Mechanisms of enhanced or impaired DNA target selectivity driven by protein dimerization

Keyword: dwell time, dimensional reduction, transcription factors, binding kinetics, DNA binding, facilitated diffusion

Mankun Sang, Margaret E. Johnson*

TC Jenkins Department of Biophysics, Johns Hopkins University, 3400 N Charles St, Baltimore, MD 21218.

*Corresponding Author: margaret.johnson@jhu.edu ph: 410-516-2376

Abstract

Successful DNA transcription demands coordination between proteins that bind DNA while simultaneously binding to one another into dimers or higher-order complexes. Measurements that report on the lifetime or occupancy of an individual protein on DNA thus represent a convolution over the protein interactions with specific DNA, nonspecific DNA, or protein partners on DNA. For some DNA-binding proteins, dimerization is considered an essential step for stable DNA association, but here we show that protein dimerization can also reduce dwell times on specific DNA targets, enhance or impair occupancy on target sequences, and spatially redistribute proteins on DNA. We use mass-action kinetic models of pairwise association reactions between proteins and DNA (specific and nonspecific) and protein dimers, along with theory and spatial stochastic simulations to isolate the role of dimerization on observed dwell times and occupancy. For proteins binding a spatially localized cluster of targets, dimerization can drive up dwell time by 1000-fold and produce high selectivity for clustered over isolated targets. However, this effect can become negligible when proteins outnumber target sequences. In contrast, for isolated DNA targets, dimerization often reduces dwell times by sequestering proteins from their target sites, in some cases thus reducing overall occupancy. The ability of these proteins to bind DNA nonspecifically and diffuse in 1D to exploit dimensional reduction is a key determinant controlling degree of enhancement, despite the presence of nucleosome barriers to 1D diffusion. By comparison with ChIP-seq data, our model explains how the distribution of the GAF pioneer proteins throughout the genome is highly selective for clustered targets due to protein interactions. This model framework will help predict when even weak dimerization can redistribute and stabilize proteins on DNA as a necessary part of transcription.

Introduction

The transcription of DNA into RNA is essential for all living systems and relies on the concerted action of a variety of DNA-binding proteins[1-4]. DNA-binding proteins must recognize specific or target DNA sequences to perform their function in not only in transcription but also in DNA repair or replication, with protein dwell times[5, 6] and occupancies on DNA thus key metrics of proper function[7-9]. Particularly in eukaryotes, these multi-domain proteins often arrive at DNA sites as monomers or small complexes and are dynamically recruited via not only specific or nonspecific DNA interactions[7, 10-13] but via protein-protein interactions[14-18]. While some proteins (such as bZIP proteins[19, 20]) contain what is effectively only one-half of a DNA-binding domain, rendering them ineffective at targeting DNA without dimerization, proteins like the GAGA Factor (GAF[21]) or λ repressor[22] bind as monomers and can also form higher-order oligomers. Experimental measurements that report on the lifetime and occupancy of multi-domain proteins on DNA thus inevitably reflect the integrated effects of both DNA (specific and nonspecific) and protein interactions. Two-state models of a protein on DNA typically capture specific and nonspecific binding modes[23] [9, 23, 24], but without incorporating protein-protein interactions, they cannot predict responses to mutations in protein interacting domains[14]; even mutations to DNA binding domains may not ablate DNA localization if proteins recruit them to DNA. Here we construct a predictive framework to understand how dimerization will quantitatively enhance or impair DNA target occupancy, dwell time, and selectivity towards clustered targets by defining the simplest model of reversible protein dimerization that also incorporates DNA binding and diffusion in 3D and 1D. Using mass-action kinetics, theory, and reaction-diffusion simulations, we show how diverse outcomes emerge dependent on the ratio of proteins to targets, dimensional gains of 1D target searchers, and relative binding rates, driving negligible or dramatic responses to dimer formation.

A key feature of many DNA-binding proteins that we capture explicitly is their nonspecific association with the charged DNA backbone[25], facilitating 1D diffusion[23]. This sliding along DNA accelerates a protein search for a target sequence and is exploited by prokaryotic[26] [27-29] and eukaryotic proteins[30, 31], despite the latter's frequent interruption along the 1D DNA path by nucleosomes. An established consequence of nonspecific association is a longer lifetime and stronger affinity to DNA, as compared to binding isolated target sequences[32-34]. With two proteins localized to DNA, their dimerization is also transformed into a 1D search as long as one partner can slide. Much like association on 2D membranes relative to 3D, this dimensional reduction can drive up the collision probability[35], promoting significantly more stable and long-lived dimer and membrane-bound populations at equilibrium[36, 37]. Building from the seminal work of Berg and Von Hippel[23], our model here allows nonspecifically adhered proteins to encounter and bind DNA targets or other 1D-associated proteins using 1D rates. To support reversible binding and retain detailed balance in all pairwise reactions, we assume rates need not be diffusion-limited. This assumption that not all encounters lead to binding is consistent with evidence of proteins sliding over targets without binding[29], and is necessary for protein-protein association that is commonly barrier-limited. With both 1D and 3D association in our model, the ratio of a protein's local volume V to the accessible DNA length L between nucleosomes represents a critical factor controlling lifetimes and occupancy. The V/L ratio has units of area and demands a corresponding ratio of on-rates, $h^2 = k_a^{3D}/k_a^{1D}$, also with units of area. To estimate 1D rates for a binding pair given a known 3D rate, one can consider the proteins as switching from a searching mode (diffusing and bound nonspecifically) to a recognition mode [38-43], with protein re-orientations on the target[44, 45], and consider how the energy landscape of nonspecific binding determines the 1D diffusion rate[46, 47]. To maintain generality, we note there is also a corresponding characterization of 3D vs 2D rates and affinities using theory[48], simulation[49], and experiment[50, 51]. A key takeaway is that if the change in equilibrium affinities is primarily from entropic restrictions to configurations bound

to their substrate[48, 52], with minimal enthalpic changes, then the corresponding lengthscale h is at the molecular or nanometer scale. The specific value of h^2 will depend on the binding pair, but we treat this as a variable parameter. Collectively we define a dimensionality factor $\gamma = \frac{V}{Lh^2}$ that is generally greater than 1 and thus enhances the bound ensemble compared to purely 3D association[36].

Although dimerization between transcription factors is known to bring together DNA-binding domains that effectively switch-on DNA binding[53], our model encompasses a much broader range of mechanisms for dimerization to impact target selectivity. We establish regimes where dimerization impairs target binding via sequestration, and where dimerization can have no measurable impact on mean dwell times, despite altering the distribution of proteins on and off DNA. Critically, we quantify how dimerization (or higher-order oligomerization) can provide strong selectivity for clusters of targets that significantly exceeds what one naively expects with any increased density of target sites. Clustered targets in the genome are known to enhance transcription factor binding[54-56], with higher concentrations of targets present in genes known to be regulated by the corresponding TF [54]. Target clusters for the same TF (homotypic) are especially common in promoter and enhancer sequences on DNA [55, 56]. Naturally, proteins will bind more frequently when more sites are available, so even monomers will select for a cluster of *n* targets over a single target. However, our model quantifies how the bridging protein-protein interactions dramatically enhance dwell times on DNA, with this effect therefore lost if the targets are too widely spaced, highlighting the spatial dependence on the cluster organization[57, 58]. The statistics of the GAF protein throughout the drosophila genome quantified through ChIP-seq corroborates this prediction[14, 59], with high selectivity for clustered targets despite their low frequency. Dimerization can thus enhance inhomogeneities in localization throughout the DNA. We show the benefits of dimerization for target selection are reduced when proteins greatly outnumber targets, $[P] \gg [S]$, which is not

true for transcription factors like GAF[21] and TFIID[60] that recognize targets that are widespread throughout the genome ($[P] \leq [S]$), but is true for others like lac repressor[61] and Hsf1[62] that are highly gene specific ($[P] \gg [S]$). Nonetheless, we find the clustering of the DNA-binding domains afforded via dimerization will, in this regime, only enhance dwell time or occupancy on the target sequence. With our model, the impact of dimerization can be separated out from the stabilizing effects of specific and nonspecific DNA binding, providing a means to improve interpretation of single-particle tracking measurements and predict how both DNA-binding and protein-protein interaction domains control DNA target selection.

