRMs-TFs. Further, $\lim_{L\to X_0}\eta_P=0$ since its numerator part goes to zero much faster than the denominator part (**Figs. S2** and **S3**). The total time required by the combinatorial TFs to form a synaptosome complex via propulsion mechanism will be $\tau_P=\tau_S+T_B\left(X\right)$.

Predictions of the propulsion model

The persistence length of DNA under *in vitro* conditions is a ~ 150 bp and the radius of gyration for most of the eukaryotic TFs will be around ~ 5 bp. When the number of TFs involved in the combinatorial regulation is 3-5, then the maximum r_P ~ 15-25 bp. Therefore, one can set the initial $X=2\pi r_P$ ~100-150 bp where we have set $\delta = 2\pi$ here (58, 59). Simulations (**Fig. 2A**) of the expression for η_P (**Eq. 7**) at different values of X_0 and, L from X_0 to 10^5 suggested that L_{opt} ~ $3X_0$ (**Figs. 2C** and **2D**). When a ~150 bp and X_0 ~ 100-150 bp, then L_{opt} ~ 300-450 bp. Remarkably, this is the most probable range of the distances between the CRMs and promoters of various genes

observed across several genomes (60). The efficiency of the propulsion will be maximum at L_{opt}. Although L_{opt} is not much affected by a, the maximum of η_P is positively correlated with a. This is logical since the stored elastic energy is directly proportional to the persistence length of the polymer. Remarkably, at the optimum L_{opt} the speed of interactions between CRM-TFs complex with the promoters will be ~10-25 times faster than the normal 1D sliding.

Tethered sliding models

When the number of TFs involved in the combinatorial regulation is less such that $n_{IV} < n_{III}$, then the cumulative nonspecific electrostatic interactions will not be enough to bend the DNA around the TFs complex as in the case of propulsion model (**Fig. 3**). Therefore, stable initial nonspecific contacts between DBD2s and DNA will be formed slightly away from CRMs (S1) i.e. the radius of curvature of the initial looped-out DNA structure will be higher than the radius of gyration of the TFs complex r_P. Here the tethered random walker (DBD2s of CRMs-TFs complex, which is actually tied with the DNA thread at DBD1-S1) wanders over 3D space and randomly forms nonspecific contacts with other segments of the same DNA polymer analogous to the ring-closure events of intersegmental transfer dynamics. Before dissociation, DBD2s of CRMs-TFs may scan the DNA of random length for the presence of promoter P. In pure tethered sliding, the nonspecific electrostatic interactions between DBD2s and DNA are strong enough to keep the dynamic loop structure intact until reaching the promoter via sliding dynamics. In case of tethered binding-sliding-unbinding, the tied DBD2s perform multiple cycles of binding-sliding-unbinding before reaching the promoter.

We first consider the pure tethered sliding model. When the length of DNA that connects DBD1-S1 and the landing position of DBD2s is X for an arbitrary nonspecific contact of DBD2s, then the potential energy barrier acting on such random sliding will be $E \simeq \left(2\pi^2 a/X\right) + \left(3/2\right)\ln\left(\pi X/6b\right)$ where $b = 1$ bp is the distance between two consecutive nucleotides of DNA. This potential energy barrier attains a minimum value as $E_{\min} = [3/2]\left(1 + \ln\left(2\pi^3 a/9b\right)\right)$ at $X_C = 4\pi^2 a/3$. Forward and reverse movement of such *tethered random walker* drives X to $X + 1$ or $X - 1$. Contrasting from the propulsion model, here we have not ignored the entropy component of the potential function E since the initial interconnecting DNA segment is larger than the radius of gyration of the TFs complex and also in the free loop form. As a result, one cannot ignore the entropic barrier associated with the loop formation. Moreover, sliding of CRMs-TFs with dynamic loop structure will always be impeded by the chain entropy. The force generated by such potential will be $F\left(X\right) = 2\pi^2 a/X^2 - 3/2X$. Upon inserting this force term into **Eq. 5** one finally obtains the following integral solution.

$$T_U\left(X\right) = \left(1/D_C\sqrt{2\pi a}\right)\int_X^L Z^{3/2}\exp\left(2\pi^2 a/Z\right)\left(\mathrm{Erf}\left(\sqrt{2\pi^2 a/Z}\right) - \mathrm{Erf}\left(\sqrt{2\pi^2 a/X_0}\right)\right)dZ \tag{9}$$

Here $\mathrm{Erf}\left(v\right) = \left(2/\sqrt{\pi}\right)\int_0^v \exp\left(-z^2\right)dz$ is the error function integral (57), and $T_U\left(X\right)$ is the MFPT required by a pure tethered random walker to find its specific site located at L starting from X (this is the initial loop length) anywhere within (X_0, L) where X_0 is a reflecting boundary and L is an

absorbing boundary. Since the potential function has a minimum at X_C, one can consider the following two different limiting regimes.

$$\lim_{X \gg X_C} T_U(X) \simeq 4\left(L^{5/2} - X^{5/2}\right)/5D_C\sqrt{X_0} - \left(L^2 - X^2\right)/D_C\,;\ \lim_{X \ll X_C} T_U(X) \simeq T_B(X) \qquad (10)$$

One can define the number of times the target finding rate of TFs can be accelerated by the tethered sliding of TF as $\eta_S = T_N(X)/T_U(X)$ (here the subscript 'S' denotes the tethered sliding model) which is clearly independent of D_C of TF and solely depends on (L, a, and X_0). Contrasting from the propulsion model, one finds that $\lim_{L\to\infty} \eta_S = 0$. In these calculations we have not included the looping mediated nonspecific association time required by the DBD2s of the CRMs-TFs complex. This in fact further increases the overall MFPT of the tethered sliding model. The rate associated with the formation of the initial (nonspecific contact) loop with contour length X can be written as $k_{NL} \simeq k_t \exp(-E)$ where k_t (s^{-1}) is the maximum achievable rate under zero potential. Clearly, k_{NL} will be a maximum at X_C which is the most probable initial landing position of the tethered DBD2s of TFs via DNA-looping. In this model, the total time required by the CRMs-TFs to form the final synaptosome complex will be $\tau_{TS} = \tau_S + 1/k_{NL} + T_U(X)$ which attains a local minimum near $X = X_C$. One can also define $\eta_{NL} = k_{NL}/k_t$ which attains the maximum $\eta_{NL} \sim 6.7$ at X_C. Since there are several unbinding events in the tethered binding-sliding-unbinding model, the overall time required by the DBDs to form the synaptosome complex will be much higher than τ_{TS}.

Predictions of the tethered sliding models

Tethered sliding model predicts the most probable distance of the CRMs of TFs from the transcription start sites as X_C. At this distance, the rate of looping mediated synaptosome complex formation of TFs will be at maximum. Upon setting $X = X_C$ in η_S and numerically iterating L from 3000 to 10000 bp with $a \sim 150$ bp, one can observe the following results. When the left reflecting boundary was at X_0, then one finds the critical distance L_C such that $\eta_S > 1$ when $L < L_C$ and approximately $\eta_S < 1$ when $L > L_C$. Particularly, when $X_0 < 100$ bp then one can define the critical distance of transcription start sites from CRMs in the tethered sliding model as $L_C \sim 3X_C$. This critical distance decreases with increase in X_0. These numerical results are demonstrated in **Fig. S4** of the Supporting Materials.

COMPUTATIONAL ANALYSIS

Tethered sliding model predicted the most probable distance of CRMs from the transcription start site (promoter, location S1 on DNA) as $X_C = 4\pi^2 a/3 \sim 2000$ bp for $a \sim 150$ bp. To check the validity of this prediction, we analyzed the upstream 5000 bp sequences of various genes of human and mouse genome. We used the position weight matrices of various transcription factors of human and mouse available with the JASPAR database and scanned upstream sequences of all the genes of the respective genomes.

Datasets and analysis

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) (61, 62) of various TFs of mouse and human were obtained from the publicly available JASPAR database (63, 64). There were 21929 sequences from mouse genome and 28824 sequences from the human genome. Using the PWMs of various available TFs we generated the score table for various upstream sequences based on the following equation (61).

$$S_{v,i} = -\sum_{u=i+1}^{q}\left[\sum_{b=A}^{T} f_{b,u}\log_e\left(f_{b,u}/f_b\right)\right];\ b=\{A,C,G,T\} \tag{11}$$

In this equation $S_{v,i}$ is the score value of PWM at i^{th} position upstream of the transcription start site (TSS) 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 specific binding sites of TFs. Here f_b was calculated from the random sequences of the given genome available with the UCSC database. We constructed the distribution of the distances of S1 from the transcription start site. There is a strong positive correlation between the score value and the binding energy of TFs (61). Computing the distribution of the distances of S1 from the TSS will confirm the validity of the tethered sliding model.

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 weight matrix corresponding to a given p-value. In our calculations, we have set the p-value < 10^{-6} for defining the putative specific binding sites of TFs. 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 x 10^6 bps and fragmented it into 10^3 number of sequences with length of 5000 bps. Then we scanned each random sequence with the same PWM and obtained the number of putative CRMs (false positives). The probability of observing a CRM site by chance will be calculated as p_{NF} = number of false positives / 1000.

RESULTS AND DISCUSSION

Various possible modes of transcription activation via combinatorial binding of several TFs are summarized in **Fig. S5** of the Supporting Materials. The main limitation of the propulsion model is the requirement of huge energy input involved in the initial bending of DNA around the TFs complex of interest. This needs to be derived either in the form of ATP hydrolysis or in the form of binding energy derived from the combinatorial TFs. For example, bending of a linear DNA with size of 100-150 bp into loop requires the hydrolysis of at least 2-3 ATPs (using $E_{\text{bend}} = E_{\text{elastic}} + E_{\text{entropy}}$, here hydrolysis of 1 ATP molecule releases ~12 k_BT). Investment of such energy input is required by CRMs-TFs system to actively slide in a directional dependent manner towards the promoter. Instead, tethered sliding of TFs does not require such huge energy input since there is no restriction on the initial loop length. As a result, directional dependent movement of TFs is not possible in the tethered sliding model. However, the probability density function associated with the initial loop length will be dictated by the bending energy profile. Actually, E_{bend} will be a minimum at $X_C \simeq 4\pi^2 a/3$ where the average search time required to form the final synaptosome complex will be at minimum (52). When $X < X_C$ then $E_{\text{bend}} \propto X^{-1}$. When $X > X_C$ then one finds that $E_{\text{bend}} \propto \ln(X)$. When a ~ 150 bp and X_C ~ 2000 bp then the minimum of E_{bend} ~ 13 k_BT which requires the hydrolysis of at least 1 ATP. These results are depicted in **Fig. 4**.

The analysis results on the upstream sequences of various genes of human and mouse are shown in **Figs. 5**. The most probable location of S1 seems to be around ~2500 bp away from the promoter in both mouse and human genome. The distributions of the distances of CRMs from the respective TSS are shown in **Figs. 5A** and **B**. These results are in line with the tethered sliding model which predicted the critical distance of CRMs from the promoter to be around $X_C \sim 2000$ bp. Including various models presented here, one can consider the following four possible modes associated with the formation of synaptosome complex.

a) *Propulsion mechanism*. This requires huge free energy input in the initial loop formation with a possibility of directional dependent movement of TFs towards the promoter. Eventually this mechanism warrants several combinatorial TFs.

b) *Tethered sliding mechanism*. This required minimal free energy input in the formation of initial loop. Although the directional dependent movement of TFs is not possible here, the free energy barrier involved in the initial dynamic loop formation stage restricts the initial landing position of DBD2s of TFs close to the promoters.

c) *Tethered binding-unbinding mode*. This mechanism is similar to the tethered sliding mode with restrictions on the sliding dynamics. Here the searching for the promoters is achieved via repeated binding-unbinding of the tethered TFs. Directional dependent movement of TFs along DNA is not possible in this mode.

d) *Parallel searching of two DBDs of TFs*. Here two different types of DBDs of the protein-protein complex of combinatorial TFs (DBD1 and DBD2) search for their cognate sites on DNA (S1 and P respectively) independently through a combination of 1D and 3D diffusions. When these DBDs binds their cognate sites simultaneously, then the looping of the intervening DNA segment automatically occurs as a result. However, this mechanism works well only for the single TF based transcription activation such as Lac I system and it is almost improbable for the combinatorial binding of TFs in the gene regulation of eukaryotes. However, this mode can be a parallel (but slow) pathway of loop formation for the above said mechanisms.

Nonspecific binding and dissociation of TFs

In multiprotein mediated DNA looping, there is always a possibility for two different TFs interact with S1 and P respectively and the looping is mediated via protein-protein interactions among these TFs. In both propulsion and pure tethered sliding models, we have assumed that the nonspecifically bound DBD2s of CRMs-TFs do not dissociate until reaching the promoter. Nevertheless, earlier studies suggested that this assumption is valid only for the average sliding length of TFs $L_S \simeq \sqrt{2D_C/k_r}$ where k_r is the dissociation rate constant (8, 21). On can define this quantity as $k_r \simeq k_r^0 \exp\left(-\mu_{NS}\right)$ where $k_r^0 \sim 10^6$ s^{-1} is the protein folding rate limit (65) and μ_{NS} is the free energy barrier related to the dissociation of DBD2s of TFs from DNA. Here $-\mu_{NS}$ is approximately the nonspecific binding energy. Clearly $\mu_{NS} \geq 11\ k_BT$ is required to attain a sliding length $L_S \sim 300$ bp which can be achieved via combinatorial binding of several TFs. In these calculations we assumed the 1D diffusion coefficient as $D_C \sim 10^6$ bp^2 s^{-1} (29, 42).

