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

Rajamanickam Murugan

Department of Biotechnology, Indian Institute of Technology Madras
Chennai, India. Email: rmurugan@gmail.com

ABSTRACT

We demonstrate that DNA-loops can stochastically propel the site-specifically bound transcription factors (TFs) towards the promoters. The gradual release of elastic energy stored on the DNA-loops is the source of propulsion. The speed of looping mediated interaction of TFs with promoters is several times faster than the sliding mode. Elastic and entropic energy barriers associated with the looping actually shape up the distribution of distances between TF binding sites and promoters. The commonly observed multiprotein binding in gene regulation is acquired through evolution to overcome the looping energy barrier. Presence of nucleosomes on the genomic DNA of eukaryotes is required to reduce the entropy barriers associated with the looping.

INTRODUCTION

Looping of DNA is critical for the activation and expression of various genes across prokaryotes to eukaryotes [1-6]. Binding of transcription factors (TFs) with their specific *cis*-regulatory motifs (CRMs) on the genomic DNA activates the downstream promoters of genes via looping of the intervening DNA segment to form a synaptosome type complex [7]. In most of the biological processes, looping of DNA is warranted for the precise protein-protein interactions required for the gene expression and recombination [8]. The statistical mechanics of looping and cyclization of DNA has been studied extensively [4, 9, 10]. However, it is still not clear why the DNA-looping is an integral part of the transcription activation and repression although such underlying site-specific protein-protein and protein-DNA interactions can also be catered via a combination of one (1D) and three-dimensional (3D) diffusions of TFs [11-14]. It is also not clear how exactly the DNA-loop is formed between the CRMs and promoters via TFs though Rippe et.al., [3] had already taken the snapshots of the looping intermediates. Schleif [2] 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 promoters. In this letter, we will show that the DNA-looping combined with an asymmetric binding energy profile can stochastically propel TFs towards the promoters along DNA. We further demonstrate that the physics behind the looping mediated propulsion of TFs along DNA actually shapes up the genomic architecture.

THEORY

In our model, we assume that 1) TFs of interest (or multiprotein complex) has two different DNA binding domains (DBDs) corresponding to CRMs (DBD1) and the promoters (DBD2) similar to the synaptic complexes of transcription activation in eukaryotes and 2) TF reaches its specific binding site on DNA via a combination of 3D and 1D diffusions within the theoretical framework of Berg-Winter-Hippel [12-16]. Here 1D diffusion is always slower than the 3D diffusion. Therefore, any factor which speeds up the sliding will eventually speed up the overall searching of TFs. Site-specific binding of TFs with their CRMs causes bending of DNA [3, 4]. The *site-specific binding energy* (E_{bind}) released at the DNA-TF interface will be dissipated partially as the elastic energy required to bend the DNA (E_{elastic}), partially to form specific non-covalent bonds (E_{bond}) and partially as the energy required to compensate the chain entropy loss (E_{entropy}) at the specific binding site. The energy stored by the specific DNA-TF complex is $E \simeq E_{\mathrm{bond}} + E_{\mathrm{elastic}}$. In these settings, the DBD2 of TF needs to distally interact with the promoter and activate the transcription via looping of the intervening DNA segment.

Let us assume that the radius of gyration of TF is r_P. Upon binding with its cognate stretch of DNA with size of X_0 base-pairs (bp, where 1 bp = l_p = 3.4 x $10^{-10\ \mathrm{m}}$) located in between S1 to S2, the TF bends the DNA segment into a circle around its spherical solvent shell surface such that $X_0 = 2\pi r_P$ as shown in **Fig 1A**. We set $X = 0$ at S1 and $X = X_0$ at S2 where X is the current location of the DBD2 of TF on DNA that spans over (0, L) as in **Fig. 1B**. Here X is also the loop-length. The energy required to bend a linear DNA will be $E_{\mathrm{bend}} = E_{\mathrm{elastic}} + E_{\mathrm{entropy}}$. For a radius of curvature of r_P, $E_{\mathrm{elastic}} \simeq aX/2r_P^2$ (k_BT units) where a is the persistence length of DNA [9, 17]. Clearly, E_{elastic} required to bend X length of DNA into a loop will be $E_{\mathrm{elastic}} \simeq 2\pi^2 a/X$. This energy has to be derived either from the specific binding energy of TFs or via an external energy input in the form of ATP hydrolysis [18]. We will investigate E_{entropy} later.

The energy stored in the site-specific DNA-TF complex can undergo three different modes of dissipation viz. 1) thermal induced physical dissociation of TF from DNA in which both bonding and elastic energies dissipate into the heat bath along with increase in the chain entropy, 2) physical dissociation of only DBD2 from S2 and re-association somewhere via looping over 3D space while S1-DBD1 is still intact as modelled by Shvets and Kolomeisky [19] and 3) stochastic propulsion of TF on DNA via sliding of DBD2 which can be achieved by gradual increase in the value of X from X_0. Here mainly the elastic energy dissipates that in turn causes bulging of DNA-loop around TF. The chain entropy does not increase much here since the intervening DNA is still under looped conformation. This is similar to the sliding of nucleosome via bulge induced reptation dynamics of DNA [20-22]. The probability of spontaneous dissociation will be inversely correlated with the bonding energy and it is an endothermic process. Clearly, physical dissociation will not be the most probable route of dissipation of the energy stored in the site-specific DNA-TF complex.