Models and Theories

Model Description

Our model contains two components, proteins, and DNA. Protein species P are diffusible and have three interaction sites, one for binding to another protein partner, one for binding to DNA nonspecifically, and one for binding to DNA specifically. DNA contains two species, specific sites S and nonspecific sites N. Specific sites are stationary, and do not diffuse. Nonspecific sites are diffusible, travelling along a 1D line that also contains the specific sites. Hence, we have replaced the continuous DNA sequence with discrete binding sites, with their number determined from the genome size and the size of proteins. This definition of nonspecific sites is needed for mass-action kinetic models that will allow proteins to localize to DNA and still retain the ability to slide (diffuse) along the backbone in 1D to encounter the immobile S sites. The nonspecific sites therefore diffuse at the rate of a protein diffusing while associate to the DNA backbone. In our spatial stochastic, particle-based reaction-diffusion simulations[63], we also treat nonspecific sites as discrete sites, although one can also use a model of adsorption to a 1D substrate.

The fundamental pairwise or bimolecular interactions between our species are protein-protein dimerization $P + P \rightleftharpoons PP$ with rates k_{on}^{P} and k_{off}^{P}, protein binding nonspecific DNA ($P + N \rightleftharpoons PN$) with rates k_{on}^{N} and k_{off}^{N}, and protein binding to specific DNA ($P + S \rightleftharpoons PS$) with rates k_{on}^{S} and k_{off}^{S} (Figure 1a). These are all 3D association rates. For these reversible interactions, these macroscopic rates account for the effect of diffusion on association/dissociation rates[64], and they are derived from the microscopic rates k_{a} and k_{b} that are independent of diffusion and used in the particle-based reaction-diffusion model[65], as we describe below. We note that the equilibrium state is independent of which rates are used, as $K_{\text{eq}} = \frac{k_{\text{on}}}{k_{\text{off}}} = \frac{k_{\text{a}}}{k_{\text{b}}}$. Because each protein P has three distinct and non-competing interaction sites, it can combine with P, S, and N to form multiple states, with some of these reactions occurring in 1D. In Figure 1b we illustrate the thermodynamic cycle connecting the unbound protein monomer P to the monomer bound specifically and nonspecifically to the DNA, PSN, which can be reached through two distinct pathways depending on whether S or N binds first. The distinction between the states PSN and PS thus emerges so that we can ensure detailed balance is maintained as our proteins can bind and unbind S from both 3D and 1D. From a molecular perspective, the $PS + N$ reaction is like a conformational change[44, 45], or unimolecular reaction, but because our N species is diffusible, it must be represented via a bimolecular reaction.

To capture the effect of dimensional reduction, or 1D localization, on our reactions, we must specify how the system size changes, $\frac{V}{L}$, where L is the continuous length of DNA and V is the volume surrounding this continuous DNA, and how the on-rates change from 3D to 1D. The rates can be defined most compactly via

$$k_a^{1D} = \frac{k_a^{3D}}{h^2},$$

where h^2 has units of area. The off rates can also be distinct in 3D vs 1D, but only by a scalar, $k_b^{1D} = ck_b^{3D}$. The equilibrium constants are thus related by $h'^2 = ch^2$, $K_{eq}^{1D} = K_{eq}^{3D}/h'^2$. For spatial simulations, these two aspects of dimensional reduction, the $3D$ vs $1D$ search spaces and the use of $3D$ vs $1D$ rates, are both explicitly captured. For rate equations, which are nonspatial, these two effects are combined into a dimensionality factor,

$$\gamma = \frac{V}{h^2 L} \tag{1}$$

For all 1D association reactions, the 3D on-rates must be scaled by γ to capture both the change in reactant concentrations (V/L) and the change in on-rates (see SI for detailed derivation). For simplicity, we assume that the off-rates are the same from 3D to 1D, $k_b^{1D} = k_b^{3D}$, and we use the same value of h^2 for all 1D reactions. We discuss the order-of-magnitude in SI, noting that both entropic and enthalpic restrictions to a DNA-bound protein can impact this squared length. An important relation is the strength of partitioning a 1D-localized protein (localized via a protein-protein interaction) to a nonspecific site N: $\gamma K_{\mathrm{eq}}^{PN,3D}[N] = \frac{V}{h^2 L} K_{\mathrm{eq}}^{PN,3D} \frac{n_{non} L}{l \cdot V} = \frac{K_{\mathrm{eq}}^{PN,1D} n_{\mathrm{non}}}{l}$, where l is the length of the linker DNA and n_{non} is the number of N sites on this linker DNA. n_{non}/l is the line density of nonspecific sites, e.g. 3 nm^{-1}, and $K_{\mathrm{eq}}^{PN,1D}$ is the affinity of the protein for DNA when it is held by a protein partner to a 1D search. Thus, a higher $\gamma K_{\mathrm{eq}}^{PN,3D}[N]$ implies either a higher line density of nonspecific sites, which we don't expect to change, or a stronger affinity for the DNA during a 1D search.

Since nucleosomes block protein sliding on linker DNA, the whole genome is separated to small segments. Each segment has a DNA sequence that may have different protein-binding environments with various distribution of targets. Therefore, we contrast three environments of DNA sequences (Figure 1c). 1) The DNA environment has no specific sites on the DNA segment. 2) The DNA+targ environment, the DNA segment includes two targets, but they are

distant from one another, such that a single protein dimer cannot bind both targets simultaneously. 3) The DNA+clusTarg environment has two clustered specific sites, therefore a single dimer can bind both sites simultaneously. All three models have the same linker length and the same volume-to-length ratio. Limited by diffusion, within a finite time period, the probability for a protein interacting with a far-away DNA segment is very low. Thus, for any species A, we use local concentrations for all models, namely $[A] = N_A / \left(\frac{V}{L} \cdot l\right)$, where N_A is the number of A in the corresponding model. For proteins and N sites, the local concentration are the same as the global concentration in nucleus, since $N_A \propto l$. However, in our model, N_S varies by segment, thus the local concentration $[S]$ does not equal the global concentration.

Model DNA contains 6 species: $[P], [PP], [N], [PN], [PPN], [PNPN]$, that participate in 6 reactions (see Methods). For model DNA+targ, in addition to the 6 species, it contains an additional 7 species: $[S]$, $[PS]$, $[PSN]$, $[PPSN]$, $[PPS]$, $[PNPS]$, $[PNPSN]$, where the last two species are distinct because $PPSN$ has one protein bound twice to DNA with the other bound to the protein only, and $PNPS$ has each protein bound to DNA through one site. These species have the 6 reactions from A, plus an additional 15 reactions that involve monomer addition reactions (e.g. $P + PSN \rightleftharpoons PPSN$) or two-two reactions (e.g. $PN + PS \rightleftharpoons PNPS$) or two-three reactions (e.g. $PSN + PN \rightleftharpoons PNPSN$ (see SI). For model DNA+clusTarg, since the targets are clustered, denoted S_2, the possible target-bound species are different from model DNA+targ. Besides the 6 species in model DNA, the other 10 species are: $[S_2]$, $[PS_2]$, $[PS_2N]$, $[PPS_2]$, $[PPS_2N]$, $[PNPS_2]$, $[PNPS_2N]$, $[PSPS]$, $[PSPSN]$, $[PSNPSN]$. The DNA+clusTarg model has 27 reactions (see SI).