Number of TFs required for the propulsion model

Earlier studies suggested that the DBDs of TFs behave like downhill folding proteins at their mid-point denaturation temperatures (21). Predominantly, DBDs exhibit at least two different conformational states viz. stationary and mobile (20, 21). Stationary conformational state is more sensitive to the sequence information of DNA but with less sliding mobility. Whereas, the mobile conformational state is less sensitive to the sequence information but with more sliding mobility. The average free energy barrier which separates these stationary and mobile states will be close to ~3 k_BT which is approximately the energy (μ_{NS}) corresponding to the nonspecific binding of single DBD with DNA i.e. the nonspecific binding energy per TF will be around -3 k_BT. When there are at least four different TFs (4 DBD2s, as demonstrated in **Fig. 1B**) involved in a combinatorial regulation, then their cumulative nonspecific binding energy will be around -12 k_BT which is sufficient to support the sliding of CRMs-TFs over ~300 bp without dissociation. In general, we have the scaling as $\mu_{NS} \sim 3n$ where n is the number of TFs involved in the combinatorial regulation. This further suggests the following relationship between n and L.

$$n \simeq [1/3]\ln\left(k_r^0 L^2 / 2D_C\right); \;\; \therefore n \propto \ln(L) \tag{12}$$

Here L is the distance between CRMs and promoter. Behavior of this functional relationship is demonstrated in **Fig. 6**. When the number TFs is less than four or distance of promoter from CRMs is higher than ~300 bp, then the CRMs-TFs complex will take either the pure tethered sliding or tethered binding-sliding-unbinding mode of dynamics to form the final synaptosome complex. Clearly, when n = 1 to 2 then the tethered binding-sliding-unbinding mode is the only choice for the looping mediated transcription activation.

Feasibility of the propulsion model

When the number of TFs (n) involved in the combinatorial regulation is ~3-5 and the average radius of gyration of individual TF is ~5 bp, then the maximum radius of gyration of the TFs complex will be around r_P ~15-25 bp. Our propulsion model will work only when the sequential recruitment of TFs at CRMs can bend the CRMs-DNA around them so that the radius of curvature is same as that of r_P. This in turn requires X_0 ~ 100-150 bp. When a ~ 150 bp, then the bending of a DNA segment of size X_0 into a circular loop requires $E_{\text{elastic}} \simeq 2\pi^2 a / X_0$ ~ 20-30 k_BT. The average energy associated with the site-specific binding of TFs with CRMs (DBD1-S1 interactions) seems to be around –(15-20) k_BT (66) and the energy associated with the cumulative nonspecific interactions of all the DBD2s of TFs (~3n as in **Eq. 12**) will be around – (10-15) k_BT. When the average energy associated with the site-specific DBD1-S1 interactions via hydrogen binding network is -15 k_BT, then one obtains the connection between n and X_0 as $X_0 \simeq 2\pi^2 a / (15 + 3n)$. The overall specific and cumulative nonspecific binding energies of 3-5 numbers of combinatorial TFs will be around – (25-35) k_BT. This amount of binding energy is more than enough not only to bend the DNA of size X_0 into a circular loop around the TFs complex but also to support the sliding of DBD2s over ~300 bp without dissociation from DNA. These calculations along with **Eq. 12** evidently substantiate the feasibility of the directional-dependent propulsion model.

Limitations and Justifications

Recently, Agrawal et. al. in Ref. (56) had shown (using a two-dimensional model and numerical simulations) that the equilibrium probability associated with the protein mediated looping of a

linear DNA fragment attained a minimum around $X \sim 0.7\ a$ for an end-to-end distance of ~50 bp. This minimum point seems to increase linearly with respect to the end-to-end distance of the DNA polymer. On the other hand, the equilibrium looping probability attained a maximum at $X \sim 5a$ which was approximately independent of the end-to-end distance (56). Since the bending energy profile described by **Eq. 1** decides the equilibrium probability density function associated with the loop formation, one can conclude that apart from the existence of the minimum in the bending energy profile at $X = X_C$, there should also be a maximum in E_{bend} at some $X = X_D$ as X tends towards zero. Therefore **Eq. 1** will be valid only when the length of the DNA polymer is such that $X > X_D$. Here one should note $X_C \sim 13a$ which corresponds to a DNA segment with end-to-end distance equal to the length of the DNA.

In our propulsion and tethered sliding models, we have simplified our calculations by assuming circular shape for the dynamic loop structure of DNA (**Figs. 1** and **3**). When the loop length is comparable with that of the persistence length i.e. $a \sim 150$ bp, then the DNA segment behaves like a stiff rod. Bending such an elastic rod by pulling their terminals generally results in a tear drop or half-lemniscate like loop rather than a circular one (67). Here one should note that the bending energy at a given point on the DNA polymer is inversely proportional to the radius of curvature of that location. In a circular loop, the radius of curvature is equal in all the points. However, in the tear-drop loop, the bending energy will be higher at the focus (where the radius of curvature is less) than the terminal regions of the DNA segment. In the propulsion model, initial shape of the DNA loop well within the CRMs-TFs complex is a circular one. This possess the maximum elastic stress to propel the DBD2s of the TFs complex. As the CRMs-TFs complex slides towards the promoter, most of the elastic stresses were already released out and the lemniscate type bulging of the DNA loop will not contribute much to the propulsion of DBD2s. Therefore, it will not change our main results much. In the tethered sliding models, the dynamic loop lengths of DNA are much higher than the persistence length. Under these conditions, DNA loops take random conformations owing to increase in the chain entropy term. Therefore, assuming a circular loop of DNA is a very good approximation for the tethered sliding models.

In our calculations, we have also assumed that the TFs complex is a rigid sphere on which DNA is wrapped around. The conformational changes which occurs on the surface of the combinatorial TFs are mostly related to the cooperative protein-protein interactions associated with the recruitment processes. Since the propulsion and the tethered sliding models assume a complete assembly of all the TFs at CRMs as the initial condition, one can ignore the effects of the conformational fluctuations related to the surface level protein-protein interactions and we can consider the TFs complex as a rigid sphere.

In general, biological systems can overcome the looping energy barrier via three possible ways viz. **a**) multiprotein binding (38) which could be the origin of the combinatorial TFs in the process of evolution, **b**) placing sequence mediated kinetic traps corresponding to DBD2 in between CRMs and promoters (18) and, **c**) placing nucleosomes all over the genomic DNA to decrease the E_{entropy} component. All these aspects are observed in the natural systems. In multiprotein binding, the free energies associated with the DNA-protein and protein-protein interactions among TFs will be utilized in a cooperative manner for the looping of DNA. Here DBD1 and DBD2 may come from different proteins. Vilar and Saiz (38) had shown that the looping of DNA would be possible even with small concentrations of TFs when the number TFs in a combination is sufficiently large.

Multiprotein binding eventually increases X_0 values. However, increasing X_0 will eventually decreases both the maximum possible acceleration of TFs search dynamics and the energy barrier associated with the DNA-looping. As a result, natural systems have optimized X_0 between these two-opposing factors for maximum efficiency via manipulating the number of TFs involved in the combinatorial regulation.

CONCLUSIONS

We have shown that DNA-loops can propel the complex of combinatorial transcription factors along DNA from their specific binding sites towards the promoters. The source of propulsion is the elastic energy stored on the bent DNA of the site-specifically bound DNA-protein complex. We considered directional-dependent propulsion, tethered sliding and tethered binding-sliding-unbinding models on the looping mediated transcription activation.

In directional-dependent propulsion model, the first arrived transcription factor at CRMs further recruits other TFs via protein-protein interactions. Such TFs complex has two different types of DNA binding domains viz. DBD1 which forms tight site-specific complex with CRMs of DNA via hydrogen bonding network and promoter specific DBD2s which form nonspecific electrostatic interactions around CRMs. When the overall energy associated with the specific as well as cumulative nonspecific interactions are sufficient enough, then the DNA segment around CRMs will be bent into a circle around the TFs complex. The number of TFs involved in the combinatorial regulation plays critical role here. Since the site-specific interactions at CRMs and the cumulative nonspecific interactions between DBD2s of TFs and DNA are strong enough to resist the dissociation, sliding of DBD2s well within the Onsager radius associated with the DBD2s-DNA interface towards the promoter is the only possible way to release the elastic stress of bent DNA. When the DBD2s reach the promoter without dissociation, then they form tight synaptosome complex which is necessary for the transcription activation.

When the number of TFs involved in the combinatorial regulation is not enough to bend the DNA in to a circle around the TFs complex, then one can consider tethered sliding or tethered binding-sliding-unbinding models. In the tethered sliding model, the specifically bound TFs complex at CRMs forms nonspecific contacts with DNA via dynamic loop structure and then slide along DNA without dissociation until finding the promoter. In case of tethered binding-sliding-unbinding model, the site-specifically bound TFs complex performs several cycles of nonspecific binding-sliding-unbinding before finding the promoter.

Actually, elastic and entropic energy barriers associated with the looping of DNA seem to shape up the distribution of distances between TF binding sites and promoters in the process of genome evolution. We argued that the commonly observed multiprotein binding in gene regulation might have been acquired over evolution to overcome the looping energy barrier. Presence of nucleosomes on the genomic DNA of eukaryotes might have been evolved to reduce the entropy barrier associated with the looping.

APPENDIX A

The energy component E_{entropy} that is required to compensate the chain entropy loss for a Gaussian chain can be computed as follows. Let us assume that the looping of DNA occurs when $\left|\overline{R}\right| \leq \xi$ where $\left|\overline{R}\right| \in \left(\xi, Ql_d\right)$ (in m) is the end-to-end distance vector, ξ is the minimum looping-distance (in m) and Ql_d is the maximum length of the DNA polymer. Here Q is the total number of nucleotide monomers present in the DNA polymer and $l_d \sim 3.4 \times 10^{-10}$ m = 1 bp is the distance between two consecutive nucleotides. The probability density function associated with the end-to-end vector $\overline{R}$ can be approximately written as $p\left(\overline{R}\right) \simeq \left(3/2\pi Ql_d^2\right)^{3/2} \exp\left(-3\overline{R}^2/2Ql_d^2\right) d\overline{R}$ (68, 69). The entropy loss upon looping of DNA is $\Delta S_{\text{loop}} \simeq \ln\left(P_l/P_\Omega\right)$ (in k_B units) where $P_l \simeq \int_0^{\xi} p\left(\overline{R}\right) d\overline{R}$ is the probability of finding the loops is and $P_\Omega = \int_0^{Ql_d} p\left(\overline{R}\right) d\overline{R}$ is the cumulative probability of finding all the possible configurations including loops. Explicitly one can write this down as,

$$\Delta S_{\text{loop}} \simeq 3\ln\left(\text{Erf}\left(\sqrt{3\xi^2/2Ql_d^2}\right)\Big/\text{Erf}\left(\sqrt{3Q/2}\right)\right) . \tag{A1}$$

Here $\text{Erf}\left(s\right) = \sqrt{4/\pi}\int_0^s \exp\left(-Y^2\right) dY$ is the error function (57). When $\xi \simeq l_d$ is very small then $\Delta S_{\text{loop}} \simeq 3\ln\left(\text{Erf}\left(\sqrt{3Q/2}\right)\sqrt{6/\pi Q}\right) \simeq -\left(3/2\right)\ln\left(\pi Q/6\right)$ for large Q (52). This expression for the entropy is closely linked with the Jacobson-Stockmayer factor, or *J*-factor associated with polymer looping (37). Since $Q = X/b$ where X is the length of the DNA polymer in bp and b = 1 bp is the distance between two consecutive nucleotide monomers of the DNA polymer, one finally obtains $E_{\text{bend}} \simeq 2\pi^2 a/X + \left(3/2\right)\ln\left(\pi X/6b\right)$.