When the binding energy profile of TF is such that the bonding energy near S1 is much higher than S2, then the bending energy stored in the site-specific TF-DNA complex can be gradually released via bulging of DNA-loop around TF which in turn stochastically propels the sliding DBD2 of TF towards the promoter located at L as shown in **Fig. 1B**. Rippe et.al [3] have studied NtrC system where binding of NtrC at its specific site activates the downstream closed complex of *glnA* promoter-RNAP-σ^{54} via looping out of the intervening DNA. They have shown that the transition from inactive-closed to an active-open promoter complex involved an increase in the bending angle of the intervening DNA which in turn is positively correlated with an increase in the radius of curvature. This is represented as bulging of DNA-loop in our model. Therefore, our assumption that the propulsion of TFs via increase in the radius of curvature of the bent DNA is a logical one. Here the asymmetric binding energy profile is essential to break the symmetry of the stochastic force acting on the sliding TFs [23]. This is also a logical assumption since S1-DBD1 is a strong site-specific interaction and S2-DBD2 is a nonspecific interaction. **Fig. 1C** shows another possibility in the formation of DNA-loop which is common in case of silencer TFs. Based on these, the position X of TF on DNA obeys the following Langevin equation [24-26].

$$dX/dt = D_C F(X) + \sqrt{2D_C}\Gamma_t;\ \langle \Gamma_t \rangle = 0;\ \langle \Gamma_t \Gamma_{t'} \rangle = \delta(t-t')\ . \tag{1}$$

In **Eq. 1**, $F(X) = -dE/dX = 2\pi^2 a/X^2$ is the force acting on TF that is generated by the bending potential $E \sim E_{\text{elastic}}$ upon bulging of the DNA-loop, Γ_t is the Δ-correlated Gaussian white noise and D_C is the 1D diffusion coefficient of the sliding of TF. The energy involved in the bonding interactions will be a constant one so that it will not contribute to the force term. Here we ignore the energy dissipation via chain entropy mainly because binding of TFs at their specific sites attenuates the conformational fluctuations at the DNA-TF interface [12, 15, 27]. The Fokker-Planck equation describing the probability of observing a given X at time t with the conditions that $X = X_0$ at $t = t_0$ can be written as follows [24, 25].

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

The form of $F(X)$ suggests that it can propel the DBD2 of TF 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. Initial condition for **Eq. 2** 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,

$$P(L,t \mid X_0,t_0) = 0;\ \left[F(X)P(X,t \mid X_0,t_0) - \partial_X P(X,t \mid X_0,t_0)\right]_{X=X_0} = 0\,. \tag{3}$$

Here X_0 acts as a reflecting boundary for a given size of TF and L is the absorbing boundary where the promoter is located. The asymmetric energy profile with respect to S1 and S2 is required for the validity of the reflecting boundary condition at X_0. Upon reaching the promoter via loop-expansion of the intervening DNA segment, TFs subsequently activate the transcription. The mean first passage time $T_B(X)$ associated with the DBD2 of TF to reach L starting from $X \in (X_0, L)$ obeys the following backward type Fokker-Planck equation along with the appropriate boundary conditions [15, 28].

$$D_C\left[F(X)d_X T_B(X) + d_X^2 T_B(X)\right] = -1;\ T_B(L) = 0;\ d_X T_B(X_0) = 0\,. \tag{4}$$

The solution of **Eqs. 4** can be expressed as follows.

$$T_B\left(X\right)=\left(2\pi^2 a/D_C\right)\int_X^L \left\{\begin{matrix} Z/2\pi^2 a+\exp\left(2\pi^2 a/Z\right)\left[\mathrm{Ei}_1\left(2\pi^2 a/X_0\right)-\mathrm{Ei}_1\left(2\pi^2 a/Z\right)\right] \\ -\left(X_0/2\pi^2 a\right)\exp\left(2\pi^2 a\left(X_0-Z\right)/X_0 Z\right)\end{matrix}\right\} dZ\ . \qquad (5)$$

Here $\mathrm{Ei}_1\left(Y\right)=\int_1^\infty\left[\exp\left(-sY\right)/s\right]ds$ [29] 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 DBD2 of TF to reach L via pure 1D sliding in the absence of DBD1 which is a solution of the following differential equation [12, 15, 28].

$$D_C d_X^2 T_N\left(X\right)=-1;\ T_N\left(L\right)=0;\ d_X T_N\left(X_0\right)=0;\ T_N\left(X\right)=\left(L^2-X^2\right)/2D_C-X_0\left(L-X\right)/D_C\ . \qquad (6)$$

To obtain the target finding time, one needs to set $X = X_0$ in **Eqs. 5** and **6**. One can define the number of times the target finding rate of TF can be accelerated by the looping mediated propulsion over 1D sliding as $\eta=T_N\left(X_0\right)/T_B\left(X_0\right)$ which is clearly independent of D_C of TF and solely depends on (L, a, and X_0). Explicitly one can write it as,

$$\eta=\left(L-X_0\right)^2\left(\left(L^2-X_0^2\right)+4\pi^2 a\int_{X_0}^L\left\{\begin{matrix}\exp\left(2\pi^2 a/Z\right)\left[\mathrm{Ei}_1\left(2\pi^2 a/X_0\right)-\mathrm{Ei}_1\left(2\pi^2 a/Z\right)\right]\\ -\left(X_0/2\pi^2 a\right)\exp\left(2\pi^2 a\left(X_0-Z\right)/X_0 Z\right)\end{matrix}\right\}dZ\right)^{-1}. \qquad (7)$$

This is the central result of this letter. Detailed numerical analysis (Supporting Material) suggests that there exists a maximum of η at which $\partial\eta/\partial L=0$ with $L = L_{\text{opt}}$ and clearly, we have $\lim_{L\to\infty}\eta=1$ (**Figs. 2A** and **B**). This is logical since when $L > L_{\text{opt}}$ then $\eta\to 1$ and when $L < L_{\text{opt}}$ then the stored energy is not completely utilized to propel the DBD2 of TF. Further, $\lim_{L\to X_0}\eta=0$ since its numerator part goes to zero much faster than the denominator (**Fig. S1**, Supporting Material). The persistence length of typical DNA under *in vitro* conditions is a ~ 150 bp and the radius of gyration for most of the eukaryotic TFs will be r_P ~ 10-15 bp. Therefore, one can set the initial $X = 2\pi r_P$ ~ 50-100 bp [30, 31]. Simulations (**Fig. 2A**) on η at different values of X_0 and, L from X_0 to 10^5 suggested that L_{opt} ~ $3X_0$ (see **Figs. 2C** and **2D**). When a ~150 bp and X_0 ~ 50-100 bp, then L_{opt} ~ 150-300 bp. Remarkably, this is the most probable range of the distances between the CRMs and promoters of various genes observed across several genomes [32].