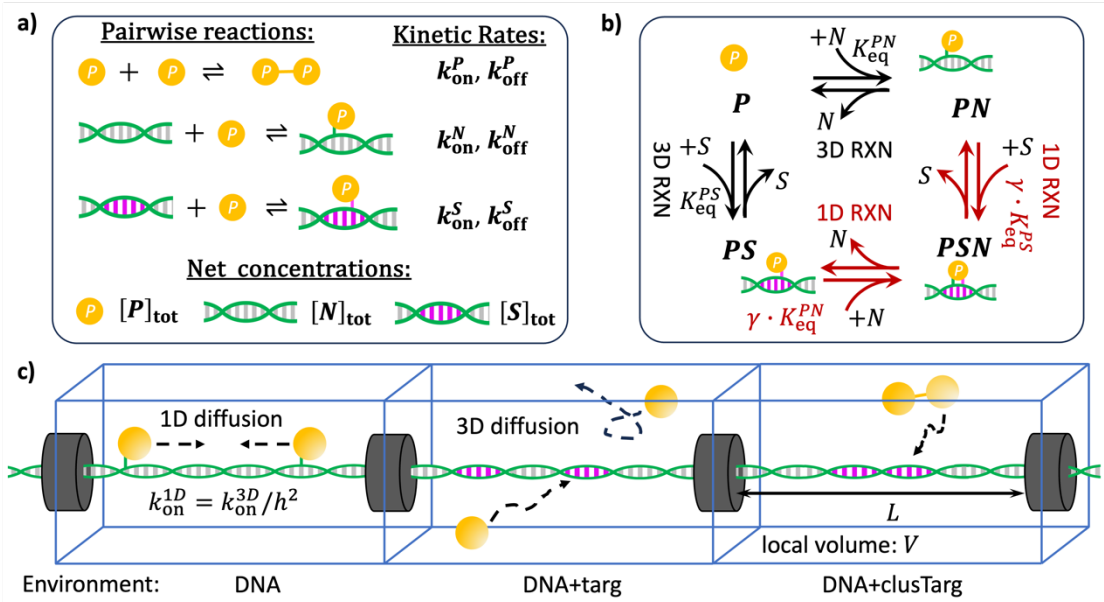

Figure 1 A schematic graph of our mass-action-based protein-DNA interaction model. a) The parameters and setup of our model. Here the yellow objects, gray bases, and magenta bases are for TFs, random DNA sequences, and target sequences, respectively. The parameters include concentrations, kinetic rates, and the dimensionality factor γ. b) The reaction network for monomers including 3D reactions (black arrows) and 1D reactions (red arrows). PSN labels target bound proteins that meanwhile electrostatically attracted to DNA. It enables transit between PN and PS without leaving DNA. c) 3 environments describe 3 distributions of targets (magenta bases) on linker DNA. The black cylinders and yellow spheres are nucleosomes and proteins, respectively. Proteins bound to random DNA sequences can perform 1D diffusion, but proteins bound to targets are fixed.

Theory

Two-state model for calculating the mean residence time.

All our models contain multiple protein-DNA bound states (e.g. PN, PS, and PSN) and unbound states (e.g. P and PP). To define convenient metrics that compare between models and capture what is typically measured in experiment, we can map these many individual states to a two-state model. In our primary two-state model, we compare proteins bound to DNA (including both specifically and nonspecifically bound proteins) vs proteins unbound from DNA. We are interested in the timescales and occupancies at an equilibrium steady state, where the flux in and out of each pairwise binding reaction is equivalent, e.g. $k^N_{on}[P]_{eq}[N]_{eq} = k^N_{off}[PN]_{eq}$. The sum over sets of pairwise reactions is thus also equal. To maintain generality, here we denote

proteins in unbound states as A_i and proteins in bound states as AB_j. Each A_i or AB_j state has proteins with a_i or a_j stoichiometry, respectively. The proteins in $\{A_i\}$ can transit to $\{AB_j\}$ by binding with B_k which can be DNA sites (e.g. S and N) or proteins on DNA (e.g. PN). Then, we can derive the mean residence time in the bound state of this two-state model (see methods),

$$\tau = \frac{\sum_{j=1}^{Nb} a_j [AB_j]}{\sum_{r=1}^{R} k_{\text{on}}^{r} a_r [A_r][B_r]} \tag{2}$$

A perhaps more familiar definition of the mean residence time comes from the survival time distribution $S(t)$, with $\tau = \int_0^\infty t \frac{-dS(t)}{dt} dt$. For the differential, the initial condition is that the protein has just bound to the DNA from solution, and it survives until it returns to solution, so $S(0) = 1$ and $S(t \to \infty) = 0$, and we are ignoring rebinding events because they represent new events. The survival time can be defined relative to the lifetime or exit-time distribution on the DNA, $S(t) = 1 - \int_0^t dt' p_{\text{exit}}(t')$, where $p_{\text{exit}}(t)$ is a normalized probability density function of exiting the DNA at exactly time t (or a lifetime on DNA of time t), given arrival at time 0. Given $S(t \to \infty) \to 0$ faster than t^{-1}, integration by parts also gives $\tau = \int_0^\infty S(t) dt$. This process is the same as calculating the mean first-passage time (MFPT). We show in the method that calculating τ via the survival distribution produces the same analytical result as Eq. (2). The advantage of using our two-state method via Eq. (2) is that the definitions are much simpler than solving first for the survival time distribution, which requires computing eigenvalues and eigenvectors. Furthermore, as the number of intermediate states grows, the eigenvalues required for calculating $S(t)$ cannot be defined analytically because of the Abel-Ruffini theorem.

Residence time of monomers with both nonspecific and specific binding

To provide a background for the role of specific binding before we consider dimerization, we consider the model illustrated by the thermodynamic cycle in Fig 1. A single protein monomer P is either free in solution, or bound to DNA via either N, S or both N and S simultaneously. The

residence time of the protein on the DNA without specific sites is given by $1/k_{\mathrm{off}}^{N}$ and that of protein on target only is $1/k_{\mathrm{off}}^{S}$. For proteins that can bind both specifically and nonspecifically, we use Eq. 2 to calculate the residence time based on the equilibrium steady state given by: $\tau_{\mathrm{monomer}}^{\mathrm{DNA+targ}} = \frac{[PN]_{\mathrm{eq}}+[PS]_{\mathrm{eq}}+[PSN]_{\mathrm{eq}}}{k_{\mathrm{on}}^{N}[P]_{\mathrm{eq}}[N]_{\mathrm{eq}}+k_{\mathrm{on}}^{S}[P]_{\mathrm{eq}}[S]_{\mathrm{eq}}}$. If we simplify this expression, we recover:

$$\tau_{\mathrm{monomer}}^{\mathrm{DNA+targ}} = \frac{K_{\mathrm{eq}}^{S}[S]_{\mathrm{eq}} + K_{\mathrm{eq}}^{N}[N]_{\mathrm{eq}} + \gamma K_{\mathrm{eq}}^{N}[N]_{\mathrm{eq}} K_{\mathrm{eq}}^{S}[S]_{\mathrm{eq}}}{k_{\mathrm{on}}^{S}[S]_{\mathrm{eq}} + k_{\mathrm{on}}^{N}[N]_{\mathrm{eq}}} \tag{3}$$

We can assume in some regimes that the flux into the specific state is higher than into the nonspecific state, or $k_{\mathrm{on}}^{N}[N]_{\mathrm{eq}} \ll k_{\mathrm{on}}^{S}[S]_{\mathrm{eq}}$, and the probability of being in specific vs nonspecific is also higher, $K_{\mathrm{eq}}^{N}[N]_{\mathrm{eq}} \ll K_{\mathrm{eq}}^{S}[S]_{\mathrm{eq}}$. Hence the residence time for DNA binding compared to purely specific binding is approximately given by,

$$\tau_{\mathrm{monomer}}^{\mathrm{DNA+targ}} \approx \tau_{S}\left(1 + \gamma K_{\mathrm{eq}}^{N}[N]_{\mathrm{eq}}\right) \tag{4}$$

Importantly, this result clearly emphasizes how the residence time for the protein bound to a specific motif will increase as the affinity or density of nonspecific sites accessible through 1D increases, as is seen experimentally with specific motifs flanked by longer nonspecific regions. It is also dependent on the dimensionality factor, as a larger γ better promotes remaining nonspecifically bound to DNA after dissociating from specific sites.

Protein bound ratio and target occupancy.

The ratio of proteins bound to DNA can be defined from the concentrations of free monomers and free dimers and varies between 0 (all proteins are in solution) and 1 (all proteins are localized to DNA). It is given by

$$\theta = \frac{\sum_{i=1}^{\mathrm{NP}} a_i[P_i]}{[P]_{\mathrm{tot}}} = 1 - \frac{[P]_{\mathrm{eq}} + 2K_{\mathrm{eq}}^{PP}[P]_{\mathrm{eq}}^{2}}{[P]_{\mathrm{tot}}}. \tag{5}$$

where NP sums over all protein species bound to the DNA. In our continuous model, the local protein copy number may be less than one, namely $N_P = [P] \cdot \left(\frac{V}{L}\right) \cdot l < 1$, since the linker DNA can be as short as $l \sim 20\ \mathrm{nm}$. In this case, the protein bound ratio represents the probability that a protein is bound to DNA in corresponding environment at equilibrium.