REFERENCES

1. Alberts, B. 2002. Molecular biology of the cell. Garland Science, New York.
2. Ptashne, M. 1986. A genetic switch : gene control and phage [lambda]. Cell Press ; Blackwell Scientific Publications, Cambridge, Mass.; Palo Alto, Calif.
3. Ptashne, M., and A. Gann. 2002. Genes & signals. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
4. Atkins, P. W. 1978. Physical chemistry. W.H. Freeman, San Francisco.
5. Berg, O. G., R. B. Winter, and P. H. von Hippel. 1981. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. Biochemistry 20:6929-6948.
6. Winter, R. B., O. G. Berg, and P. H. von Hippel. 1981. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor--operator interaction: kinetic measurements and conclusions. Biochemistry 20:6961-6977.
7. Murugan, R. 2004. DNA-protein interactions under random jump conditions. Phys Rev E Stat Nonlin Soft Matter Phys 69:011911.
8. Murugan, R. 2007. Generalized theory of site-specific DNA-protein interactions. Phys Rev E Stat Nonlin Soft Matter Phys 76:011901.
9. Niranjani, G., and R. Murugan. 2016. Generalized theory on the mechanism of site-specific DNA–protein interactions. Journal of Statistical Mechanics: Theory and Experiment 2016:053501.
10. Koslover, E. F., M. A. Diaz de la Rosa, and A. J. Spakowitz. 2011. Theoretical and computational modeling of target-site search kinetics in vitro and in vivo. Biophys J 101:856-865.
11. Slutsky, M., and L. A. Mirny. 2004. Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. Biophys J 87:4021-4035.
12. Wunderlich, Z., and L. A. Mirny. 2008. Spatial effects on the speed and reliability of protein-DNA search. Nucleic Acids Res 36:3570-3578.
13. Shvets, A. A., and A. B. Kolomeisky. 2016. Crowding on DNA in Protein Search for Targets. J Phys Chem Lett 7:2502-2506.
14. Murugan, R. 2018. Theory of Site-Specific DNA-Protein Interactions in the Presence of Nucleosome Roadblocks. Biophysical Journal 114:2516-2529.
15. Kulic, I. M., and H. Schiessel. 2003. Nucleosome repositioning via loop formation. Biophys J 84:3197-3211.
16. Schiessel, H., J. Widom, R. F. Bruinsma, and W. M. Gelbart. 2001. Polymer reptation and nucleosome repositioning. Phys Rev Lett 86:4414-4417.
17. Amitai, A. 2018. Chromatin Configuration Affects the Dynamics and Distribution of a Transiently Interacting Protein. Biophys J.
18. Niranjani, G., and R. Murugan. 2016. Theory on the mechanism of site-specific DNA–protein interactions in the presence of traps. Physical Biology 13:046003.
19. Lange, M., M. Kochugaeva, and A. B. Kolomeisky. 2015. Dynamics of the Protein Search for Targets on DNA in the Presence of Traps. J Phys Chem B 119:12410-12416.
20. Kalodimos, C. G., N. Biris, A. M. Bonvin, M. M. Levandoski, M. Guennuegues, R. Boelens, and R. Kaptein. 2004. Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. Science 305:386-389.
21. Murugan, R. 2010. Theory of site-specific DNA-protein interactions in the presence of conformational fluctuations of DNA binding domains. Biophys J 99:353-359.
22. Zhou, H. X. 2011. Rapid search for specific sites on DNA through conformational switch of nonspecifically bound proteins. Proc Natl Acad Sci U S A 108:8651-8656.
23. Ando, T., and J. Skolnick. 2014. Sliding of proteins non-specifically bound to DNA: Brownian dynamics studies with coarse-grained protein and DNA models. PLoS Comput Biol 10:e1003990.

24. Murugan, R. 2009. Packaging effects on site-specific DNA-protein interactions. Phys Rev E Stat Nonlin Soft Matter Phys 79:061920.
25. Khazanov, N., A. Marcovitz, and Y. Levy. 2013. Asymmetric DNA-search dynamics by symmetric dimeric proteins. Biochemistry 52:5335-5344.
26. Marcovitz, A., and Y. Levy. 2013. Obstacles may facilitate and direct DNA search by proteins. Biophys J 104:2042-2050.
27. Marcovitz, A., and Y. Levy. 2013. Weak frustration regulates sliding and binding kinetics on rugged protein-DNA landscapes. J Phys Chem B 117:13005-13014.
28. Vuzman, D., Y. Hoffman, and Y. Levy. 2012. Modulating protein-DNA interactions by post-translational modifications at disordered regions. Pac Symp Biocomput:188-199.
29. Wang, Y. M., R. H. Austin, and E. C. Cox. 2006. Single molecule measurements of repressor protein 1D diffusion on DNA. Phys Rev Lett 97:048302.
30. Schleif, R. 1988. DNA looping. Science 240:127-128.
31. Schleif, R. 1992. DNA looping. Annu Rev Biochem 61:199-223.
32. Rippe, K., M. Guthold, P. H. von Hippel, and C. Bustamante. 1997. Transcriptional activation via DNA-looping: visualization of intermediates in the activation pathway of E. coli RNA polymerase x sigma 54 holoenzyme by scanning force microscopy. J Mol Biol 270:125-138.
33. Mulligan, P. J., Y. J. Chen, R. Phillips, and A. J. Spakowitz. 2015. Interplay of Protein Binding Interactions, DNA Mechanics, and Entropy in DNA Looping Kinetics. Biophys J 109:618-629.
34. Lewin, R. A., D. M. Crothers, D. L. Correll, and B. E. Reimann. 1964. A Phage Infecting Saprospira Grandis. Can J Microbiol 10:75-85.
35. Murugan, R. 2010. Theory on the mechanism of distal action of transcription factors: looping of DNA versus tracking along DNA. Journal of Physics A: Mathematical and Theoretical 43:415002.
36. Grindley, N. D., K. L. Whiteson, and P. A. Rice. 2006. Mechanisms of site-specific recombination. Annu Rev Biochem 75:567-605.
37. Zhang, Y., A. E. McEwen, D. M. Crothers, and S. D. Levene. 2006. Statistical-mechanical theory of DNA looping. Biophys J 90:1903-1912.
38. Vilar, J. M., and L. Saiz. 2006. Multiprotein DNA looping. Phys Rev Lett 96:238103.
39. Murugan, R. 2011. Theory on thermodynamic coupling of site-specific DNA–protein interactions with fluctuations in DNA-binding domains. Journal of Physics A: Mathematical and Theoretical 44:505002.
40. Mitarai, N., I. B. Dodd, M. T. Crooks, and K. Sneppen. 2008. The generation of promoter-mediated transcriptional noise in bacteria. PLoS Comput Biol 4:e1000109.
41. Murugan, R. 2011. Theory on the dynamic memory in the transcription-factor-mediated transcription activation. Phys Rev E Stat Nonlin Soft Matter Phys 83:041926.
42. Elf, J., G. W. Li, and X. S. Xie. 2007. Probing transcription factor dynamics at the single-molecule level in a living cell. Science 316:1191-1194.
43. Hammar, P., M. Wallden, D. Fange, F. Persson, O. Baltekin, G. Ullman, P. Leroy, and J. Elf. 2014. Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. Nat Genet 46:405-408.
44. Lewin, B., J. E. Krebs, S. T. Kilpatrick, E. S. Goldstein, and B. Lewin. 2011. Lewin's genes X. Jones and Bartlett, Sudbury, Mass.
45. Ravasi, T., H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, P. Carninci, C. O. Daub, A. R. Forrest, J. Gough, S. Grimmond, J. H. Han, T. Hashimoto, W. Hide, O. Hofmann, A. Kamburov, M. Kaur, H. Kawaji, A. Kubosaki, T. Lassmann, E. van Nimwegen, C. R. MacPherson, C. Ogawa, A. Radovanovic, A. Schwartz, R. D. Teasdale, J. Tegner, B. Lenhard, S. A. Teichmann, T. Arakawa, N. Ninomiya, K. Murakami, M. Tagami, S. Fukuda, K. Imamura, C. Kai, R. Ishihara, Y. Kitazume, J. Kawai, D. A. Hume, T. Ideker,

and Y. Hayashizaki. 2010. An atlas of combinatorial transcriptional regulation in mouse and man. Cell 140:744-752.
46. Murugan, R. 2010. Theory of site-specific interactions of the combinatorial transcription factors with DNA. Journal of Physics A: Mathematical and Theoretical 43:195003.
47. Halford, S. E. 2009. An end to 40 years of mistakes in DNA-protein association kinetics? Biochem Soc Trans 37:343-348.
48. Halford, S. E., and J. F. Marko. 2004. How do site-specific DNA-binding proteins find their targets? Nucleic Acids Res 32:3040-3052.
49. Murugan, R. 2009. Directional dependent dynamics of protein molecules on DNA. Phys Rev E Stat Nonlin Soft Matter Phys 79:041913.
50. Zhang, Y., and D. M. Crothers. 2003. Statistical mechanics of sequence-dependent circular DNA and its application for DNA cyclization. Biophys J 84:136-153.
51. Spirin, A. S. 2009. How does a scanning ribosomal particle move along the 5'-untranslated region of eukaryotic mRNA? Brownian Ratchet model. Biochemistry 48:10688-10692.
52. Shvets, A. A., and A. B. Kolomeisky. 2016. The Role of DNA Looping in the Search for Specific Targets on DNA by Multisite Proteins. J Phys Chem Lett 7:5022-5027.
53. Gardiner, C. W. 1985. Handbook of stochastic methods for physics, chemistry, and the natural sciences. Springer-Verlag, Berlin; New York.
54. Risken, H. 1989. The Fokker-Planck equation : methods of solution and applications. Springer-Verlag, Berlin; New York.
55. Kampen, N. G. v. 1981. Stochastic processes in physics and chemistry. North-Holland ; Sole distributors for the USA and Canada, Elsevier North-Holland, Amsterdam; New York; New York.
56. Agrawal, N. J., R. Radhakrishnan, and P. K. Purohit. 2008. Geometry of mediating protein affects the probability of loop formation in DNA. Biophys J 94:3150-3158.
57. Abramowitz, M., and I. A. Stegun. 1965. Handbook of mathematical functions, with formulas, graphs, and mathematical tables. Dover Publications, New York.
58. Wingender, E., P. Dietze, H. Karas, and R. Knuppel. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 24:238-241.
59. Khan, A., O. Fornes, A. Stigliani, M. Gheorghe, J. A. Castro-Mondragon, R. van der Lee, A. Bessy, J. Cheneby, S. R. Kulkarni, G. Tan, D. Baranasic, D. J. Arenillas, A. Sandelin, K. Vandepoele, B. Lenhard, B. Ballester, W. W. Wasserman, F. Parcy, and A. Mathelier. 2018. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res 46:D1284.
60. Koudritsky, M., and E. Domany. 2008. Positional distribution of human transcription factor binding sites. Nucleic Acids Res 36:6795-6805.
61. Stormo, G. D. 2000. DNA binding sites: representation and discovery. Bioinformatics 16:16-23.
62. Kreiman, G. 2004. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. Nucleic Acids Research 32:2889-2900.
63. Mathelier, A., O. Fornes, D. J. Arenillas, C. Y. Chen, G. Denay, J. Lee, W. Shi, C. Shyr, G. Tan, R. Worsley-Hunt, A. W. Zhang, F. Parcy, B. Lenhard, A. Sandelin, and W. W. Wasserman. 2016. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 44:D110-115.
64. Sandelin, A., W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32:D91-94.
65. Yang, W. Y., and M. Gruebele. 2003. Folding at the speed limit. Nature 423:193-197.
66. Barnes, S. L., N. M. Belliveau, W. T. Ireland, J. B. Kinney, and R. Phillips. 2019. Mapping DNA sequence to transcription factor binding energy in vivo. PLoS Comput Biol 15:e1006226.

67. Coleman, B. D., and E. H. Dill. 1992. Flexure waves in elastic rods. The Journal of the Acoustical Society of America 91:2663-2673.
68. Doi, M., and S. F. Edwards. 1988. The theory of polymer dynamics. Clarendon Press, Oxford [Oxfordshire].
69. Gennes, P.-G. d. 1979. Scaling concepts in polymer physics. Cornell University Press, Ithaca, N.Y.