RESULTS AND DISCUSSION

The efficiency of the stochastic propulsion will be a maximum at L_{opt}. Although L_{opt} is not much affected by a, the maximum of η is positively correlated with a. This is logical since the stored elastic energy is directly proportional to a. Remarkably at L_{opt} the speed of interactions between CRM-TFs complex with the promoters will be ~10-25 times faster than the normal 1D sliding. These results are demonstrated in **Figs. 2**. Here we assumed that the nonspecifically bound DBD2 of TF does not dissociate until reaching the promoter which is valid only for $L\simeq\sqrt{2D_C/k_r}$ where k_r is the dissociation rate constant [15] that is defined as $k_r\simeq k_r^0\exp\left(-\mu_{NS}\right)$ where k_r^0~10^6 s^{-1} and μ_{NS} is the nonspecific binding energy associated with DBD2. Clearly $\mu_{NS} \geq 12\ k_BT$ is required to attain L ~ 300 bp which can be achieved via multiprotein binding.

Noting that $E_{\text{elastic}} \simeq 2\pi^2 a/X \sim 3000/X$ (for $a \sim 150$ bp), E_{entropy} component for a Gaussian chain can be computed as follows. Let us assume that looping occurs when $|\bar{R}| \le \xi$ where $|\bar{R}| \in (\xi, Xl_d)$ is the end-to-end distance vector, ξ is the minimum looping-distance (in m) and Xl_d is the maximum length of the DNA polymer. The density function of $\bar{R}$ is $p(\bar{R}) \simeq (3/2\pi X b^2)^{3/2} \exp(-3\bar{R}^2/2Xb^2)$ [33, 34] where X is the number of monomers in the polymer and b is the average distance between the monomers. The entropy loss upon looping of DNA is $\Delta S_{\text{loop}} \simeq \ln(P_l/P_{all})$ (k_B units) where $P_l \simeq \int_0^{\xi} p(\bar{R})\, d\bar{R}$ is the probability of finding loops and $P_{all} = \int_0^{Xl_d} p(\bar{R})\, d\bar{R}$ is the probability of finding all the configurations including loops. Explicitly one can write down ΔS_{loop} as follows.

$$\Delta S_{\text{loop}} \simeq 3\ln\left(\text{Erf}\left(\sqrt{3\xi^2/2Xb^2}\right)\Big/\text{Erf}\left(\sqrt{3Xl_d^2/2b^2}\right)\right) \quad . \tag{8}$$

Here $\text{Erf}(Z) = \sqrt{4/\pi}\int_0^Z \exp(-Y^2)\, dY$ is the error function [29]. When $\xi \simeq b \simeq l_d$ is small then $\Delta S_{\text{loop}} \simeq 3\ln\left(\text{Erf}\left(\sqrt{3X/2}\right)\sqrt{6/\pi X}\right) \simeq -(3/2)\ln(\pi X/6)$ for large values of X [19]. This expression is closely linked with the Jacobson-Stockmayer factor, or J-factor associated with polymer looping [9]. One finally obtains that $E_{\text{entropy}} \simeq (3/2)\ln(\pi X/6)$.

Clearly, bending of linear DNA with size of 50-100 bp into loops requires the hydrolysis of at least 3-5 ATP molecules (using $E_{\text{bend}} = E_{\text{elastic}} + E_{\text{entropy}}$, 1 ATP ~ 12 k_BT). 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 synaptosome will be at minimum [19]. When $X < X_C$ then $E_{\text{bend}} \propto X^{-1}$. When $X > X_C$ then $E_{\text{bend}} \propto \ln(X)$. When $a \sim 150$ bp and $X_C \sim 2$ kbp then the minimum of $E_{\text{bend}} \sim 13\ k_BT$ which requires the hydrolysis of at least 1 ATP. These results are demonstrated in **Fig. 3**. To simplify our model, we have ignored the entropic barriers imposed by the flanking regions of DNA. However, it increases only in a logarithmic manner along the chain length compared to the elastic energy. In the absence of energy input, biological systems can overcome the looping energy barrier via three possible ways viz. 1) multiprotein binding [10] which could be the origin of the combinatorial regulatory TFs in the process of evolution, 2) placing sequence mediated kinetic traps corresponding to DBD2 in between CRMs and promoters [35] and 3) the 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 [10] 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 TF search dynamics and the energy barrier associated with the DNA-looping. As a result, natural systems optimize X_0 between these two-opposing factors for maximum efficiency via manipulating the number of TFs in the combinatorial binding. We conclude here with the open fundamental question. What is the exact mechanism of DNA-loop mediated transcription activation in the real systems? Is it via the stochastic propulsion of our model or via the repeated association-dissociation of DBD2 as proposed [19] by Shvets and Kolomeisky? Future single molecule experiments need to address these basic questions.

CONCLUSION

In summary, for the first time we have shown that DNA-loops can stochastically propel the transcription factors along DNA from their specific binding sites towards the promoters. We have shown that the source of propulsion is the elastic energy stored on the specific looped DNA-protein complex. Actually, elastic and entropic energy barriers associated with the looping of DNA shape up the distribution of distances between TF binding sites and promoters. 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 is required to reduce the entropy barrier associated with the looping.

REFERENCES

1. Schleif R. DNA looping. Science. 1988;240(4849):127-8. PubMed PMID: 3353710.
2. Schleif R. DNA looping. Annu Rev Biochem. 1992;61:199-223. doi: 10.1146/annurev.bi.61.070192.001215. PubMed PMID: 1497310.
3. Rippe K, Guthold M, von Hippel PH, Bustamante C. 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. 1997;270(2):125-38. PubMed PMID: 9236116.
4. Mulligan PJ, Chen YJ, Phillips R, Spakowitz AJ. Interplay of Protein Binding Interactions, DNA Mechanics, and Entropy in DNA Looping Kinetics. Biophys J. 2015;109(3):618-29. doi: 10.1016/j.bpj.2015.06.054. PubMed PMID: 26244743; PubMed Central PMCID: PMCPMC4572505.
5. Ptashne M, Gann A. Genes & signals. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 2002.
6. Lewin RA, Crothers DM, Correll DL, Reimann BE. A Phage Infecting Saprospira Grandis. Can J Microbiol. 1964;10:75-85. PubMed PMID: 14124864.
7. Murugan R. Theory on the mechanism of distal action of transcription factors: looping of DNA versus tracking along DNA. Journal of Physics A: Mathematical and Theoretical. 2010;43(41):415002.
8. Grindley ND, Whiteson KL, Rice PA. Mechanisms of site-specific recombination. Annu Rev Biochem. 2006;75:567-605. doi: 10.1146/annurev.biochem.73.011303.073908. PubMed PMID: 16756503.
9. Zhang Y, McEwen AE, Crothers DM, Levene SD. Statistical-mechanical theory of DNA looping. Biophys J. 2006;90(6):1903-12. doi: 10.1529/biophysj.105.070490. PubMed PMID: 16361335; PubMed Central PMCID: PMCPMC1386771.
10. Vilar JM, Saiz L. Multiprotein DNA looping. Phys Rev Lett. 2006;96(23):238103. doi: 10.1103/PhysRevLett.96.238103. PubMed PMID: 16803410.
11. Murugan R. Theory on thermodynamic coupling of site-specific DNA–protein interactions with fluctuations in DNA-binding domains. Journal of Physics A: Mathematical and Theoretical. 2011;44(50):505002.
12. Murugan R. Theory of site-specific DNA-protein interactions in the presence of conformational fluctuations of DNA binding domains. Biophys J. 2010;99(2):353-9. doi: 10.1016/j.bpj.2010.04.026. PubMed PMID: 20643052; PubMed Central PMCID: PMCPMC2905069.
13. Berg OG, Winter RB, von Hippel PH. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. Biochemistry. 1981;20(24):6929-48. PubMed PMID: 7317363.