For target binding systems, we can also define the fraction of specific DNA targets occupied by proteins, or target occupancy , which also varies from 0 (all targets are free) to 1 (all targets are occupied) via:

$$\phi = 1 - \frac{[S]_{\mathrm{eq}}}{[S]_{\mathrm{tot}}}. \tag{6}$$

In a local environment, because proteins are assumed homogeneously distributed in the nucleus and targets are fixed in a local environment, it is possible that $[S]_{\mathrm{tot}} \gg [P]_{\mathrm{tot}}$ and $\phi \leq [P]_{\mathrm{tot}}/[S]_{\mathrm{tot}} \ll 1$.

Results

Driven by 1D reactions, dimerization enhances protein occupancy and usually lengthens residence time on nonspecific DNA

We first consider the simplest DNA environment where proteins can only bind DNA nonspecifically, as it nonetheless illustrates how dimensional reduction can enhance DNA localization properties. The occupancy of proteins on the DNA, which we measure via the protein bound ratio, θ^{DNA}, is given by $\frac{\chi_N}{1+\chi_N}$ for monomers, where $\chi_N = K_{\mathrm{eq}}^{PN}[N]_{\mathrm{tot}}$ measures the strength of partitioning from solution to nonspecific DNA. According to Eq. $(\mathrm{M12})$ and Eq. $(\mathrm{M13})$ (see Methods), the bound ratio will always increase due to dimerization (Fig 2a).

$$\frac{\theta^{DNA}}{\theta_{\mathrm{monomer}}^{\mathrm{DNA}}} = \frac{1+\chi_N}{\chi_N} \frac{\chi_N + 2K_{\mathrm{eq}}^{PP}[P]_{\mathrm{eq}}(2\chi_N + \gamma {\chi_N}^2)}{\chi_N + 1 + 2K_{\mathrm{eq}}^{PP}[P]_{\mathrm{eq}}(2\chi_N + \gamma {\chi_N}^2 + 1)} \geq 1, \tag{7}$$

where it is exactly 1 when $K_{\mathrm{eq}}^{PP} P_{eq} = 0$ (no dimers) and reaches maximal enhancement for irreversible dimers, $\frac{\theta_{\mathrm{dimer}}^{DNA}}{\theta_{\mathrm{monomer}}^{\mathrm{DNA}}} = \frac{(1+\chi_N)}{\chi_N} \frac{2\chi_N + \gamma {\chi_N}^2}{2\chi_N + \gamma {\chi_N}^2 + 1}$. The maximal enhancement is determined by the dimensionality factor γ, with larger values always driving larger enhancement since $\frac{\partial}{\partial \gamma} \frac{\theta^{DNA}}{\theta_{\mathrm{monomer}}^{\mathrm{DNA}}}$ is always positive. When $\gamma \leq 2$, the dimensionality effect is too small to enhance protein dimers binding DNA. For a fixed $\gamma > 2$, the enhancement is not monotonic with χ_N, meaning the biggest enhancement occurs at a specific value ${\chi_N}^*$ calculated by taking the derivative, $\frac{\partial}{\partial \chi_N} \frac{\theta_{\mathrm{dimer}}^{DNA}}{\theta_{\mathrm{monomer}}^{\mathrm{DNA}}} = 0$, to get:

$$ {\chi_N}^* = \frac{-1 + \sqrt{\gamma - 1}}{\gamma} \tag{8} $$

For stronger partitioning, $\chi_N \gg {\chi_N}^*$, monomers do not need 'help' to occupy DNA, and for values that are much weaker, dimerization cannot rescue very weak DNA binding (Fig 2a). The maximal occupancy can increase by up to 2-16 fold dependent on γ, with $\max\left(\frac{\theta_{\mathrm{dimer}}^{DNA}}{\theta_{\mathrm{monomer}}^{\mathrm{DNA}}}\right) = \frac{\gamma(\sqrt{\gamma-1}+2)-2}{2(\gamma-1)}$. Physically, higher γ facilitates PPN to form the second protein-DNA bond, which stabilizes dimers on DNA. Because nonspecific sites are more numerous than proteins, we assume they are never depleted, and we expect the bound ratio to always reach one with strong enough nonspecific binding and/or dimerization enhancement. For binding with specific targets, this will not always be the case. Conceptually, it is useful to consider what factors will change γ. Lowering h^2 (increasing the 1D affinity) will increase it without changing any other system property. Because of the dimensionality differences and what we assume is a fixed line density of nonspecific sites on DNA, changes to Volume that increase gamma will be compensated by changes to χ_N, because the nonspecific copies are fixed, their concentration will go down. For the same protein concentrations, this increases the copies that can localize to DNA, although it

will also lower χ_N, because the same number of nonspecific sites are in a larger volume, and copies of nonspecific sites does not change without a corresponding change in length.

For the residence time, we find that counterintuitively, dimerization can both enhance and reduce lifetimes relative to monomers. The mean residence time for monomers is $1/k_{\text{off}}^{N}$. By allowing proteins to dimerize, the mean residence time can be written as

$$\tau^{\text{DNA}} = \frac{1}{k_{\text{off}}^{N}}\alpha. \tag{9}$$

Where the scalar $\alpha > 1$ indicates dimerization has lengthened the residence time, and vice-versa. For proteins that form irreversible dimers, we can show dimerization always lengthens residence time, as following Eq 2:

$$\alpha_{\text{irr}} = \left(1 + \frac{\gamma K_{\text{eq}}^{PN}[N]_{eq}}{2}\right). \tag{10}$$

is always > 1. (Fig 2b) This maximal increase in residence time resulting from dimerization can be as large as 3 orders-of-magnitude, with the primary driving force being that a protein recruited to DNA by another protein can exploit 1D localization to bind DNA, with dissociation now requiring both proteins to unbind before rebinding occurs. In the more general case where protein-protein interactions are reversible, then,

$$\alpha = \frac{\frac{1}{K_{\text{eq}}^{PP}[P]_{\text{eq}}} + 4 + 2\gamma K_{\text{eq}}^{PN}}{\frac{1}{K_{\text{eq}}^{PP}[P]_{\text{eq}}} + 4 + 2\frac{k_{\text{off}}^{P}}{k_{\text{off}}^{N}}}. \tag{11}$$

Counterintuitively, here we see that we can have $\alpha < 1$, or a shortening of residence time due to dimerization. This occurs if $\gamma k_{\text{on}}^{N}[N]_0 < k_{\text{off}}^{P}$, that is, for fast k_{off}^{P} and low values of $\gamma k_{\text{on}}^{PN}[N]_0$ (Fig 2c). The origin of this is that proteins are still observed to be 'DNA bound' even when not bound explicitly if they are localized to DNA by another bound protein ($P + PN \rightarrow PPN$). The lifetimes of

these proteins that are only co-localized to the neighborhood of DNA can be controlled by the protein-protein lifetime if it is shorter than the time to form a second protein-DNA bond. Under this condition, stronger dimerization increases the fraction of these co-localized proteins that have a short residence time. Physiologically, this is less likely, as we expect nonspecific bonds to DNA are much shorter lived than protein-protein interactions, $k^P_{\text{off}} < k^N_{\text{off}}$, resulting in DNA-colocalized proteins more rapidly recruiting a second nonspecific site ($PPN + N \rightarrow NPPN$) rather than returning to solution ($PPN \rightarrow P + PN$). However, it establishes that by co-localizing proteins to an observed rather than explicit DNA-bound state, dimerization does not always ensure enhanced residence time on DNA.