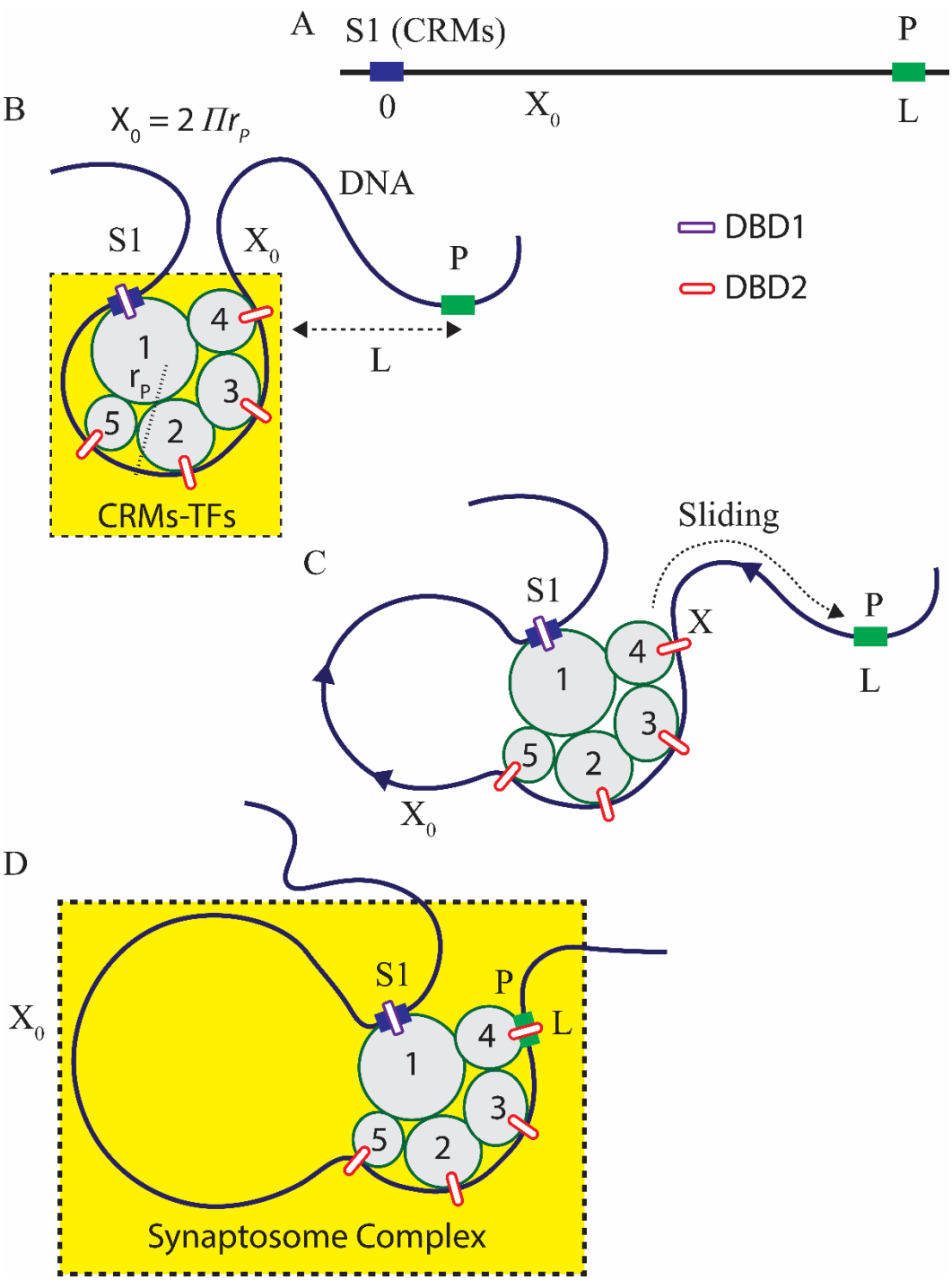

FIGURE 1. **Directional-dependent propulsion of combinatorial TFs in the looping mediated transcription activation**. **A**. We denote the location of CRMs-TFs complex on DNA as X. CRMs spans a length $X = (0, X_0)$ and the promoter is located at $X = L$. **B**. Looping mediated propulsion of combinatorial TFs complex with the overall radius of gyration of r_P bp along DNA. In this example, there are five different TFs (denoted as TF-1, 2, 3, 4, 5 respectively) involved in the combinatorial regulation. Here DBD1 type DBD of TF-1 specifically binds at S1 (CRMs) and

DBD2 of TF-4 specifically binds at the promoter. TFs 2, 3, and 5 involve in the stabilization of the CRMs-TFs complex via protein-protein interactions. TF-1 which arrives first at S1 will further recruit other four TFs in a cooperative manner. This complete CRMs-TFs complex is stabilized by the site-specific hydrogen bonding interactions between DBD1 of TF-1 and the CRMs located at S1 of DNA, protein-protein interactions among all the five TFs and the cumulative nonspecific electrostatic interactions between DBD2s of other TFs (2, 3, 4 and 5) with DNA. Binding of these TFs near CRMs (that spans for a length of X_0 from S1 ($X = 0$) to $X = X_0$) bends the DNA segment into a loop around them such that $X_0 = 2\pi r_P$. The elastic stress on DNA can be incrementally released via bulging of DNA around CRMs-TFs complex. **B**. Since the specific binding of DBD1 of TF-1 at S1 as well as the cumulative nonspecific interactions of all the TFs are strong enough to resist the dissociation, the elastic stress of DNA in the CRMs-TFs complex can be released only via thermally induced sliding of DBD2 of TFs towards the promoter well within the Onsager radius associated with the DBD2s-DNA interface. Upon reaching the promoter, DBD2 of TF-4 interacts with the promoter to form the final synaptosome complex. **D**. Synaptosome complex where the combinatorial TFs are bound with both CRMs and the promoter via DNA-loop.

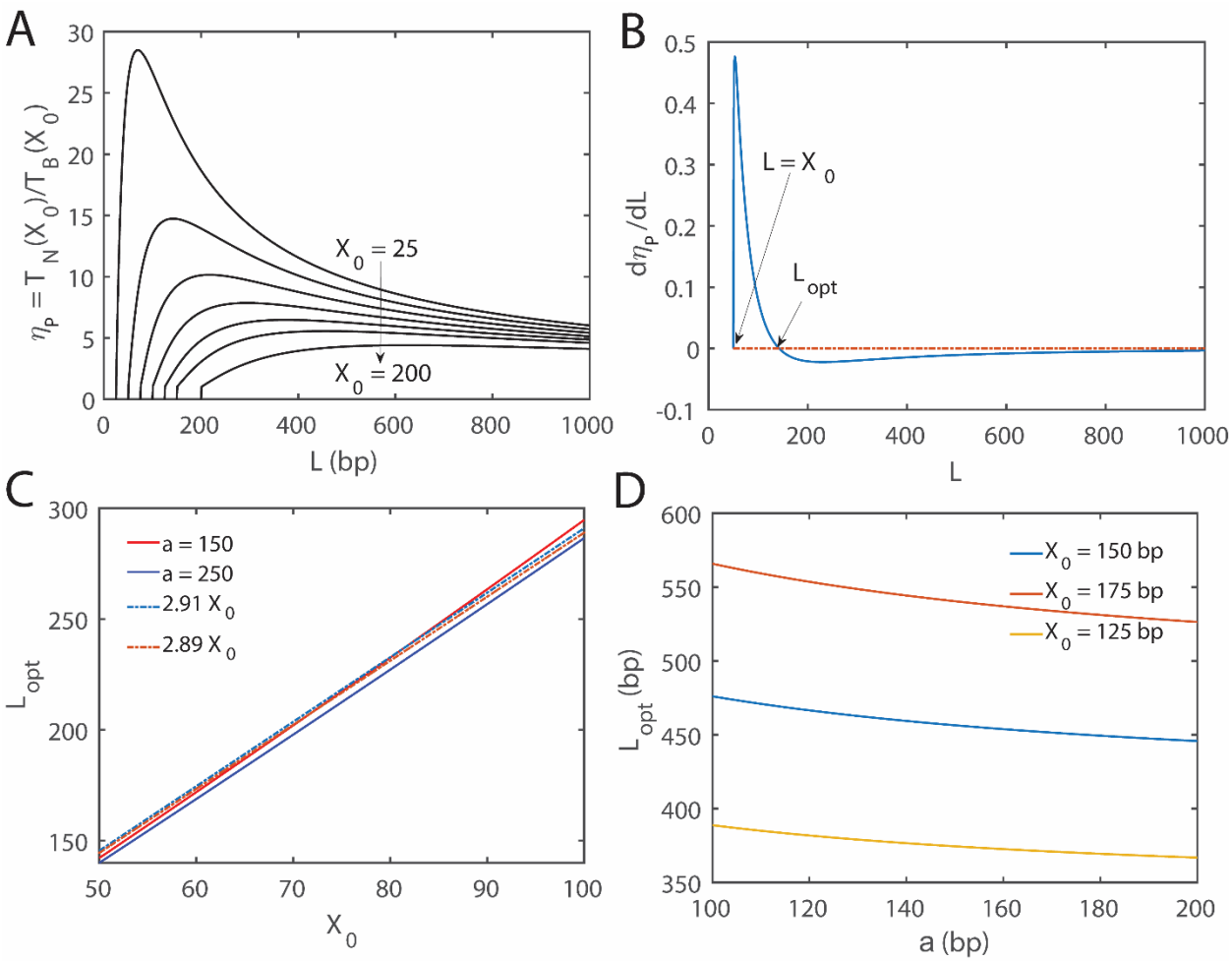

FIGURE 2. **Numerical analysis of the propulsion model**. **A.** Relative efficiency of looping mediated stochastic propulsion of TFs versus normal 1D sliding along DNA. $T_N(X_0)$ is the mean first passage time that is required by TFs to reach the promoter that is located at L, starting from X_0 via 1D sliding. $T_B(X_0)$ is the mean first passage time required by TFs to reach L starting from X_0 via looping mediated stochastic propulsion mechanism. X_0 was iterated as (25, 50, 75, 100, 125, 150, 200) along the arrow while iterating L from X_0 to 1000. The efficiency of looping mediated sliding is strongly dependent on the persistence length of DNA (a), L and X_0 and it is a maximum at $L_{opt} \sim 3X_0$. **B**. Plot of $d\eta_P/dL$ with respect to L. Here the settings are $a \sim 150$ bp and $X_0 \sim 50$ bp and L was iterated from 50 to 1000 bp. Upon solving $d\eta_P/dL = 0$ for L numerically one finds that $L_{opt} \sim 142.2$ bp. **C**. Variation of L_{opt} with respect to X_0. Clearly $L_{opt} \sim 3X_0$, is slightly dependent on the persistent length a. Here we have iterated X_0 from 50 to 100 bp and a = (150, 250) bp. The solution for L was searched within the interval (50, 1000) bp. **D**. Variation of L_{opt} with respect to changes in a. Here we have iterated a from 100 to 200 bp and X_0 = (125, 150, 200) bp. The solution for L was searched within the interval (50, 1000) bp. The error in the approximation $L_{opt} \sim 3X_0$ seems to be < 10% over wide range of a values.

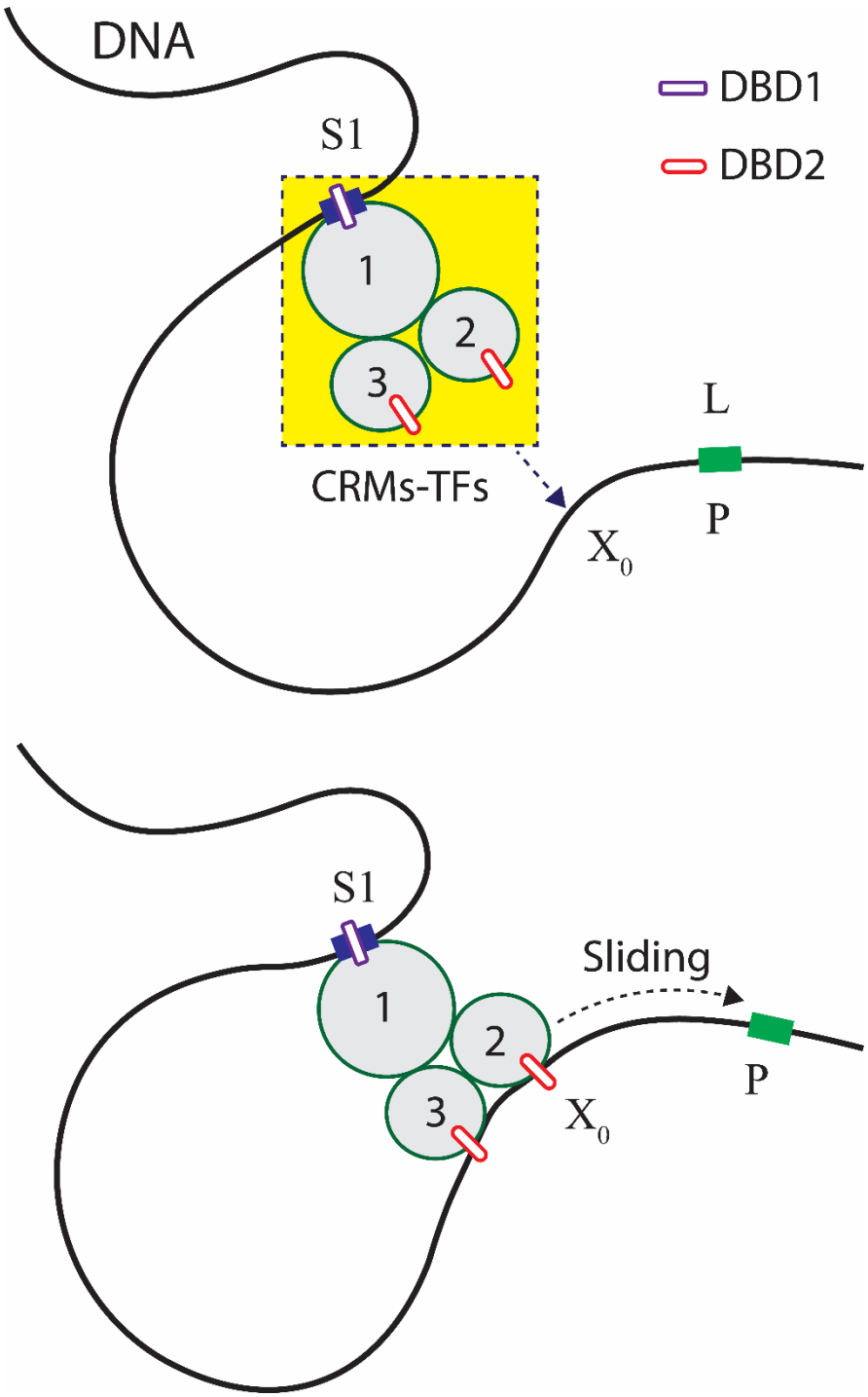

FIGURE 3. **Tethered sliding model**. In this example, there are three combinatorial TFs involved in the regulation viz. TF-1, 2 and 3. Binding of TF-1 at the CRM located at S1 subsequently recruits TF-2 and 3 near the location of S1 via protein-protein interactions. The DBD2s of TF-2 and 3 are specific to the promoter located at P whereas the DBD1 of TF-1 is specific to the CRMs located at S1. With intact hydrogen bonding interactions at S1, the CRMs-TFs complex wander over 3D space and forms nonspecific contacts with DNA randomly at X_0. Depending on the strength of the

cumulative nonspecific interactions of DBD2s, one can consider pure tethered sliding or tethered binding-sliding-unbinding. In pure tethered sliding, upon forming the first nonspecific contact with DNA at X_0 via dynamic loop structure, DBD2s of TFs search for the promoter via sliding without dissociation. In case of tethered binding-sliding-unbinding, nonspecifically bound DBD2s of TFs perform several cycles of nonspecific binding-sliding-dissociation before finding the promoter. When the DBD2s of TFs find the promoter that is located at P, then it forms the final synaptosome type complex.