14. Winter RB, Berg OG, von Hippel PH. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor--operator interaction: kinetic measurements and conclusions. Biochemistry. 1981;20(24):6961-77. PubMed PMID: 7032584.

15. Murugan R. Generalized theory of site-specific DNA-protein interactions. Phys Rev E Stat Nonlin Soft Matter Phys. 2007;76(1 Pt 1):011901. doi: 10.1103/PhysRevE.76.011901. PubMed PMID: 17677488.

16. Murugan R. Directional dependent dynamics of protein molecules on DNA. Phys Rev E Stat Nonlin Soft Matter Phys. 2009;79(4 Pt 1):041913. doi: 10.1103/PhysRevE.79.041913. PubMed PMID: 19518262.

17. Zhang Y, Crothers DM. Statistical mechanics of sequence-dependent circular DNA and its application for DNA cyclization. Biophys J. 2003;84(1):136-53. doi: 10.1016/S0006-3495(03)74838-3. PubMed PMID: 12524271; PubMed Central PMCID: PMCPMC1302599.

18. Spirin AS. How does a scanning ribosomal particle move along the 5'-untranslated region of eukaryotic mRNA? Brownian Ratchet model. Biochemistry. 2009;48(45):10688-92. doi: 10.1021/bi901379a. PubMed PMID: 19835415.

19. Shvets AA, Kolomeisky AB. The Role of DNA Looping in the Search for Specific Targets on DNA by Multisite Proteins. J Phys Chem Lett. 2016;7(24):5022-7. doi: 10.1021/acs.jpclett.6b02371. PubMed PMID: 27973894.

20. Schiessel H, Widom J, Bruinsma RF, Gelbart WM. Polymer reptation and nucleosome repositioning. Phys Rev Lett. 2001;86(19):4414-7. Epub 2001/05/01. doi: 10.1103/PhysRevLett.86.4414. PubMed PMID: 11328188.

21. Kulic IM, Schiessel H. Nucleosome repositioning via loop formation. Biophys J. 2003;84(5):3197-211. Epub 2003/04/30. doi: 10.1016/S0006-3495(03)70044-7. PubMed PMID: 12719249; PubMed Central PMCID: PMCPMC1302880.

22. Murugan R. Theory of Site-Specific DNA-Protein Interactions in the Presence of Nucleosome Roadblocks. Biophysical Journal. 2018;114(11):2516-29. doi: 10.1016/j.bpj.2018.04.039.

23. Lee Y, Allison A, Abbott D, Stanley HE. Minimal Brownian ratchet: an exactly solvable model. Phys Rev Lett. 2003;91(22):220601. doi: 10.1103/PhysRevLett.91.220601. PubMed PMID: 14683223.

24. Gardiner CW. Handbook of stochastic methods for physics, chemistry, and the natural sciences. Berlin; New York: Springer-Verlag; 1985.

25. Risken H. The Fokker-Planck equation : methods of solution and applications. Berlin; New York: Springer-Verlag; 1989.

26. Kampen NGv. Stochastic processes in physics and chemistry. Amsterdam; New York; New York: North-Holland ; Sole distributors for the USA and Canada, Elsevier North-Holland; 1981.

27. Kalodimos CG, Biris N, Bonvin AM, Levandoski MM, Guennuegues M, Boelens R, et al. Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. Science. 2004;305(5682):386-9. doi: 10.1126/science.1097064. PubMed PMID: 15256668.

28. Murugan R. DNA-protein interactions under random jump conditions. Phys Rev E Stat Nonlin Soft Matter Phys. 2004;69(1 Pt 1):011911. doi: 10.1103/PhysRevE.69.011911. PubMed PMID: 14995651.

29. Abramowitz M, Stegun IA. Handbook of mathematical functions, with formulas, graphs, and mathematical tables. New York: Dover Publications; 1965.

30. Wingender E, Dietze P, Karas H, Knuppel R. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res. 1996;24(1):238-41. PubMed PMID: 8594589; PubMed Central PMCID: PMCPMC145586.

31. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018;46(D1):D1284. doi: 10.1093/nar/gkx1188. PubMed PMID: 29161433; PubMed Central PMCID: PMCPMC5753202.

32. Koudritsky M, Domany E. Positional distribution of human transcription factor binding sites. Nucleic Acids Res. 2008;36(21):6795-805. doi: 10.1093/nar/gkn752. PubMed PMID: 18953043; PubMed Central PMCID: PMCPMC2588498.

33. Doi M, Edwards SF. The theory of polymer dynamics. Oxford [Oxfordshire]: Clarendon Press; 1988.

34. Gennes P-Gd. Scaling concepts in polymer physics. Ithaca, N.Y.: Cornell University Press; 1979.

35. Niranjani G, Murugan R. Theory on the mechanism of site-specific DNA–protein interactions in the presence of traps. Physical Biology. 2016;13(4):046003.