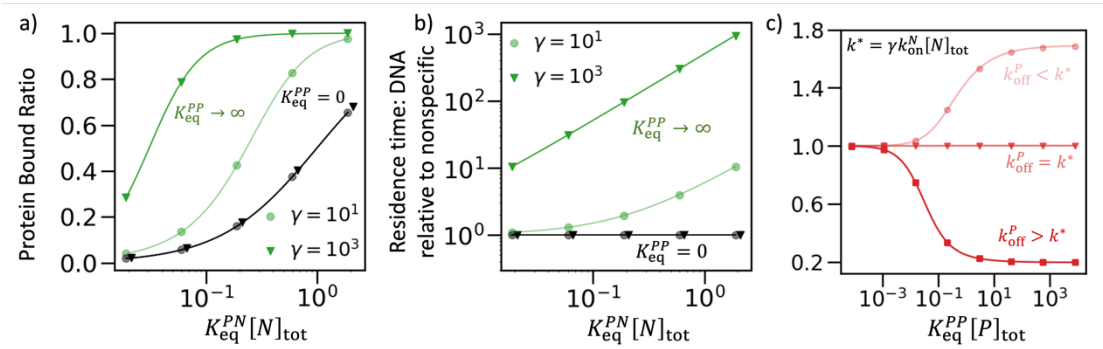

Figure 2 **Dimerization enhances occupancy but can both lengthen or shorten residence times on nonspecific DNA**. a) The occupancy/protein bound ratio depends on the strength of nonspecific partitioning, $K^{PN}_{eq}[N]_{tot}$ for both monomers (black curves) and dimers (green curves). Dimers have better occupancy than monomers, with largest improvement when $K^{PN}_{eq}[N]_{tot} = \frac{1}{\sqrt{\gamma}}$, and with larger $\gamma = 1000$ (triangles) improving occupancy relative to γ =10 (circles). b) The residence time of proteins on DNA relative to the monomer value, $1/k^N_{off}$, does not change for monomers, as expected (black curves). For strong dimers, the same trend is observed as for occupancy, with orders-of-magnitude increases in residence time possible for larger γ and more stable partitioning to DNA. c) The residence time can also decrease due to dimerization, where we define $k^* \equiv \gamma k^N_{on}[N]_{tot}$, and no change occurs when $k^P_{off} = k^*$ (triangles) and reduced times when $k^P_{off} > k^*$ (squares). The degree of enhancement or reduction depends on the strength of dimerization ($K^{PP}_{eq}[P]_{tot}$). Here $\gamma = 8.77$, $\gamma k^N_{on}[N]_{tot} = 1s^{-1}$, and $[P]_{tot} = 7.9\ \mu M$. From light to dark, k^P_{off} is $0.1\ s^{-1}$ (circle), $1\ s^{-1}$ (triangle), and $10\ s^{-1}$(square).

Separated targets enhance protein-DNA binding, but dimerization has two-sided influence on lifetime and occupancy.

With the addition of specific target sites to the DNA (Fig 1b), we assume target bindings are stable ($\chi_S = K_{\mathrm{eq}}^{PS}[S] > 1$), such that in models $\mathrm{DNA + targ}$ and model $\mathrm{DNA + clusTarg}$, the protein bound ratio for monomers to targets is already measureable (e.g. dimerization is not essential for DNA binding). From Eq. $(\mathrm{M18})$, $\theta_{\mathrm{monomer}}^{\mathrm{DNA+targ}} = \theta_{\mathrm{monomer}}^{\mathrm{DNA+cluTarg}} \geq \chi_S/(1+\chi_S)$, which leads to over 50% proteins bound to the DNA even when they are monomers (Figure 3a). With the addition of dimerization, the protein bound ratio is even higher and can be more than 85% even when nonspecific binding is weak. This makes χ_N always in the regime that protein-DNA binding is strong enough that the enhancement of dimerization is not a key factor for stabilizing more proteins on DNA. Different from model DNA, where $\gamma > 2$ is required for proteins to fully utilize the dimensionality effect, target binding helps to enhance protein-DNA binding. When K_{eq}^{PN} is weak, the flanking random sequence facilitates proteins to find and bind to targets, increasing occupancy on DNA.

Because targets are separated, a protein dimer can bind with only one target. In this case, dimerization can be harmful for target occupancy. Given by Eq. $(\mathrm{M19})$ and Eq. $(\mathrm{M25})$, the target occupancy of irreversible dimer is less than monomer, namely $\phi_{\mathrm{irr}}^{\mathrm{DNA+targ}} - \phi_{\mathrm{monomer}}^{\mathrm{DNA+targ}} < 0$, when target binding is strong enough to hold proteins bound with themselves (Figure 3b),

$$\chi_S > \chi^* = \frac{\chi_N(\gamma - 1)}{(1+\gamma\chi_N)^2}. \tag{12}$$

Instead, when $\chi_S < \chi^*$, a significant fraction of proteins is not target-bound which means proteins can take advantage of all the separated targets. When $\gamma \leq 1$, $\chi^* \leq 0 < \chi_S$, forming dimers is punished by the dimensionality effect which destabilizes the binding between proteins and targets surrounded by N sites. For $\gamma > 1$ such that dimensionality effect stabilizes proteins

on DNA, a greater χ^* results in a larger range of χ_S where dimerization enhances target occupancy. When χ_N and γ are very strong, dimerization is not essential for recruiting proteins on DNA. Now stronger dimerization reduces the target occupancy by restricting two proteins from binding two targets. Only with moderate $\gamma\chi_N$, dimerization enhances target occupancy. Now, dimerization enables target bound proteins stabilized by binding nonspecifically. Although forming dimers can restrict the number of proteins on targets, a dimer is more stable than a monomer and its lifetime on its target is longer.

Intuitively, introducing specific DNA binding targets will significantly increase the mean residence time for monomers, even by orders-of-magnitude (Figure 3c), as they have a longer lifetime bound to specific vs nonspecific sequences. This residence time (given by Eq. (M17), see Methods) also exploits the nonspecific binding, as shown here:

$$\tau_{\text{monomer}}^{\text{DNA+targ}} \approx \frac{1}{k_{\text{off}}^{S}} \cdot \frac{\chi_N + \chi_S(1+\gamma\chi_N)}{\chi_S\left(1+\frac{k_{\text{on}}^{N}[N]_{tot}}{k_{\text{on}}^{S}[S]_{tot}}\right)}. \tag{13}$$

This expression is the same for both DNA+targ and DNA+clusTarg since we do not allow dimerization here. Different from random DNA, since 1D diffusion helps proteins to find the target, stronger nonspecific binding results in stronger enhancement to $\tau_{\text{monomer}}^{\text{DNA+targ}}$ (Figure 3c). However, the flanking random sequence may also reduce $\tau_{\text{monomer}}^{\text{DNA+targ}}$ to be shorter than $1/k_{\text{off}}^{S}$, the residence time on an isolated target. This is because nonspecific binding can compete for proteins. To benefit from nonspecific binding, proteins must stay on random DNA long enough to find targets.

In $\text{DNA} + \text{targ}$, the mean residence time is very sensitive to protein dissociation (k_{off}^{P}). The mean residence time for irreversible dimer, $\tau_{\text{irr}}^{\text{DNA+targ}}$, can be much larger than that for strong but reversible dimers, $\tau_{\text{rev-strong}}^{\text{DNA+targ}}$, which is given by Eq 2. The difference $\tau_{\text{rev-strong}}^{\text{DNA+targ}} - \tau_{\text{irr}}^{\text{DNA+targ}}$ is not always positive. They are equal when $k_{\text{off}}^{P} = k^*$, where

$$k^* \approx \frac{\left(k_{\mathrm{off}}^N \chi_N + k_{\mathrm{off}}^S \chi_S\right)\left(2\gamma\chi_N\chi_S + \gamma\chi_N^2 + 2(\gamma\chi_N)^2\chi_S\right)}{(\chi_N + \chi_S + \gamma\chi_N\chi_S)^2}. \tag{14}$$

Assuming $\chi_S \gg \chi_N$, we have,

$$k^* \approx \frac{2\gamma\chi_N\left(k_{\mathrm{off}}^N \chi_N + k_{\mathrm{off}}^S \chi_S\right)}{\chi_S(1 + \gamma\chi_N)}. \tag{15}$$

For fast dimer dissociation ($k_{\mathrm{off}}^P > k^*$), dimerization reduces the mean residence time (Figure 3d) since a dimer can only bind with one target (e.g. PPS), but two monomers can bind two targets (e.g. PS and PS). With fast enough k_{off}^P, the dissociation of protein from DNA-bound dimers reduces the mean residence time. This phenomenon is clearer shown by Eq. $(\mathrm{S}3.5)$. We learn that $\partial k^*/\partial\gamma > 0$, which means it is harder to observe that dimerization harms residence time with strong dimensionality effect. Faster protein-DNA dissociation rates also increase k^*. When K_{eq}^{PS} and K_{eq}^{PN} are constants, faster k_{off}^S or k_{off}^N mean faster k_{on}^S or k_{on}^N, respectively. The faster these association rates are, the more easily the co-localized proteins bind with DNA. Noting that $\chi_S = K_{\mathrm{eq}}^{PS}[S]_{\mathrm{tot}}$, the fact that $\partial k^*/\partial\chi_S < 0$ means increasing the number of targets ($[S]_{\mathrm{tot}}$) can reduce the enhancement of dimerization to residence time. Proteins bound to these outnumbered targets will absorb other free proteins and prohibit them from finding other targets.