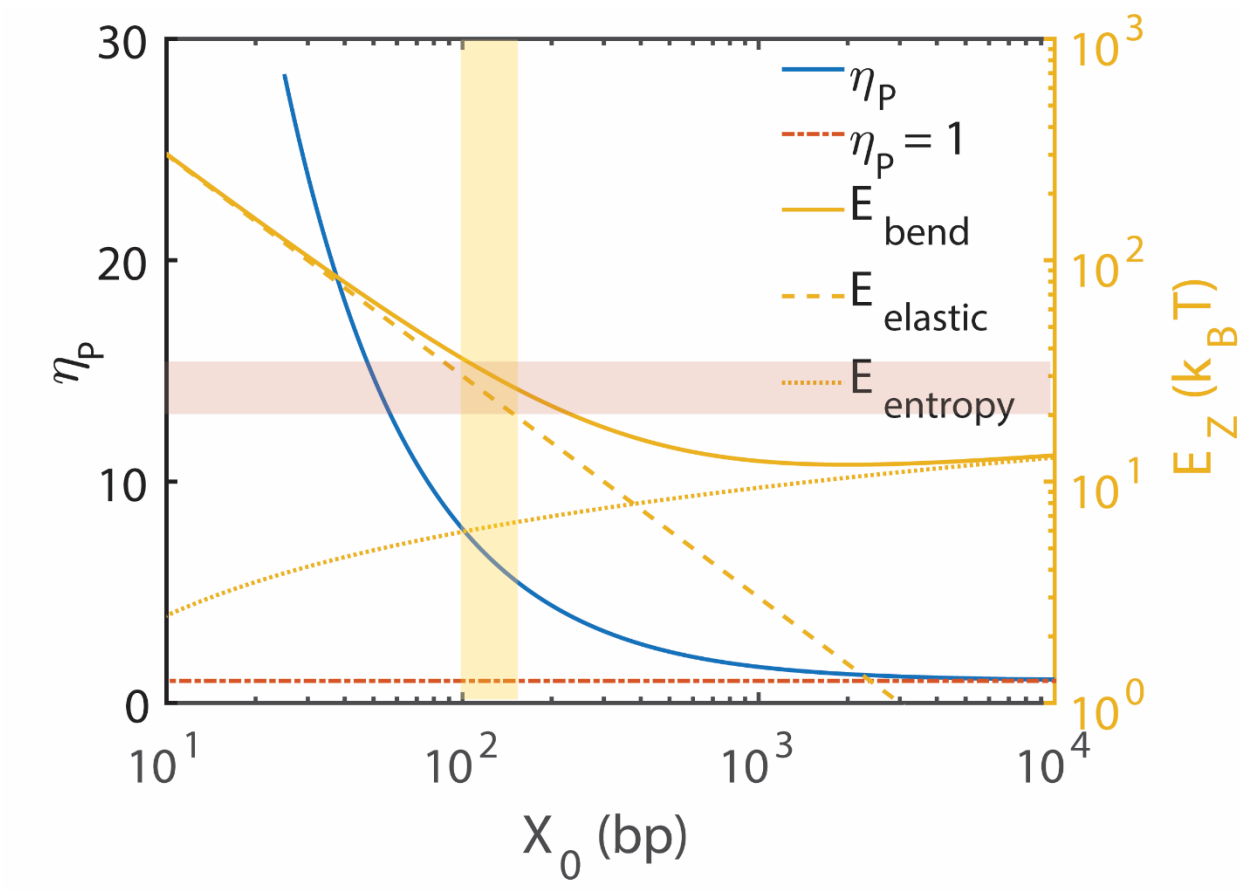

FIGURE 4. **Energy components of looping mediated transcription activation**. Variation of the propulsion efficiency η_P and bending energy with respect to changes in the initial loop length X_0. Here the settings are $a \sim 150$ bp, $L = 3X_0 \sim L_{opt}$. We computed $\eta_P = T_N(X_0)/T_B(X_0)$ with $L = 3X_0$ so that η_P will be close to its maximum. Here the subscript Z = (entropy, bend, elastic, enthalpy). $E_{bend} = E_{elastic} + E_{entropy}$ where $E_{elastic} \simeq 3000/X_0$ and $E_{entropy} \simeq 3\ln(\pi X_0/6b)/2$ which is ~12 k_BT at $X_0 \sim 2000$ bp. $E_{elastic} \leq 1\ k_BT$ when $X_0 \geq 3000$ bp. Clearly, the bending energy of the linear DNA is always $\geq$ 12 k_BT irrespective of the length. Shaded regions are the most probable X_0 values observed in the natural systems (~100-150 bp). The corresponding optimum distance between the transcription factor binding sites and promoters $L_{opt} \sim 3X_0 \sim$ 300-450 bp.

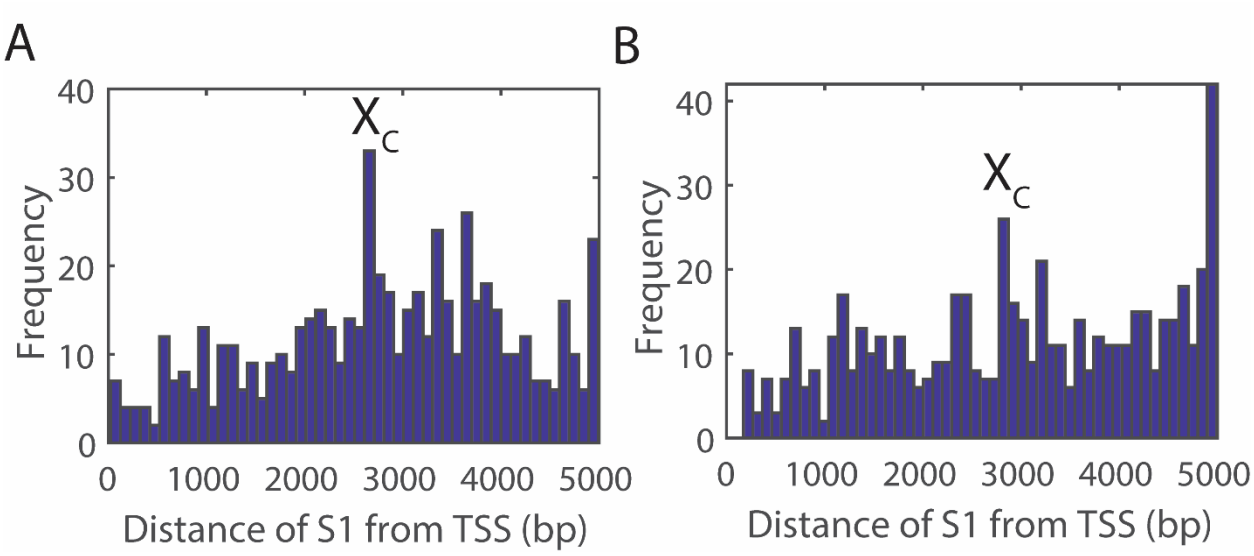

FIGURE 5. Computational data analysis results. We considered the position weight matrices of human and mouse available with the JASPAR database and scanned the upstream 5000 bp sequences of various human and mouse genes. The tethered sliding model predicted the most probable distance of the CRMs of TFs from the transcription start sites as $X_C \sim 2000$ bp for a persistence length of DNA as $a \sim 150$ bp. The distribution of the upstream location of CRMs of various TFs shows a maximum around ~2500 bp. Putative binding sites were defined with a *p*-value $< 10^{-6}$. These results are in line with the tethered sliding model. **A**. Mouse. **B**. Human.

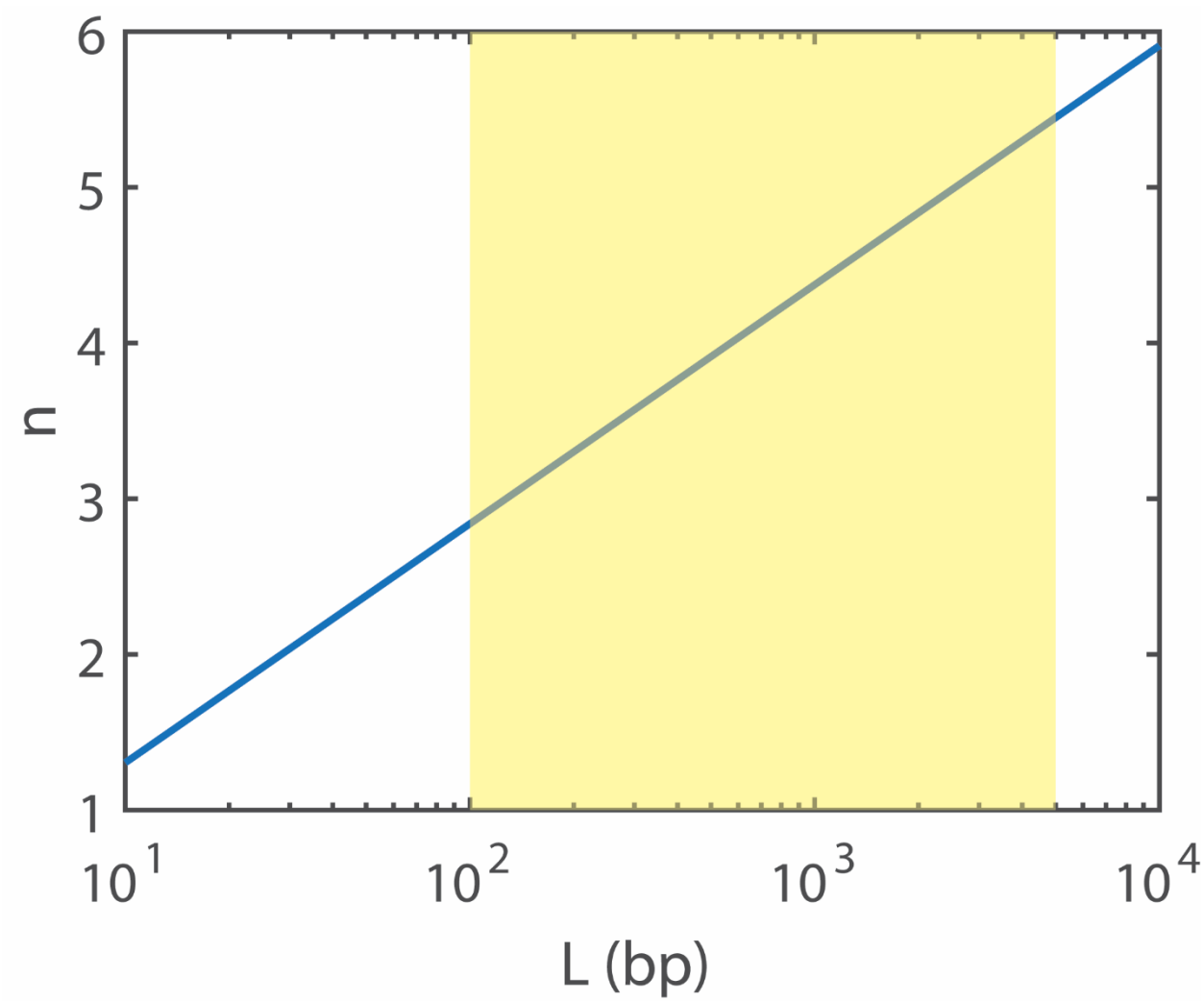

FIGURE 6. **Number of combinatorial TFs required to sustain the propulsion model**. Here L is the distance between CRMs and promoter and n is the number of combinatorial TFs required for the sliding of DBD2s of CRMs-TFs complex until reaching the promoter without dissociation. As in **Eq. 12**, $n \simeq [1/3]\ln\left(k_r^0 L^2 / 2D_C\right)$. For the purpose of calculations, we have used the value for the nonspecific dissociation rate constant of DBD2s at zero free energy barrier as $k_r^0 \simeq 10^6 s^{-1}$ and the 1D diffusion coefficient of DBD2s of TFs on DNA as $D_C \simeq 10^6 \text{bp}^2\text{s}^{-1}$. The yellow shaded region denotes the range of values of L occurring in the natural systems. These results suggest that ~3-5 TFs are enough to propel the DBD2s of TFs in most of the combinatorial regulations.