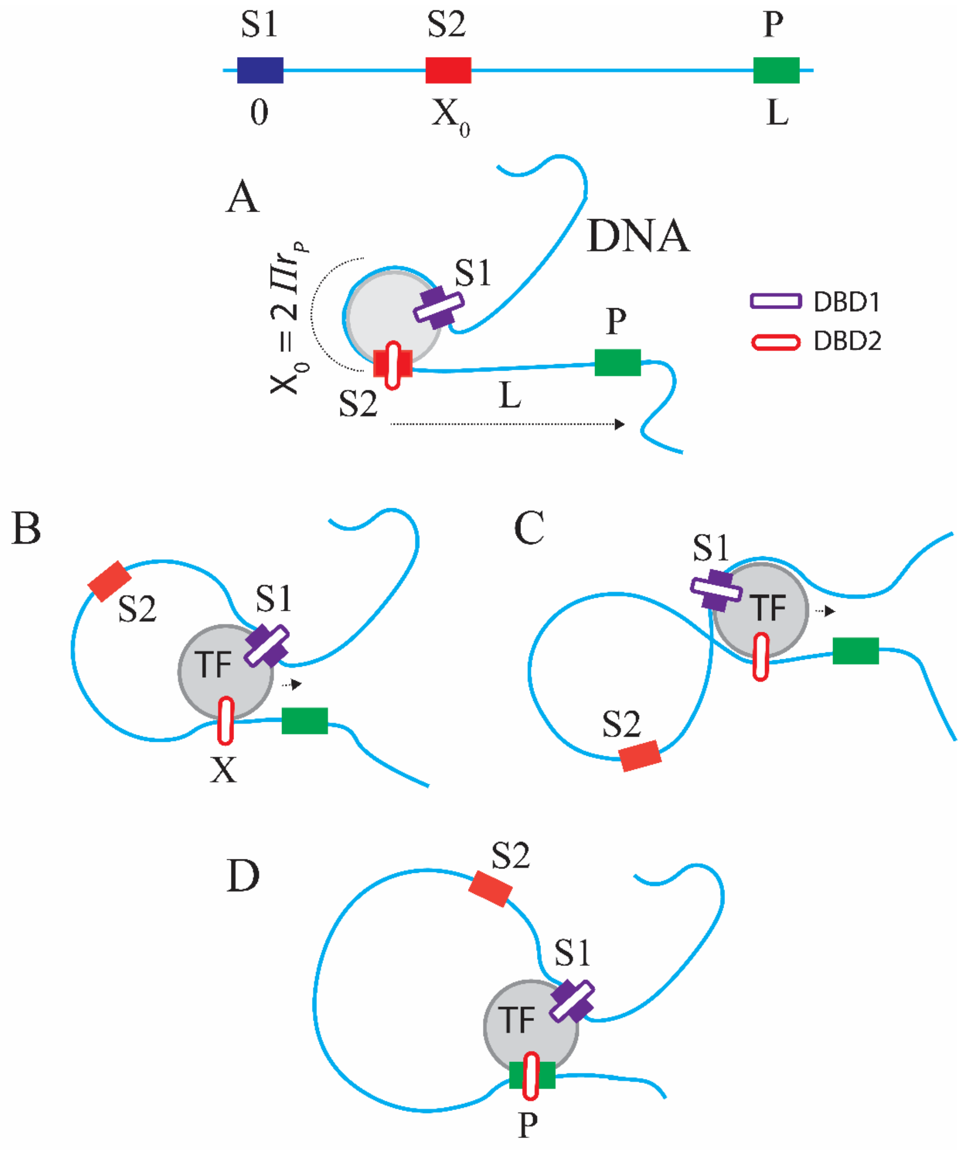

FIG 1. **A**. Looping mediated stochastic propulsion of TF with radius of gyration of r_P along DNA. Here TF has two binding sites corresponding to viz. its *cis*-regulatory module (DBD1) located between S1 and S2 and the promoter (DBD2). Binding of TF with its specific site (that spans for a length of X_0 from S1 ($X = 0$) to S_2 ($X = X_0$)) bends the DNA segment into a loop around it such that $X_0 = 2\pi r_P$. The bending energy stored in the site-specific complex will be incrementally

released via bulging of DNA around the TF. **B**. When the binding energy near S1 is stronger than S2, then the TF can be stochastically propelled towards the promoter (P) that is located at $X = L$. Upon reaching the promoter, DBD2 of TF interacts with the promoter to form a specific synaptic complex. **C**. DNA-loop configuration utilized for gene silencing. **D**. Synaptosome where TF is bound with both its specific binding site and the promoter.