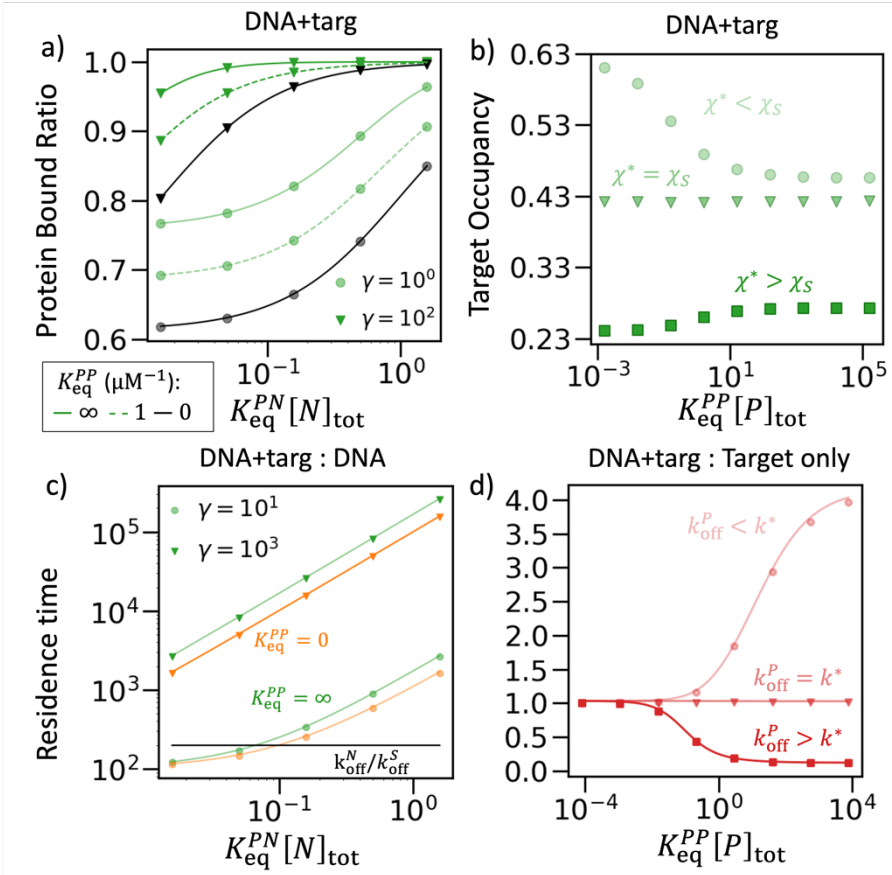

Figure 3 Separated targets stabilize proteins but dimers are not always preferred. a) shows nonspecific binding stabilizes protein-DNA binding for irreversible dimer (solid green), reversible dimer (dashed green), and monomer (black) ($[P]_{tot} = 1.6\ \mu M$ and $K_{eq}^{PS} = 10^4\ M^{-1}$). b) shows that with weak target binding ($K_{eq}^{PS} = 10^3\ M^{-1}$), $\phi_S^{DNA+targ}$ responds to K_{eq}^{PP} differently controlled by χ^*. To avoid nonphysiologically fractional binding, a high protein concentration is used and the assumption $[S]_{tot} = [S]_{eq}$ fails and the equilibrium is solved numerically with ODEs. ($\chi_S = 0.199$, $[P]_{tot} = [S]_{tot} = 0.16\ mM$). c) shows both $\tau_{monomer}^{DNA+targ}$ / $\tau_{monomer}^{DNA}$ (orange) and $\tau_{dimer}^{DNA+targ}$ /τ_{dimer}^{DNA} (green) benefits from stronger K_{eq}^{PN}. Irreversible dimers always have a longer lifetime than monomers. With very weak dimensionality enhancement and nonspecific binding, $\tau^{DNA+targ}/\tau^{DNA}$ can be smaller than $\frac{1}{k_{off}^S}/\frac{1}{k_{off}^N}$ (black). ($[P]_{tot} = 7.9\ \mu M$ and $K_{eq}^{PS} = 10^3 K_{eq}^{PN}$). d) Intuitively, $\tau^{DNA+targ} \cdot k_{off}^S$ increases as K_{eq}^{PP} gets stronger (square dots). However, when k_{off}^P is larger than $k^* = 0.1\ s^{-1}$, dimerization harms $\tau^{DNA+targ}$ (circle dots), and when $k_{off}^P = k^*$, the lifetime does not change with dimerization (triangle dots). ($[P]_{tot} = 7.9\ \mu M$, $K_{eq}^{PN} = 200\ M^{-1}$, $K_{eq}^{PS} = 2 \times 10^4\ M^{-1}$, and $\gamma = 32$).

Clustered targets cooperate with dimerization to offer dramatic increases in protein-DNA stability.

When targets are clustered, a dimer can bind both targets at the same time. In this case, a dimer that already binds one target (like PPS_2) will easily find and bind the second target ($PPS_2 \rightarrow PSPS$). This kind of fully-target-bound dimers are very stable since two proteins are locked together at the target place and their dissociation requires breaking two bonds (P- S and P-P) in a short time ($1/k_{\mathrm{on}}^{S}C_0$ or $1/k_{\mathrm{on}}^{P}C_0$). As a result, dimerization significantly increases the fraction of DNA-bound proteins. In model $\mathrm{DNA+targ}$, even irreversible dimers need help from dimensionality enhancement and nonspecific binding to reach $\theta = 1$ (Figure 3-a). However, in model $\mathrm{DNA+clusTarg}$, irreversible dimers always have $\theta = 1$ (Figure 4-a). With clustered targets, the protein bound ratio becomes sensitive to weak $K_{\mathrm{eq}}^{PP} \sim 3\ \mathrm{mM}^{-1}$, where the protein bound ratio increases halfway from the monomer limit to 100%. On the contrary, when targets are separated, the halfway increase happens when $K_{\mathrm{eq}}^{PP} \sim 1\ \mu\mathrm{M}^{-1}$.

The target occupancy and residence time always increase within physiological parameter range (Figure 4-b,c). From Eq. (M26) and Eq. (M19), we know that $\phi_{\mathrm{irr}}^{\mathrm{DNA+clusTarg}} - \phi_{\mathrm{monomer}}^{\mathrm{DNA+clusTarg}}$ is always positive as long as protein dimer's affinity to targets is higher than to random sequences. From Eq. (M17) and Eq. (M20), we can calculate the critical rate k^* like Eq. (14),

$$k^* \approx \frac{\left(k_{\mathrm{off}}^{N}\chi_N + k_{\mathrm{off}}^{S}\chi_S\right)}{(\chi_N + \chi_S + \gamma\chi_N\chi_S)^2}\left\{(2\gamma\chi_N\chi_S + \gamma\chi_N^2 + 2(\gamma\chi_N)^2\chi_S) + C_0K_{\mathrm{eq}}^{PS}\chi_S(1+\gamma\chi_N)^2\right\}. \qquad (16)$$

Assuming $\chi_S \gg \chi_N$ and $C_0K_{\mathrm{eq}}^{PS} \gg 1$, we have,

$$k^* \approx C_0\left(\frac{[N]_{\mathrm{eq}}}{[S]_{\mathrm{eq}}}k_{\mathrm{on}}^{N} + k_{\mathrm{on}}^{S}\right). \qquad (16)$$

Since $C_0 = 1\ \mathrm{M}$ and physiologically $k_{\mathrm{on}}^{S} > 10^3\ \mathrm{M}^{-1}\cdot\mathrm{s}^{-1}$, one has $k^* > C_0k_{\mathrm{on}}^{S} > 10^3\ \mathrm{s}^{-1}$. Physiologically, it is impossible to have such a fast k_{off}^{P} for the co-localized protein to prefer

dissociation but not binding DNA. When nonspecific binding is considered, the association rate to DNA for the co-localized protein is faster and k^* will be faster. Therefore, dimerization always enhances the residence time when targets are clustered.