Theory on the looping mediated stochastic propulsion of transcription factors along DNA

R. Murugan

Department of Biotechnology, Indian Institute of Technology Madras, Chennai, India

Phone: +91 44 2257 4116. E-mail: rmurugan@gmail.com

Supporting Material

1. Theory of site-specific binding of TFs with CRMs

Let us consider a linear DNA of size N bps containing a *cis*-regulatory module (**CRM**) of a transcription factor (**TF**) of interest at an arbitrary location. The overall random search time or mean first passage time (**MFPT**) τ_U required by this TF to find its CRM via a combination of 3D and 1D diffusions can be written as follows (1, 2).

$$\tau_U \simeq \left[\left[P_{BTF}\right]\left(k_{fa} + k_{fn}/\left(1 + k_r \lambda \eta_U\right)\right)\right]^{-1};\ \lambda = N/U;\ \eta_U = U^2/6\left(\chi_{TF}/l_d^2\right). \tag{S1}$$

Here, $[P_{BTF}]$ (M, mols/lit) is the concentration of TF in cytoplasm, χ_{TF} ($m^2\ s^{-1}$) is the 1D diffusion coefficient of TFs on DNA, k_{fa} ($M^{-1}\ s^{-1}$) is the Smolochowski type bimolecular rate constant associated with the direct site-specific binding of TF with CRM via 3D diffusion, $k_{fn} \simeq k_{fa}\left(N/R_P\right)$ is the overall non-specific binding rate via 3D diffusion. In this, R_P is the radius of gyration of TF and k_r (s^{-1}) is the rate of dissociation of nonspecifically bound TF from DNA. Detailed theoretical studies suggested (1) an expression for k_{fa} as $k_{fa} \simeq k_t\left(p_R \delta/8\right)$. In this expression, $k_t \simeq 8k_B T/3\varphi$ ($M^{-1}\ s^{-1}$) is the maximum possible 3D diffusion-limited bimolecular collision rate where k_B is the Boltzmann constant, T is the absolute temperature and φ is the viscosity of the aqueous medium surrounding TFs. The numerical factor 1/8 accounts for the geometry of the random coiled and relaxed conformational state of DNA. Here (1 bps = l_d = 3.4 x 10^{-10} m). Under *in vitro* laboratory conditions (T = 298K and $\varphi \sim 10^{-1}$ kg $m^{-1}\ s^{-1}$ for aqueous buffer solution), one finds that $k_t \sim 10^8$ $M^{-1}s^{-1}$ (3). Further, p_R is the equilibrium probability of observing a specific or nonspecific binding site to be free from other dynamic roadblocks (1, 4) which are all present on the same DNA and $\delta \simeq |\kappa| \exp\left(1.18\omega_D \omega_P \sqrt{\Xi}\right)/R_S$ is the factor which accounts for the overall electrostatic attractive forces and the counteracting shielding effects of the solvent and other ions operating at the DNA-protein interface (1). Here $R_S \approx R_P + R_D$ is the corresponding reaction radius, R_D is the radius of the DNA cylinder and $|\kappa|$ is the ***Onsager radius*** *which is defined as the distance between charged reactant molecules at which the overall electrostatic energy will be the same as that of the background thermal energy*. Further ω_D and ω_P are the overall charges on the DNA backbone

and the DNA binding domains (**DBDs**) of TFs respectively and Ξ is the ionic strength of the surrounding reaction medium.

The term λ in **Eq. S1** is the minimum number of association-scan-dissociation cycles required by TFs to scan the entire DNA sequence of size N (bps) and η_U is the (averaged over initial landing positions) overall MFPT that is required to scan U bps of DNA via pure 1D diffusion. Here U (bps) is a random variable which takes different values in each of the association-scan-dissociation cycles. The probability density function associated with the 1D diffusion lengths U of TFs can be written as follows (1).

$$p(U) \simeq \left[2U \exp\left(-\left(U/U_A \right)^2 \right) \Big/ U_A^2 \right]; \; U_A = \sqrt{6\chi_{TF} / l_d^2 k_r} \; . \tag{S2}$$

Here U_A is the maximum achievable 1D diffusion length of the nonspecifically bound TFs on DNA that is measured in bp. When TF moves with a jump size of k (bps) then we find that $\chi_{TF} = l_d^2 \sum_{i=-k}^{k} \left(p_{TF,i} w_{TF,i} i^2 \right)$ ($\mathrm{m^2\ s^{-1}}$) where $i \in Z$, $w_{TF,\pm i}$ are the microscopic transition rates ($1/s$) associated with the forward and reverse movements of TFs with jump size of i, and $p_{TF,\pm i}$ are the corresponding microscopic transition probabilities (5, 6). Here the dynamics of TF with jump size k means that from the current location x (measured in m) TF can hop anywhere inside ($x - k\, l_d$, $x + k\, l_d$) with equal probabilities i.e. $1/2k$. The step length k is measured in bp. Since the dynamics at the DNA-protein interface involves segmental motion of DBDs of TFs, one can assume the protein folding rate limit (7) for $w_{TF,\pm 1} \sim 10^6\ \mathrm{s^{-1}}$. Noting that $p_{TF,\pm 1} \sim ½$ for an unbiased 1D random walk, one finds that $\chi_{TF}/l_d^2 \sim 10^6\ \mathrm{bps^2 s^{-1}}$ corresponding to the sliding of TFs on DNA for which $k = 1$ bps. Approximately this is the experimental value of the 1D diffusion coefficient associated with the sliding of TFs on DNA (1, 8, 9). Using the probability density function corresponding to the sliding length, one can define the overall average search time as $\tau_S \simeq \int_0^L \tau_U p(U)\, dU$.

For an arbitrary jump size k with the average microscopic transition probabilities as $p_{TF,\pm i} = 1/2k$ and average transition rates $\langle w_{TF,i} \rangle = \phi_{TF}$, one finds that $\chi_{TF} = l_d^2 \phi_{TF} \left[(k+1)(2k+1)/6 \right]$. Recently, Amitai has shown (10) that $p_{TF,\pm i} \propto i^{-1/2}$ and $w_{TF,\pm i} \propto p_{TF,\pm i}$ (**Eqs. 2, 4-5** of Ref. (10) where we have $|n\text{-}l| = i$ in the present context) so that $\left(p_{TF,\pm i} w_{TF,\pm i} \right) \propto i^{-1}$ for large size chromatin DNA. Upon implementing this scaling, one obtains that $\chi_{TF} \propto \left[k(k+1)/2 \right]$. Therefore, our formula for χ_{TF} with respect to k is asymptotically consistent with Ref. (10). Here the jump size k is positively correlated with the degree of condensation of DNA. Densely packed DNA allows larger jump sizes. The 3D diffusion coefficient of TFs, i.e. χ_{3D} will be always ~10-10^2 times higher than χ_{TF} and in general $\chi_{TF} \leq \chi_{3D}$ irrespective of k and presence of various 1D facilitating processes such as hopping and intersegmental transfers. Various symbols and parameters used throughout this supporting material as well as in the main text are listed in **Table S1**.

Upon mixing DNA with TF in buffer, the site-specific complex can be formed via two different pathways viz. (**a**) direct site-specific binding through 3D diffusion and (**b**) through a combination

of 1D and 3D diffusions. One can describe this by a pseudo first order scheme $DNA \xrightarrow{1/\tau_U} DNAP$. Here the concentration of specific site will be equal to the concentration of DNA, and DNAP is the site-specific complex. In τ_U of **Eq. S1**, $[P_{BTF}]k_{fa}$ is the pseudo first order rate of direct site-specific binding of TF with DNA via pure 3D diffusion and $k_{fn}[P_{BTF}]/(1+k_r\lambda\eta_U)$ is the pseudo first order rate of site-specific binding of TF via a combination of 1D and 3D diffusion.

2. Looping mediated communication between CRMs-TFs with the promoter

The mean first passage time (MFPT) associated with the DNA binding domains 2 (DBD2) of CRMs-TFs complex to reach the promoter of a gene via *DNA-loop mediated propulsion mechanism* while the DBD1 of TF complex is still tightly bound with S1 of DNA (**Fig.1** of the main text and **Fig. S1** for details) can be given as follows.

$$T_B(X_0) = \frac{L^2 - X_0^2}{2D_C} + \frac{G}{2D_C} \tag{S3}$$

Here X_0 is the initial position of DBD2s of CRMs-TFs on DNA or the initial loop length, L is the location of promoter, a is the persistence length of DNA and D_C is the one-dimensional diffusion coefficient associated with the sliding of DBD2s of TFs along DNA. We have assumed here that the site-specific DBD1-S1 is strong and intact (**Fig. 1B** of the main text). The function G can be defined as follows.

$$G = 4\pi^2 a \int_{X_0}^{L} \left(\exp\left(\frac{2\pi^2 a}{Z}\right)\left[\mathrm{Ei}_1\left(\frac{2\pi^2 a}{X_0}\right) - \mathrm{Ei}_1\left(\frac{2\pi^2 a}{Z}\right)\right] - \left(\frac{X_0}{2\pi^2 a}\right)\exp\left(\frac{2\pi^2 a(X_0 - Z)}{X_0 Z}\right)\right) dZ \tag{S4}$$

Here $\mathrm{Ei}_1(Y) = \int_1^\infty [\exp(-sY)/s]\,ds$ is the E_1 exponential integral (11).

Noting that $T_N(X_0) = (L - X_0)^2/2D_C$ (from **Eq. 7** of the main text) which is the MFPT associated with the finding of promoter by TFs complex starting from X_0 via pure sliding dynamics without dynamics DNA loop structures, one can define η_P which is the number of times the dynamics loop driven sliding of CRMs-TFs towards the promoter is faster than the normal 1D sliding dynamics as follows.

$$\eta_P = \frac{T_N(X_0)}{T_B(X_0)} = \frac{(L - X_0)^2}{L^2 - X_0^2 + G} = \frac{1}{(L + X_0)/(L - X_0) + G/(L - X_0)^2} \ . \tag{S5}$$

Clearly η_P is not dependent on D_C and it depends only on the parameters (L, X_0 and a). Further, $\lim_{L\to X_0} \eta_P = 0$ since $T_N(X_0)$ approaches zero much faster than $T_B(X_0)$ as L tends towards infinity (see **Fig. S2** for details). There also exists an asymptotic limit as $\lim_{L\to\infty} \eta_P = 1$. This means that $\lim_{L\to\infty}\left[G/(L - X_0)^2\right] = 0$ (see **Fig. S3** for details). The optimum distance between CRMs

and promoter i.e. L_{opt} at which η_P is a maximum can be obtained by solving $d\eta_P/dL = 0$ for L for given a and X_0. Explicitly one can write down this as follows.

$$\frac{d\eta_P}{dL} = \frac{2\left(L - X_0\right)\left[-\frac{\left(L - X_0\right)}{2}\frac{\partial G}{\partial L} + G + X_0\left(L - X_0\right)\right]}{\left(G + L^2 - X_0^2\right)^2} = 0. \tag{S6}$$

This has a trivial solution $L = X_0$. Upon ignoring this one, L_{opt} can be obtained by numerically solving the following equation for L at given a and X_0.

$$L > X_0; \;\; \therefore \left(L - X_0\right)\frac{\partial G}{\partial L} - 2G - 2X_0\left(L - X_0\right) = 0. \tag{S7}$$