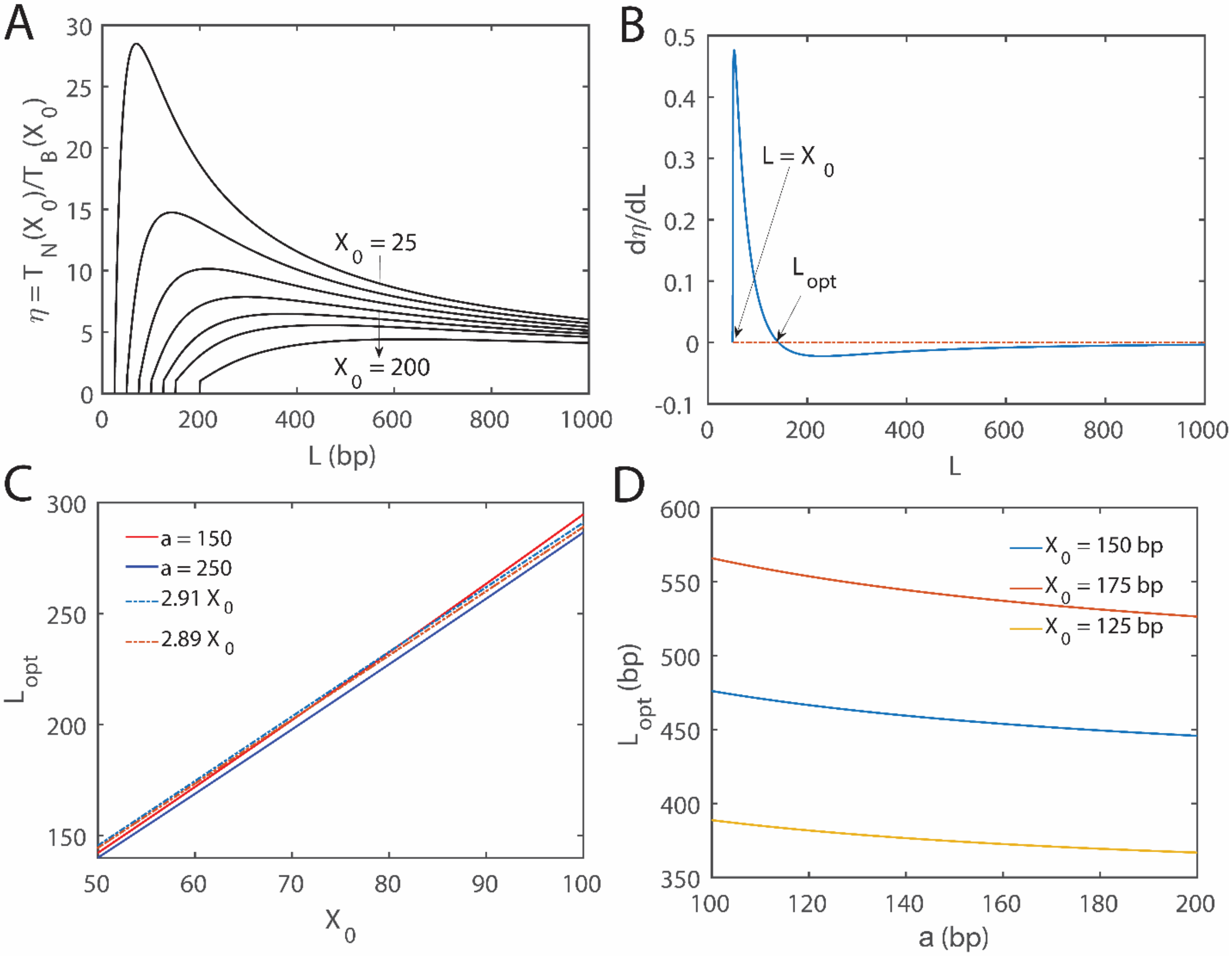

FIG 2. **A.** Relative efficiency of looping mediated stochastic propulsion of TFs versus normal 1D sliding along DNA. $T_N(X_0)$ is the mean first passage time that is required by TFs to reach the promoter that is located at L, starting from X_0 via 1D sliding. $T_B(X_0)$ is the mean first passage time 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 $d\eta/dL$ with respect to L. Here the settings are a ~ 150 bp and X_0 ~ 50 bp and L was iterated from 50 to 1000 bp. Upon solving $d\eta/dL = 0$ for L numerically one finds that L_{opt} ~ 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 $L_{\mathrm{opt}} \sim 3X_0$ seems to be < 10% over wide range of *a* values.

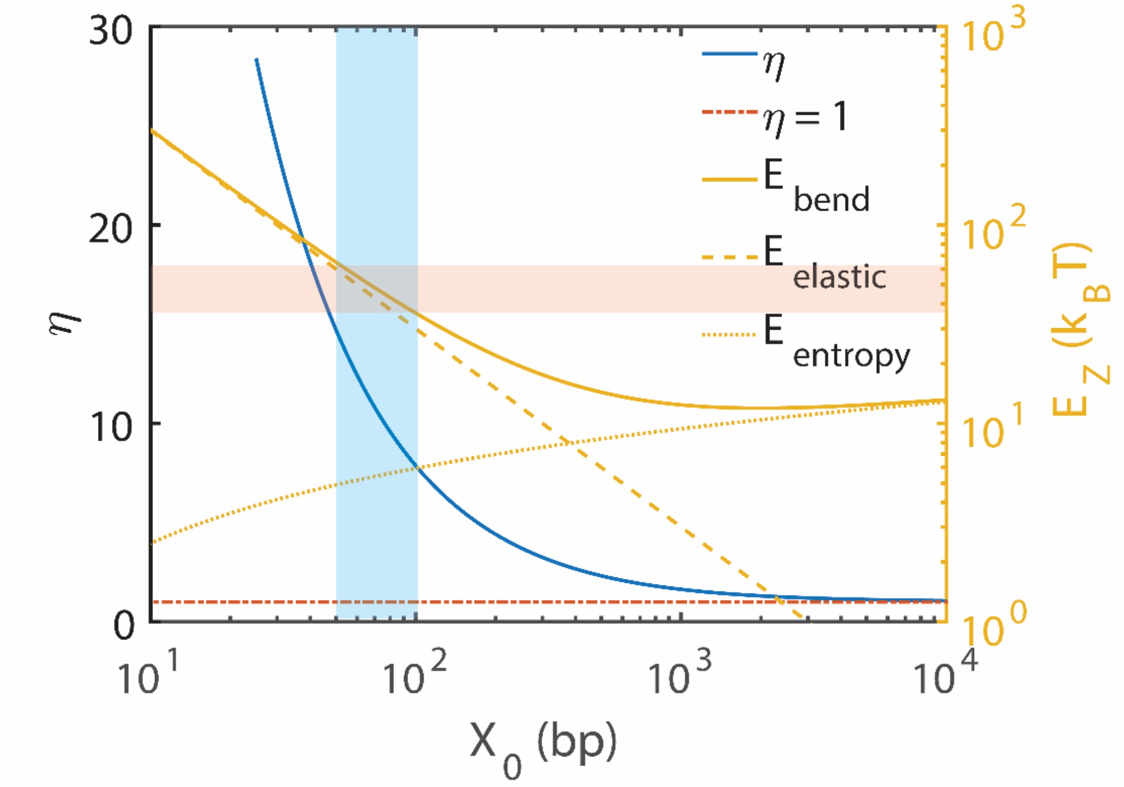

FIG 3. Variation of the propulsion efficiency η and bending energy with respect to changes in the initial loop length X_0. Here the settings are a ~ 150 bp, $L = 3X_0 \sim L_{\mathrm{opt}}$. We computed $\eta = T_N\left(X_0\right)/T_B\left(X_0\right)$ with $L = 3X_0$ so that η will be close to its maximum. $E_{\mathrm{bend}} = E_{\mathrm{elastic}} + E_{\mathrm{entropy}}$ where $E_{\mathrm{elastic}} \simeq 3000/X_0$ and $E_{\mathrm{entropy}} \simeq 3\ln\left(\pi X_0/6\right)/2$ which is ~12 k_BT at X_0 ~ 2 kb. The elastic energy $E_{\mathrm{elastic}} \leq 1$ k_BT when $X_0 \geq$ 3kb. Clearly, the bending energy of linear DNA is always ≥ 12 k_BT irrespective of the length. Shaded regions are the most probable X_0 values observed in the natural systems where the optimum distance between the transcription factor binding sites and promoters $L_{opt} \sim 3X_0 \sim$ 150-300 bp.