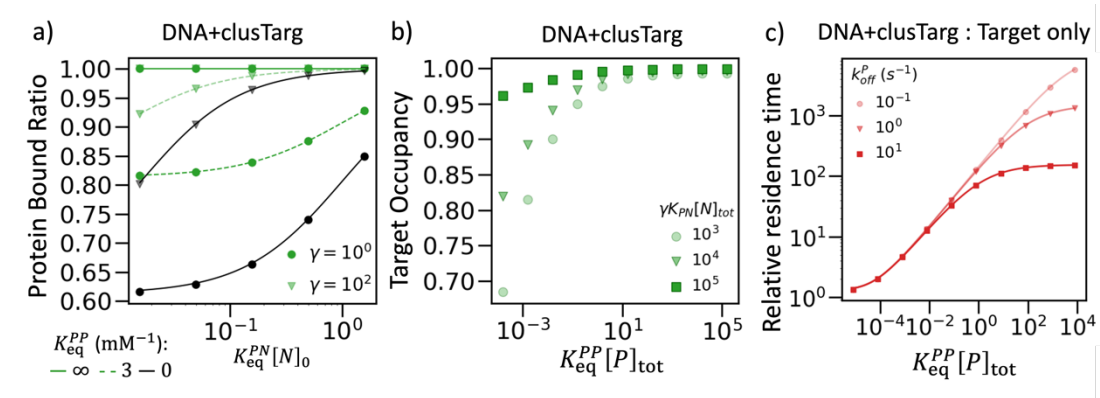

Figure 4 Cooperativity between dimers and clustered targets significantly stabilize protein-DNA binding. a) $\tau_{monomer}^{DNA+clustTarg} \cdot k_{off}^{S}$ increases as dimerization gets stronger (here $K_{eq}^{PN} = 20$, $K_{eq}^{PS} = 20000$, $\gamma = 100$, and $[P]_{tot} = 7.9\ \mu M$). b) $\phi_S^{DNA+clutTarg}$ increases as dimerization gets stronger. (here $K_{eq}^{PS} = 10000$, $\gamma = 50$, and $[P]_{tot} = 160\ \mu M$). c) $\theta_P^{DNA+clutTarg}$ increases as nonspecific binding gets stronger (here $K_{eq}^{PS} = 10000$, and $[P]_{tot} = 1.6\ \mu M$).

Higher local protein concentration enhances target occupancy but reduces the mean residence time when excess.

When proteins are bound to DNA, the number of free proteins in solution decreases, causing a depletion of $[P]_{\mathrm{eq}}$ in the local environment. Therefore, proteins in neighbor environments can diffuse to this depleted region and the total local protein concentration ($[P]_{\mathrm{tot}}$) will increase. This effect is more important in environment $\mathrm{DNA + clusTarg}$ since the high protein bound ratio. Besides, external stimuli or other processes can also cause the change of DNA-binding protein concentrations. By our numerical solution, a higher $[P]_{\mathrm{tot}}$ always increases the target occupancy in environment $\mathrm{DNA + clusTarg}$ (Figure 5-a). With the help of dimerization ($K_{\mathrm{eq}}^{PP} = 10\ \mathrm{M}^{-1}$), even for weak DNA-binding affinities ($K_{\mathrm{eq}}^{PS} = 10^3\ \mathrm{M}^{-1}$ and $K_{\mathrm{eq}}^{PN} = 1\ \mathrm{M}^{-1}$), 10-fold increase of $[P]_{\mathrm{tot}}$ can result in more than 10-fold increase of $\phi^{\mathrm{DNA+clusTarg}}$. Under this case,

monomers only reach $\phi^{\mathrm{DNA+clusTarg}} \approx 70\%$ with $[P]_{\mathrm{tot}} = 10[S]_{\mathrm{tot}}$, however, dimerization allows full occupancy with much smaller $[P]_{\mathrm{tot}}$.

The mean residence time can benefit from higher protein concentration but will decrease with excess proteins (Figure 5-b). When some proteins occupied the targets by either P-S binding or P-P binding, the excess proteins can only form short-lived bindings which decrease $\tau^{\mathrm{DNA+clusTarg}}$. For monomers, any more proteins are in excess since they cannot be stabilized by binding to target-bound proteins and co-localize with targets. However, with dimerization, $\tau^{\mathrm{DNA+clusTarg}}$ can benefit from higher $[P]_{\mathrm{tot}}$ since one target-bound protein can stabilize one another protein. This mechanism is effective until $[P]_{\mathrm{tot}}$ is high enough that the number of dimers is enough to occupy all targets, or $\phi^{\mathrm{DNA+clusTarg}} \approx 1$. This means that, only with the help of dimerization, excess proteins will not be trapped in the region where $\phi^{\mathrm{DNA+clusTarg}}$ is already high enough.

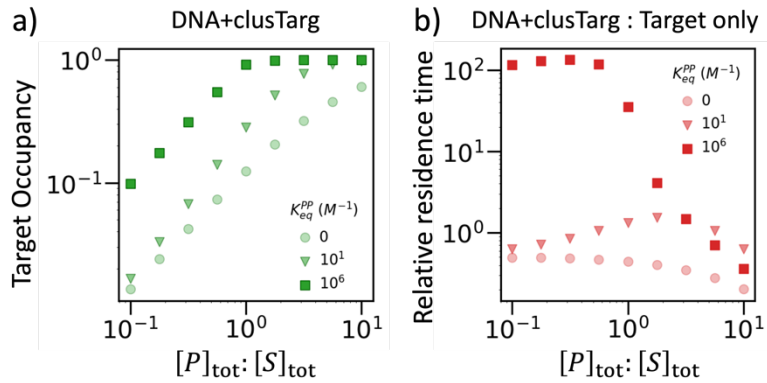

Figure 5 Numerical solution for the effect of Higher local protein concentration on environment $DNA + clusTarg$. a) Higher protein concentration cooperating with dimerization always result in higher target occupancy. b) More protein yields more dimers, which increases the mean residence time. However, excess proteins reduce the lifetime. Stronger dimerization can increase the lifetime when concentration is low or reduce the lifetime by causing excess.

Nonspecific binding and dimensionality effect controls how dimerization affects lifetime and occupancy.

Using a typical set of parameters, we examine the impact of dimerization on residence time (τ) and DNA-bound protein populations (ϕ and θ) in a generalized framework. When targets are clustered, the strong cooperativity between dimerization and specific binding ensures that τ and ϕ always benefit from dimerization under physiological conditions (Figure 6-a). However, in the absence of specific binding, increasing dimerization strength can lead to a decrease in τ. For weak nonspecific binding and weak dimensionality enhancement ($\gamma\chi_N \approx 1$), a dissociation rate of $k_{\mathrm{off}}^{P} > 0.5\ \mathrm{s}^{-1}$ is too fast, causing the dissociation of DNA-colocalized proteins becomes the dominant dissociation event (see SI). Both more stable nonspecific binding and stronger γ increase k^*. However, when targets are isolated, $k_{\mathrm{off}}^{P} > 0.1\ \mathrm{s}^{-1}$ will always result in a shorter τ with stronger dimerization (Figure 6-a).

Notably, kinetic effects do not always directly translate to occupancy behavior (ϕ) (Figure 6-b) and never affects protein bound ratios (θ) (Figure 6-c). Although τ^{DNA} and $\tau^{\mathrm{DNA+targ}}$ are easily to be impeded for $\gamma\chi_N \approx 1$, the target occupancy is still likely to benefit from dimerization. With clustered targets, ϕ always increase. With isolated targets, ϕ can be reduced when the partitioning to targets is too strong ($\chi_S > \chi^*$). This happens in two cases. Firstly, the concentration of isolated targets is very high, either because of a shorter flanking region, or because of too many targets crowded in the linker DNA. Secondly, K_{eq}^{PS} is too strong, which locks proteins on the first target they find. However, $\chi_S < \chi^*$ is also highly possible under this condition when free proteins are associating and dissociating with the target-bound protein very quickly (τ is not increased). These proteins stabilize the target-bound protein on DNA but remains free to search for other targets. In a living cell, most of the motifs are isolated and only a small fraction of motifs is clustered [66]. With a moderate $\gamma\chi_N$, these isolated motifs can

benefit from dimerization and proteins are still able to freely search for clustered targets that may need long-term promoting.