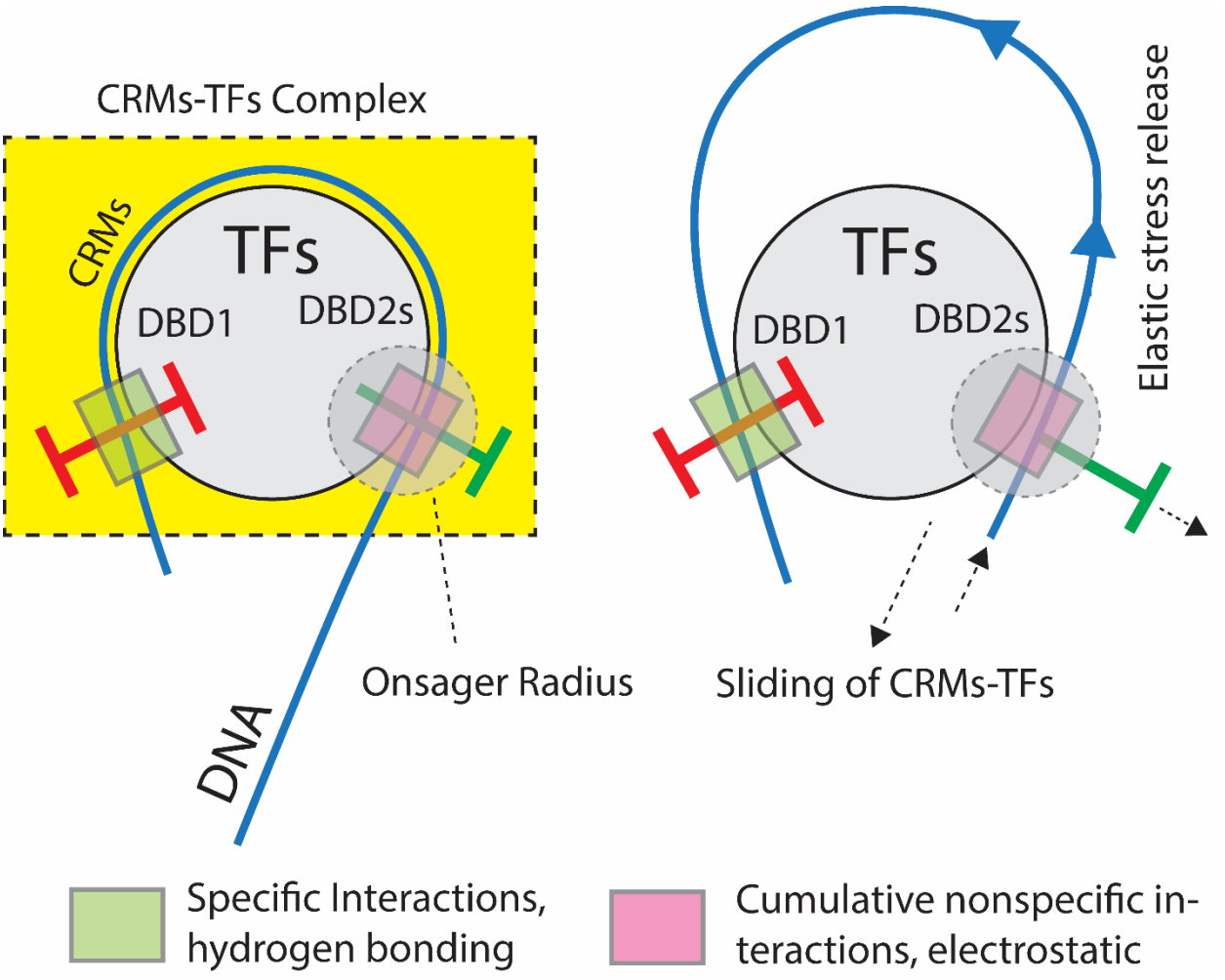

FIG S1. Schematic diagram on the mechanism of propulsion of CRMs-TFs on DNA. In the propulsion model, Combinatorial TFs assemble on the CRMs via recruitment mechanism and bend the DNA around final TFs complex. This TFs complex has two different DBDs viz. DBD1 which site-specifically interacts with the CRMs via hydrogen bonding network and several DBD2s which interact with the DNA near CRMs via nonspecific electrostatic interactions. Both DBD1-CRMs as well as DBD2s-DNA interactions are strong enough to keep the bent DNA intact. In this scenario, the elastic stress of bent DNA of CRMs-TFs complex can be released only via sliding of DBD2s of TFs along DNA well within the Onsager radius associated with the DBD2s-DNA interface. Propulsion mechanism works only when the cumulative nonspecific interactions are strong enough to keep the DBD2-DNA well within the Onsager radius until CRMs-TFs reaches the promoter. This in turn requires the binding of several combinatorial TFs.

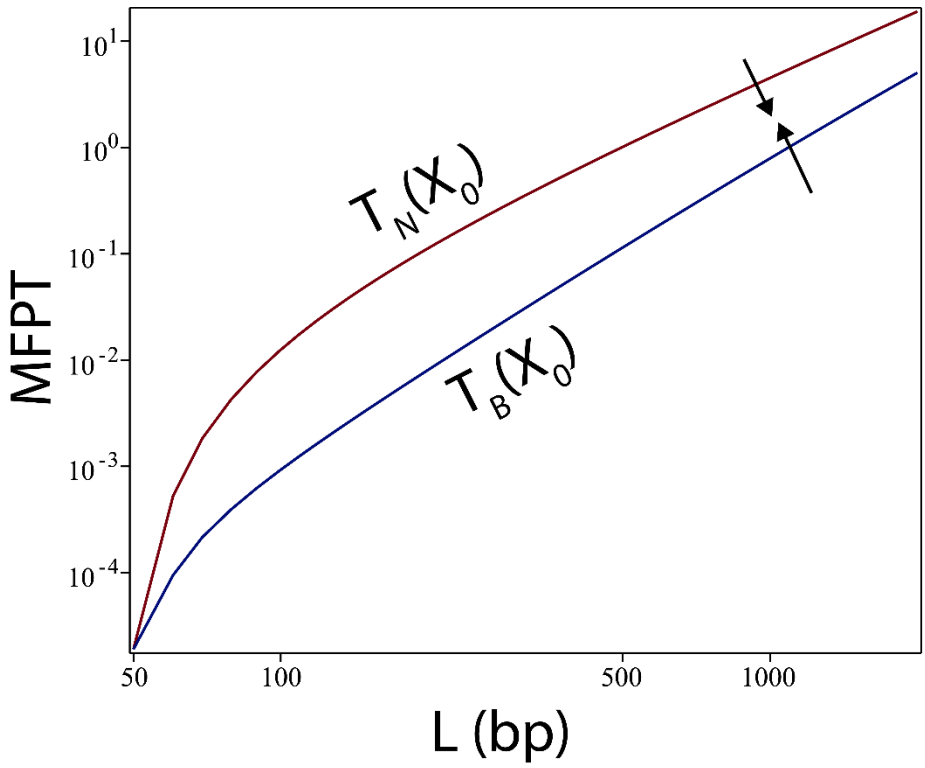

FIG S2. Showing that $T_N(X_0)$ approaches zero much faster than $T_B(X_0)$ so that $\lim_{L\to X_0}\eta_P = 0$. Here the settings are $X_0 \sim 50$ bp and $a \sim 150$ bp and L was iterated from 50 to 2000 bp. Further we also find the limit $\lim_{L\to\infty} T_B(X_0) = T_N(X_0)$ or $\lim_{L\to\infty}\left[T_N(X_0)/T_B(X_0)\right] = 1$.

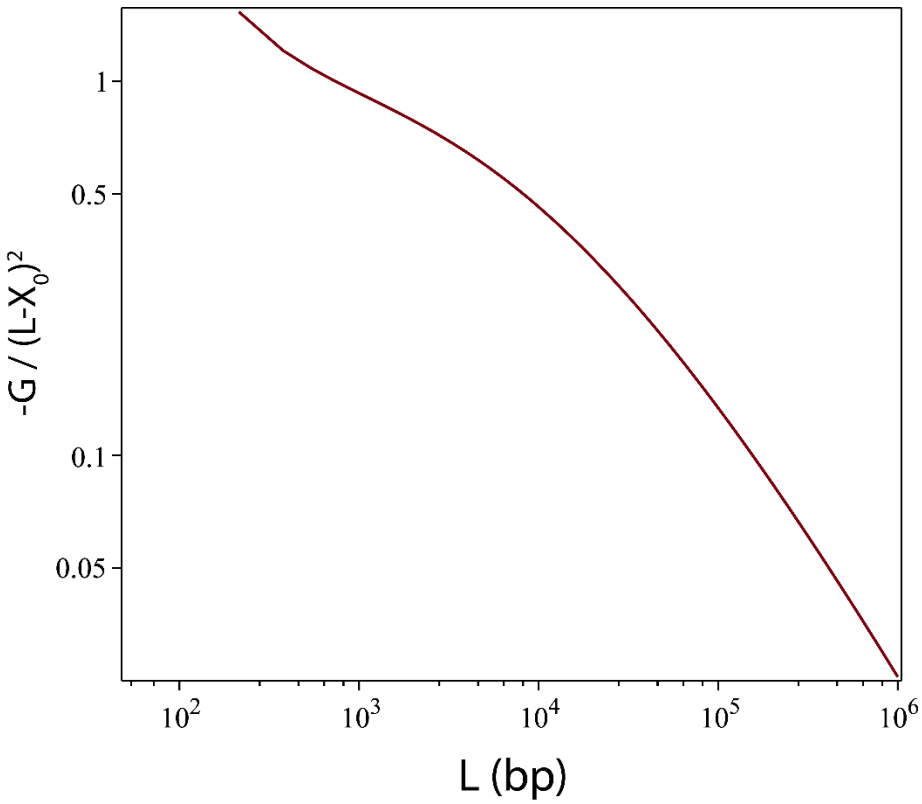

FIG S3. Showing $\lim_{L\to\infty}\left[G/\left(L-X_0\right)^2\right]\to 0$. Here the settings are $X_0 \sim 50$ bp and $a \sim 150$ bp and L was iterated from 50 to 10^6. The function G is defined as in **Eq. S4**.

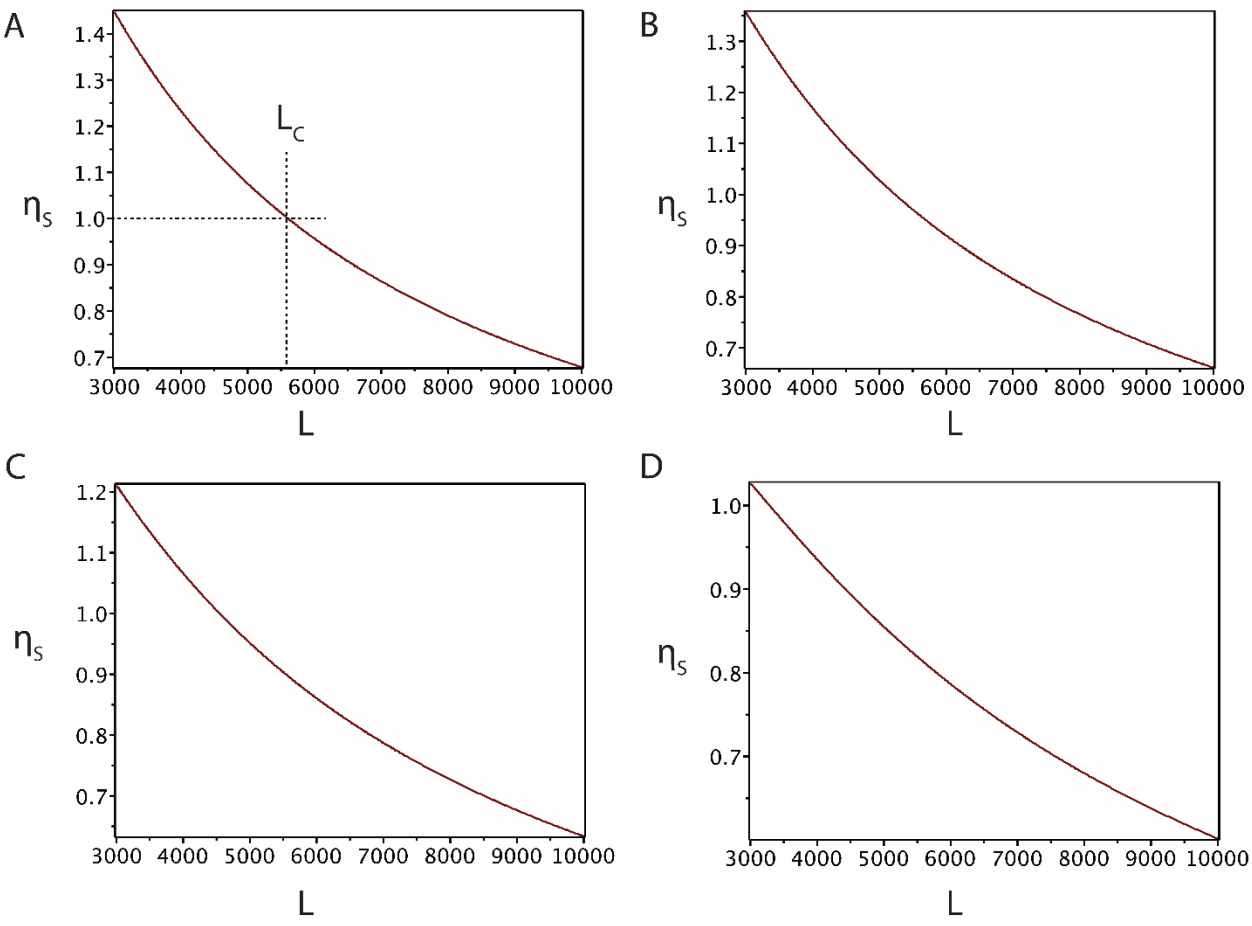

FIG S4. Variation of the critical distance (L_C) between CRMs and promoter in the tethered sliding model. Here $\eta_S = T_N(X)/T_U(X)$ as defined in **Eqs. 7** and **9**. In these MFPTs, X is the initial loop length, X_0 is the left side reflecting boundary and L is the right-side absorbing boundary. L_C is defined such that when $L < L_C$ then $\eta_S > 1$ and when $L > L_C$ then $\eta_S < 1$. Settings are (see **Eq. 9** of the main text), $X = X_C$ where $X_C = 4\pi^2 a/3$, persistence length of DNA $a \sim 150$ bp and L was iterated from 3000 to 10000 bp. **A**. $X_0 = 100$ bp. **B**. $X_0 = 250$ bp. **C**. $X_0 = 500$ bp. **D**. $X_0 = 1000$ bp. These results suggest that L_C decreases as X_0 increases. For the most probable value of $X_0 \sim 100$ bp one finds that $L_C \sim 3X_C$.