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

Rajamanickam Murugan

Department of Biotechnology, Indian Institute of Technology Madras
Chennai, India. Email: rmurugan@gmail.com

Supporting Material

The mean first passage time (MFPT) associated with the DNA binding domain 2 (DBD2) of TF 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 (see **Fig.1** of main text for details) can be given as follows.

$$T_B\left(X_0\right)=\frac{L^2-X_0^2}{2D_C}+\frac{G}{2D_C} \qquad \text{(S1)}$$

Here X_0 is the initial position of DBD2 of TF on DNA or the initial loop length, L is the location of the promoter, a is the persistence length of DNA and D_C is the one-dimensional diffusion coefficient associated with the sliding of DBD2 of TF along DNA. We have assumed here that the site-specific DBD1-S1 is strong and intact (**Fig. 1A** of main text). The function G can be defined as follows.

$$G=4\pi^2 a\int_{X_0}^{L}\left(\exp\left(\frac{2\pi^2 a}{Z}\right)\left[\mathrm{Ei}_1\left(\frac{2\pi^2 a}{X_0}\right)-\mathrm{Ei}_1\left(\frac{2\pi^2 a}{Z}\right)\right]-\left(\frac{X_0}{2\pi^2 a}\right)\exp\left(\frac{2\pi^2 a\left(X_0-Z\right)}{X_0 Z}\right)\right)dZ \qquad \text{(S2)}$$

Here $\mathrm{Ei}_1\left(Y\right)=\int_1^{\infty}\left[\exp\left(-sY\right)/s\right]ds$ is the E_1 exponential integral [1].

Noting that $T_N\left(X_0\right)=\left(L-X_0\right)^2/2D_C$ (from **Eq. 6** of the main text) which is the MFPT associated with the finding of the promoter by TF starting from X_0 via pure sliding dynamics, one can define η which is the number of times the DNA-loop driven searching of TF for the promoter is faster than the normal 1D sliding dynamics as follows.

$$\eta=\frac{T_N\left(X_0\right)}{T_B\left(X_0\right)}=\frac{\left(L-X_0\right)^2}{L^2-X_0^2+G}=\frac{1}{\left(L+X_0\right)/\left(L-X_0\right)+G/\left(L-X_0\right)^2}\ . \qquad \text{(S3)}$$

Clearly η is not dependent on D_C and it depends only on the parameters (L, X_0 and a). Further, $\lim_{L\to X_0}\eta=0$ since $T_N\left(X_0\right)$ approaches zero much faster than $T_B\left(X_0\right)$ as L tends towards infinity (see **Fig. S1** for details). There also exists an asymptotic limit as $\lim_{L\to\infty}\eta=1$. This means that

$\lim_{L\to\infty}\left[G/\left(L-X_0\right)^2\right]=0$ (see **Fig. S2** for details). The optimum distance between CRMs and promoter i.e. *L*opt at which η is a maximum can be obtained by solving $d\eta/dL=0$ for *L* for given *a* and *X0*. Explicitly one can write down this as follows.

$$\frac{d\eta}{dL}=\frac{2\left(L-X_0\right)\left[-\frac{\left(L-X_0\right)}{2}\frac{\partial G}{\partial L}+G+X_0\left(L-X_0\right)\right]}{\left(G+L^2-X_0^2\right)^2}=0. \qquad \text{(S4)}$$

This has a trivial solution *L* = *X0*. Upon ignoring this one, *L*opt can be obtained by numerically solving the following equation for *L* at given *a* and *X0*.

$$L>X_0;\ \therefore\left(L-X_0\right)\frac{\partial G}{\partial L}-2G-2X_0\left(L-X_0\right)=0. \qquad \text{(S5)}$$

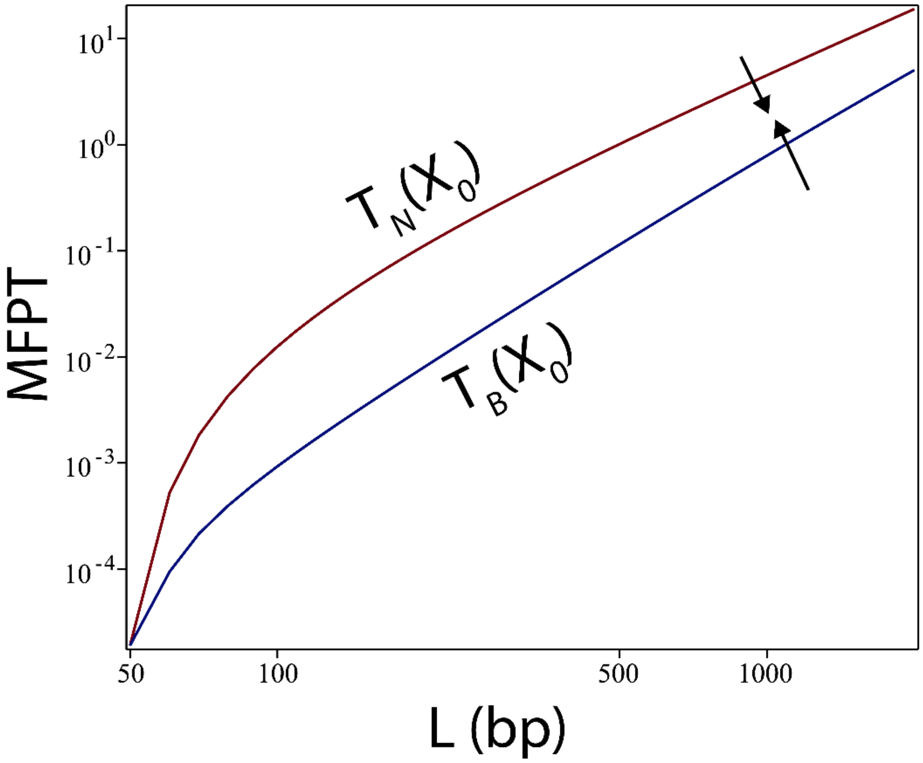

FIG S1. Showing that $T_N\left(X_0\right)$ approaches zero much faster than $T_B\left(X_0\right)$ so that $\lim_{L\to X_0}\eta=0$. Here the settings are *X0* ~ 50 bp and *a* ~ 150 bp and *L* was iterated from 50 to 2000 bp. Further we also find the limit $\lim_{L\to\infty}T_B\left(X_0\right)=T_N\left(X_0\right)$ or $\lim_{L\to\infty}\left[T_N\left(X_0\right)/T_B\left(X_0\right)\right]=1$.

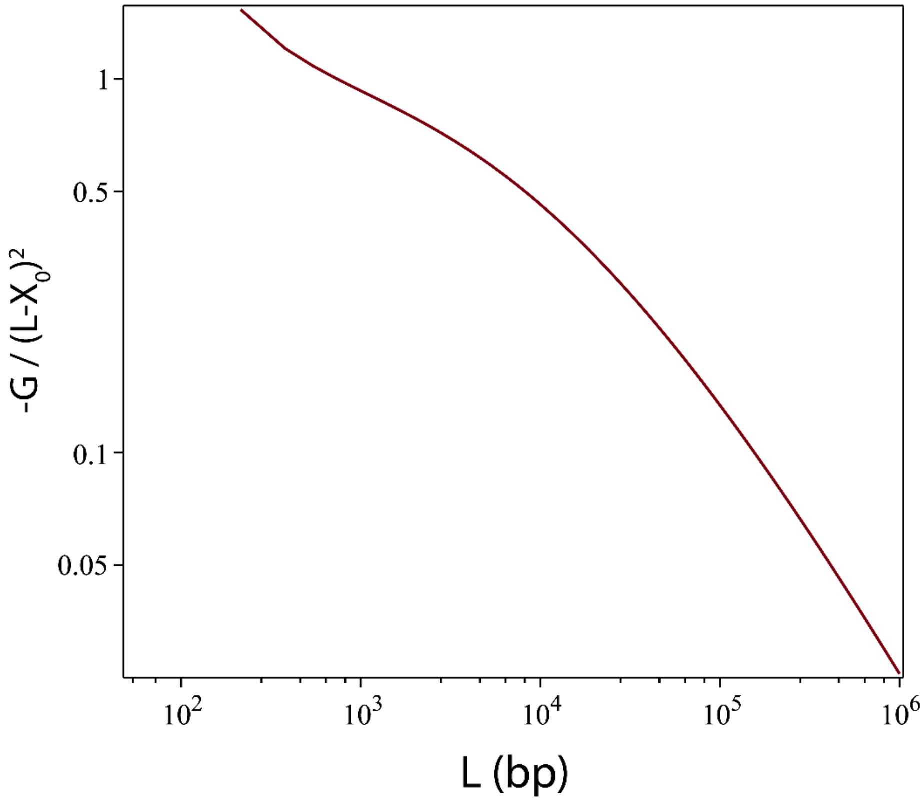

FIG S2. Showing $\lim_{L\to\infty}\left[G/\left(L-X_0\right)^2\right]\to 0$. Here the settings are X_0 ~ 50 bp and *a* ~ 150 bp and *L* was iterated from 50 to 10^6. The function *G* is defined as in **Eq. S2**.

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.	
DBD DBD1 DBD2	DNA Binding Domain of TF proteins; DBD1 of the TF complex specifically binds with CRM; DBD2 of the TF complex specifically interacts with the promoter of the gene that needs to be activated.	
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 TF 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.	
S2	DNA binding sequence present at the location of DBD2 of TFs complex, which is not a specific binding site for DBD2.	
P	Promoter sequence.	bp
a	Persistence length DNA. Typically, it is ~150 bp.	bp
E	Energy stored on the site-specific DNA-TF complex $E \sim E_{\text{bond}} + E_{\text{elastic}}$.	$k_B T$
$F(X)$	It is force generated by the potential E. $F(X) = -dE/dX = 2\pi^2 a/X^2$. This in turn can propel TF towards the promoter.	$k_B T$ / bp
E_{bind}	~ $E_{\text{bond}} + E_{\text{elastic}} + E_{\text{entropy}}$, the total site-specific binding energy associated with the DNA-TF complex dissipates into these three components.	$k_B T$
E_{entropy}	Energy required to compensate the chain entropy of linear DNA that is converted into a loop upon binding of TF, $E_{\text{entropy}} \simeq (3/2)\ln(\pi X/6)$ 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_{bend}	~ E_{elastic} + E_{entropy}, energy required to bend a linear DNA segment into a loop. $E_{\text{bend}} \simeq 2\pi^2 a/X + (3/2)\ln(\pi X/6)$.	k_BT
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_{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
E_{bond}	Resultant energy with respect to all types of bonding interactions present at the DNA-TF interface.	k_BT
D_C	One-dimensional diffusion coefficient associated with the sliding of TFs on DNA.	$\text{bp}^2\ \text{s}^{-1}$
$T_B(X_0)$	MFPT required by TF to reach the promoter starting from its CRM (X_0) via DNA-loop mediated propulsion mechanism. $T_B(X_0) = (L^2 - X_0^2)/2D_C + G/2D_C$, where the function G is defined as in **Eq. S2**.	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 $T_N(X) = (L^2 - X^2)/2D_C - X_0(L - X)/D_C$. Here $X = X_0$ is the reflecting boundary and $X = L$ is the absorbing boundary.	s
η	$\eta = [T_N(X_0)/T_B(X_0)]$	dimensionless
L_{opt}	Distance between CRM and promoter at which the ratio η is at a maximum. Approximately, $L_{\text{opt}} \sim 3X_0$	bp
ΔS_{loop}	Entropy change upon looping of DNA, $\Delta S_{\text{loop}} \simeq \ln(P_l/P_{all})$ 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	$k_r \simeq k_r^0 \exp(-\mu_{NS})$, 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}
L_S	$L_S \simeq \sqrt{2D_C/k_r}$, Average sliding length of TF on DNA before it dissociates where k_r is the dissociation rate constant.	bp

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions, with formulas, graphs, and mathematical tables* (Dover Publications, New York, 1965).