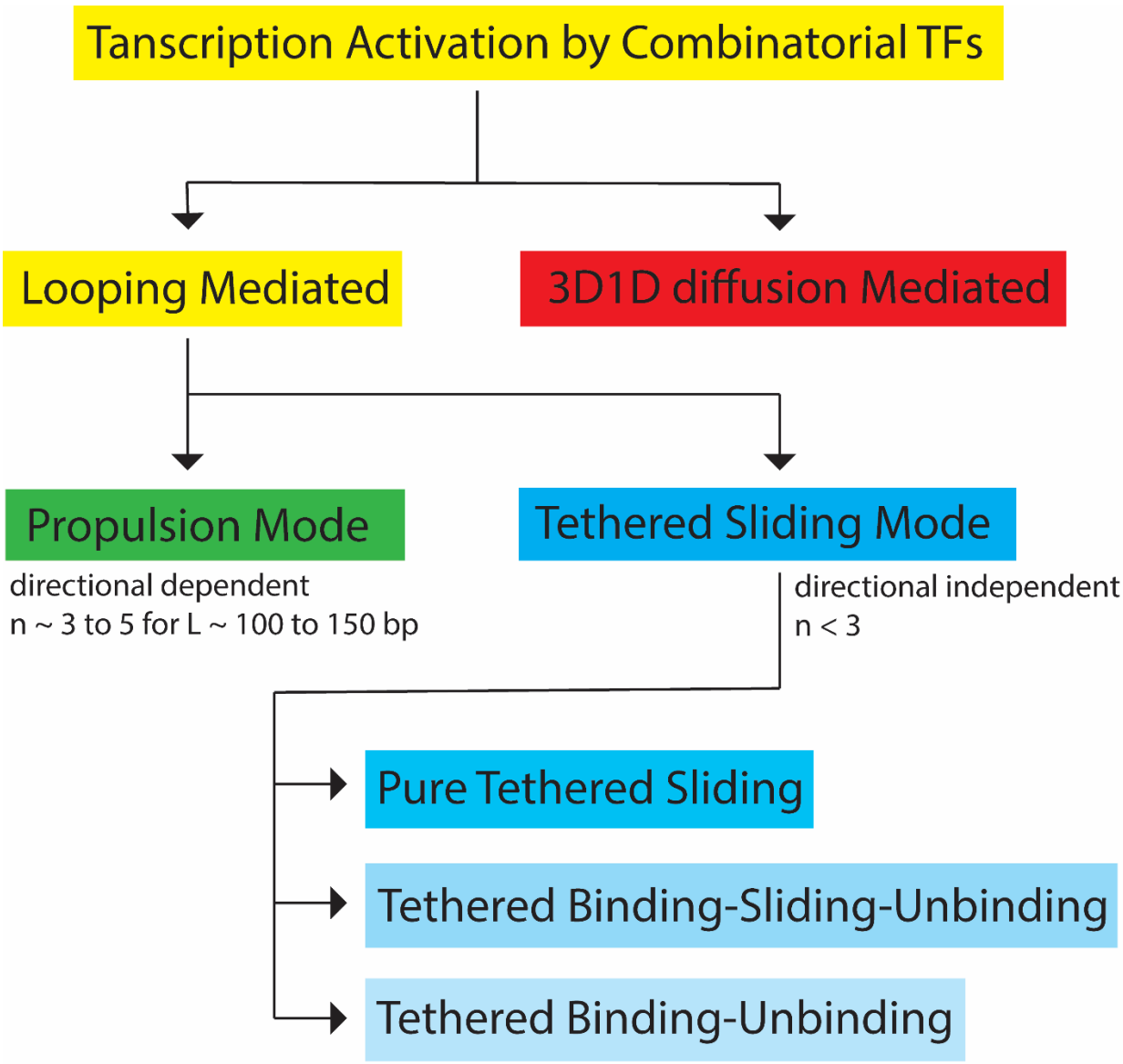

FIG S5. Summary of various models proposed on the transcription activation of combinatorial TFs. Looping mediated activation can be via the directional dependent propulsion mode or tethered sliding mode depending on the number of combinatorial TFs. Here n is the number TFs involved in the combinatorial regulation and L is the distance between CRM and promoter. Propulsion model will work only when the overall energy associated with the specific and nonspecific binding of combinatorial TFs is enough to bend the CRMs-DNA around the TFs complex and also sustain the continuous sliding of CRMs-TFs complex until reaching the promoter region. Tethered binding-unbinding model has been studied in detail in Ref. (12).

TABLE S1. List of symbols used in the main text

Symbol	Definition	Remarks
TF	Transcription factor	
CRM	*Cis*-regulatory module, specific sequence of DNA where TFs bind and then distally act on the promoters to initiate transcription.	
CRMs-TFs complex	Represents the entire complex of TFs bound with CRMs.	
Synaptosome complex	The final CRMs-TFs complex coupled with the promoter via DNA loop structure.	
DBD DBD1 DBD2	DNA Binding Domain of TF proteins; DBD1 of the TFs complex specifically binds with CRMs; DBD2s of the TFs complex specifically interacts with the promoter of the gene that needs to be activated.	
TSS	Transcription Start Site.	
MFPT	Mean First Passage Time.	s
1D, 3D	One-dimensional, three-dimensional.	
l_d	~3.4 x 10^{-10} m, equals to 1 base-pair (bp). We measure all the length parameters in bp.	bp
r_P	Radius of gyration of the TFs complex.	bp
X	Length of the DNA-loop at a given time point.	bp
L	Position of the promoter.	bp
X_0	Initial DNA-loop length. It is assumed such that $X_0 \sim 2\pi r_P$ where r_P is the radius of gyration of the TF complex.	bp
S1	DNA binding sequence (CRM) corresponding to DBD1, which is a specific binding site.	
P	Promoter sequence. This is the specific site corresponding to DBD2 by definition.	bp
a	Persistence length DNA. Typically, it is ~150 bp.	bp
E	Energy potential associated with the CRMs-TFs complex with bent DNA segment. $E \sim \lvert E_{\text{bond}} \rvert + E_{\text{elastic}}$. This is the energy required to dissociate the site-specific CRMs-TFs complex.	k_BT
n	Number of combinatorial TFs involved in the transcription activation.	dimensionless
b	Distance between two consecutive nucleotide monomers of the DNA polymer (this is equal to 1 bp).	bp
$F(X)$	It is force generated by the potential E in the propulsion model. $F(X) = -dE/dX = 2\pi^2 a/X^2$. This in turn can propel CRMs-TFs complex towards the promoter. In case of tethered sliding model $F(X) = 2\pi^2 a/X^2 - 3/2X$.	k_BT / bp
E_{bind}	$\sim E_{\text{bond}} + E_{\text{elastic}} + E_{\text{entropy}}$, the total site-specific binding energy associated with the DNA-TF complex dissipates into these three components.	k_BT $E_{\text{bind}} \leq 0$

E_{entropy}	Energy required to compensate the chain entropy loss of linear DNA that is converted into a loop upon binding of TF, $E_{\text{entropy}} \simeq (3/2)\ln(\pi X/6b)$ where X is the size of DNA that is converted into a loop upon binding of TF.	$k_B T$
E_{elastic}	Elastic energy involved in the bending of a linear segment of DNA $E_{\text{elastic}} \simeq aX/2r_P^2$, where a is the persistence length of DNA, X is the length of DNA segment and r_P is the radius of curvature after bending (this is equal to the radius of gyration of the TF complex). Elastic energy required to convert a DNA segment of size X into a loop such that $X = 2\pi r_P$ will be $E_{\text{elastic}} \simeq 2\pi^2 a/X$.	$k_B T$
E_{bond}	~$E_{\text{electrostatic}} + E_{\text{hydrogen}} + E_{\text{protein-protein}}$ is the overall bonding energy associated with the formation of CRMs-TFs complex. $E_{\text{electrostatic}}$ is associated with the cumulative nonspecific DBD2s-DNA interactions. E_{hydrogen} is component corresponding to the site-specific DBD1-S1 interactions via hydrogen bonding network. $E_{\text{protein-protein}}$ corresponds to the protein-protein interactions among the combinatorial TFs.	$k_B T$ $E_{\text{bond}} \leq 0$
E_{bend}	~ $E_{\text{elastic}} + E_{\text{entropy}}$, energy required to bend a linear DNA segment into a loop. $E_{\text{bend}} \simeq 2\pi^2 a/X + (3/2)\ln(\pi X/6b)$. This has a minimum with respect to X as $E_{\text{bend}}^{\min} = [3/2]\left(1+\ln\left(2\pi^3 a/9\right)\right)$ at $X_C \simeq 4\pi^2 a/3$.	$k_B T$ $E_{\text{bend}} \geq 0$
X_C	Critical length of DNA at which the bending energy will be at minimum. Particularly, $X_C \simeq 4\pi^2 a/3$ which can be obtained by solving $dE_{\text{bend}} / dX = 0$ for X. When $X < X_C$ then $E_{\text{bend}} \propto X^{-1}$. When $X > X_C$ then $E_{\text{bend}} \propto \ln(X)$.	bp
D_C	One-dimensional diffusion coefficient associated with the sliding of TFs on DNA.	$bp^2\ s^{-1}$
$T_B(X_0)$	MFPT required by TF to reach the promoter starting from its CRM (X_0) via DNA-loop mediated propulsion mechanism. $T_B(X_0) = \left(L^2 - X_0^2\right)/2D_C + G/2D_C$, where the function G is defined as in **Eq. S3**. Here X_0 is the initial loop length as well as the left side reflecting boundary and L is the right-side absorbing boundary.	s
$T_N(X_0)$	MFPT required by TF to reach the promoter starting from its CRM (X_0) via pure 1D sliding mechanism. For an arbitrary starting position X, we have $T_N(X) = \left(L^2 - X^2\right)/2D_C - X_0(L-X)/D_C$. Here $X = X_0$ is the reflecting boundary and $X = L$ is the absorbing boundary.	s
η_P	$\eta_P = \left[T_N(X_0)/T_B(X_0)\right]$. Here the subscript 'N' denotes normal sliding and 'B' denotes the propulsion model.	dimensionless

$T_U(X)$	MFPT required by TF to reach the promoter starting via tethered sliding mechanism. Here initial loop length is X, with left reflecting barrier at X_0 and right absorbing barrier at L.	s
η_S	$\eta_S = \left[T_N(X)/T_U(X)\right]$. Here the subscript 'N' denotes normal sliding and 'U' denoted the tethered sliding model.	dimensionless
L	Distance between the endpoint of CRMs from the promoter or transcription start site.	bp
L_C	In the tethered sliding model, when the left reflecting boundary was at X_0, then one finds the critical distance between CRMs and promoter L_C such that $\eta_S > 1$ when $L < L_C$ then $\eta_S < 1$ when $L > L_C$. When $X_0 < 100$ bp, then one finds that $L_C \sim 3X_C$.	bp
L_{opt}	Distance between CRM and promoter at which the ratio η_P is at a maximum. Approximately, $L_{\text{opt}} \sim 3X_0$	bp
ΔS_{loop}	Entropy change upon looping of DNA, $\Delta S_{\text{loop}} \simeq \ln\left(P_l/P_{all}\right)$ where P_l is the probability of finding looped conformations of DNA and P_{all} is the probability of finding all the possible conformations (equal to one).	k_B
k_r	$k_r \simeq k_r^0 \exp\left(-\mu_{NS}\right)$, dissociation rate constant related to the nonspecific DNA-TF complex where μ_{NS} is the nonspecific binding energy barrier measured in k_BT and k_r^0 is the dissociation rate at zero energy barrier.	s^{-1}
k_{NL}	Rate of loop formation with nonspecific contact $k_{NL} \simeq k_t \exp\left(-E\right)$ in tethered sliding model. Here $k_t \simeq k_{on}\xi$ where k_{on} (s^{-1} M^{-1}) is the Smolochowski type 3D diffusion-controlled bimolecular collision rate limit and ξ is the concentration (mol/lit, M) of DBD2 under *in vivo* conditions. and E is the potential associated with the loop formation.	s^{-1}
L_S	$L_S \simeq \sqrt{2D_C/k_r}$, Average sliding length of nonspecifically bound TF on DNA before it dissociates where k_r is the dissociation rate constant.	bp

SUPPORTING REFERENCES

1. Niranjani, G., and R. Murugan. 2016. Generalized theory on the mechanism of site-specific DNA–protein interactions. Journal of Statistical Mechanics: Theory and Experiment 2016:053501.
2. Murugan, R. 2018. Theory of Site-Specific DNA-Protein Interactions in the Presence of Nucleosome Roadblocks. Biophysical Journal 114:2516-2529.
3. Atkins, P. W. 1978. Physical chemistry. W.H. Freeman, San Francisco.
4. Niranjani, G., and R. Murugan. 2016. Theory on the mechanism of site-specific DNA–protein interactions in the presence of traps. Physical Biology 13:046003.
5. Gardiner, C. W. 1985. Handbook of stochastic methods for physics, chemistry, and the natural sciences. Springer-Verlag, Berlin; New York.
6. Murugan, R. 2007. Generalized theory of site-specific DNA-protein interactions. Phys Rev E Stat Nonlin Soft Matter Phys 76:011901.
7. Yang, W. Y., and M. Gruebele. 2003. Folding at the speed limit. Nature 423:193-197.
8. Hammar, P., M. Wallden, D. Fange, F. Persson, O. Baltekin, G. Ullman, P. Leroy, and J. Elf. 2014. Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. Nat Genet 46:405-408.
9. Li, G.-W., O. G. Berg, and J. Elf. 2009. Effects of macromolecular crowding and DNA looping on gene regulation kinetics. Nat Phys 5:294-297.
10. Amitai, A. 2018. Chromatin Configuration Affects the Dynamics and Distribution of a Transiently Interacting Protein. Biophys J 114:766-771.
11. Abramowitz, M., and I. A. Stegun. 1965. Handbook of mathematical functions, with formulas, graphs, and mathematical tables. Dover Publications, New York.
12. Shvets, A. A., and A. B. Kolomeisky. 2016. Crowding on DNA in Protein Search for Targets. J Phys Chem Lett 7:2502-2506.