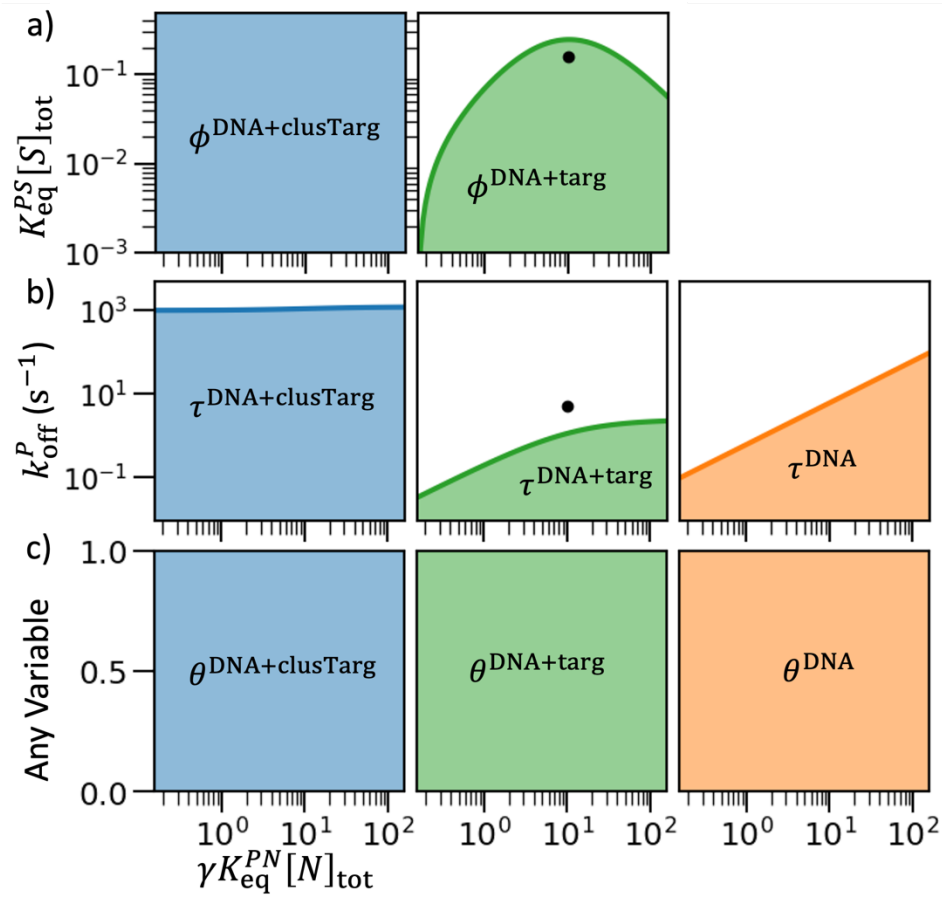

Figure 6 How lifetime and population change for reversible dimers compared with monomers. The solid lines in all plots represent no change between dimers and monomers, and the colored region means dimerization leads to enhancement. a) Target occupancy ϕ. Dimers only occupy more targets (green) in environment $DNA + targ$ with moderate nonspecific binding and dimensionality enhancement. With clustered targets (blue), stronger dimerization always increases target occupancy. The black dots in a) and b) label the same condition. b) Dwell time τ. In environment $DNA + clusTarg$ (blue), dimerization always increases the residence time of proteins on DNA under physiological conditions. For random sequences, dimerization may

reduce the residence time with fast k_{off}^{P} (orange). In environment $DNA + targ$ (green), it is more likely to observe dimerization reducing the residence time. c) All DNA occupancy θ. The protein bound ratio to both specific and nonspecific DNA always benefits from dimerization.

Dimerization helps protein to identify clustered target dimers on DNA.

Although our theory ignores spatial distribution and separates the whole chromatin to disconnected segments separated by nucleosomes, it can be altered to calculate the spatial selectivity at equilibrium (see method $(\mathrm{M27})$). Since proteins can freely diffuse in solution among adjacent segments, the side with higher θ_P has fewer proteins in solution and thus attracts proteins from the other side. Suppose two adjacent segments have clustered targets on the left and have separated targets on the right (Figure 7-a), $\theta_P^{\mathrm{clusTarg}} > \theta_P^{\mathrm{Targ}}$ suggests that more proteins will locate in the left side (Figure 7-b), which leads to preference for clustered targets. To study this, we define the selectivity ≥ 0 for one environment to the other as:

$$\text{selectivity for A to B} = \frac{\text{DNA bound proteins on A}}{\text{DNA bound proteins on B}}$$

Once proteins are attracted to the left side, the local concentration increases and further promotes θ_P by forming more dimers, which forms a positive feedback loop. When the total number of proteins, P_{tot}, is less than the number of targets on one side, $S_{\mathrm{tot}}/2$, this feedback ensures to fully utilize dimerization to enhance selectivity. Therefore, when $P_{\mathrm{tot}} < S_{\mathrm{tot}}/2$, the selectivity is rarely influenced by the number of proteins in physical range ($P_{\mathrm{tot}} > 1$). Instead, dimerization contributes a lot to the selectivity. With $K_{\mathrm{eq}}^{PP} \sim \mu\mathrm{M}^{-1}$, almost all proteins select the segment with clustered targets (Figure 7-b).

When $P_{\mathrm{tot}} > S_{\mathrm{tot}}/2$, not all proteins are able to bind to targets on the same segment. In this case, the rest proteins will be attracted by the targets on the other segment, which reduces the selectivity. Suppose $P_{\mathrm{tot}} = 3$ and $S_{\mathrm{tot}}/2 = 2$, when all proteins are bound to targets, the selectivity is just 1/3. However, with finite DNA affinity, dimerization first increase the selectivity and then reduce the selectivity. We have learnt that $K_{\mathrm{eq}}^{PP}[P]_{\mathrm{tot}} \sim 10^{-3}$ (Figure 4) and $K_{\mathrm{eq}}^{PP}[P]_{\mathrm{tot}} \sim 10^{1}$ (Figure 3) are the turning points for $\theta_P^{\mathrm{DNA+clusTarg}}$ and $\theta_P^{\mathrm{DNA+targ}}$, respectively. Therefore, when K_{eq}^{PP} increases from 0 to about $1\ \mathrm{mM}^{-1}$, the increase of $\theta_P^{\mathrm{DNA+clusTarg}}$ is fater than the increase of $\theta_P^{\mathrm{DNA+targ}}$ and selectivity increases. But for much stronger K_{eq}^{PP}, the increase of $\theta_P^{\mathrm{DNA+targ}}$ is faster, which reduces the selectivity.

We also use spatiotemporal simulations to validate our theory (Figure 7-c). Without dimerization, there is no selectivity between two segments. In this condition, many proteins (about half) are in solution. When $K_{\mathrm{eq}}^{PP} = 1\ \mu\mathrm{M}^{-1}$, there are no proteins in solution and proteins show selectivity=1 for clustered targets. The spatiotemporal simulations by NERDSS show a same trend but have different selectivity (Figure 7-c). This is the difference between spatiotemporal models and continuous models accepting fractional protein copy numbers. For cases with weak K_{eq}^{PP}, a continuous model allows forming dimers even the number of proteins is less than 2, which is rejected in spatiotemporal simulations. By capturing spatial heterogeneity, the NERDSS simulation highlights the switch-like behavior that the two-magnitude increasing of K_{eq}^{PP} from $10^{1}\ \mathrm{M}^{-1}$ to $10^{3}\ \mathrm{M}^{-1}$ significantly promotes selectivity.

For protein dimers, although the clustering of more than 2 targets attracts more proteins as the size of target clustering gets larger, the cooperativity among proteins and targets does not increase. The cooperativity is represented by the fraction of proteins bound per target,

$$\text{protein per target} = \frac{\text{protein bound to environment}}{\text{number of targets clustered}}.$$

For dimers, 0% proteins are bound to the isolated target, and for clustered targets, 10%~11% proteins are bound per target (Figure 7-d). Compared with monomers that results in 2%~3% proteins are bound per target for all environments, the increase of protein per target from 0% to 10% indicates that the clustering enables cooperative binding of dimers. However, for 2~4 targets, they have the protein per target, meaning that the cooperativity does not increase. To better utilize clustered targets in a larger size, proteins need to form larger oligomers. When proteins can form linear tetramers (see SI for structure), the protein per target increases as more targets clustered together, which indicates that the cooperativity increases.

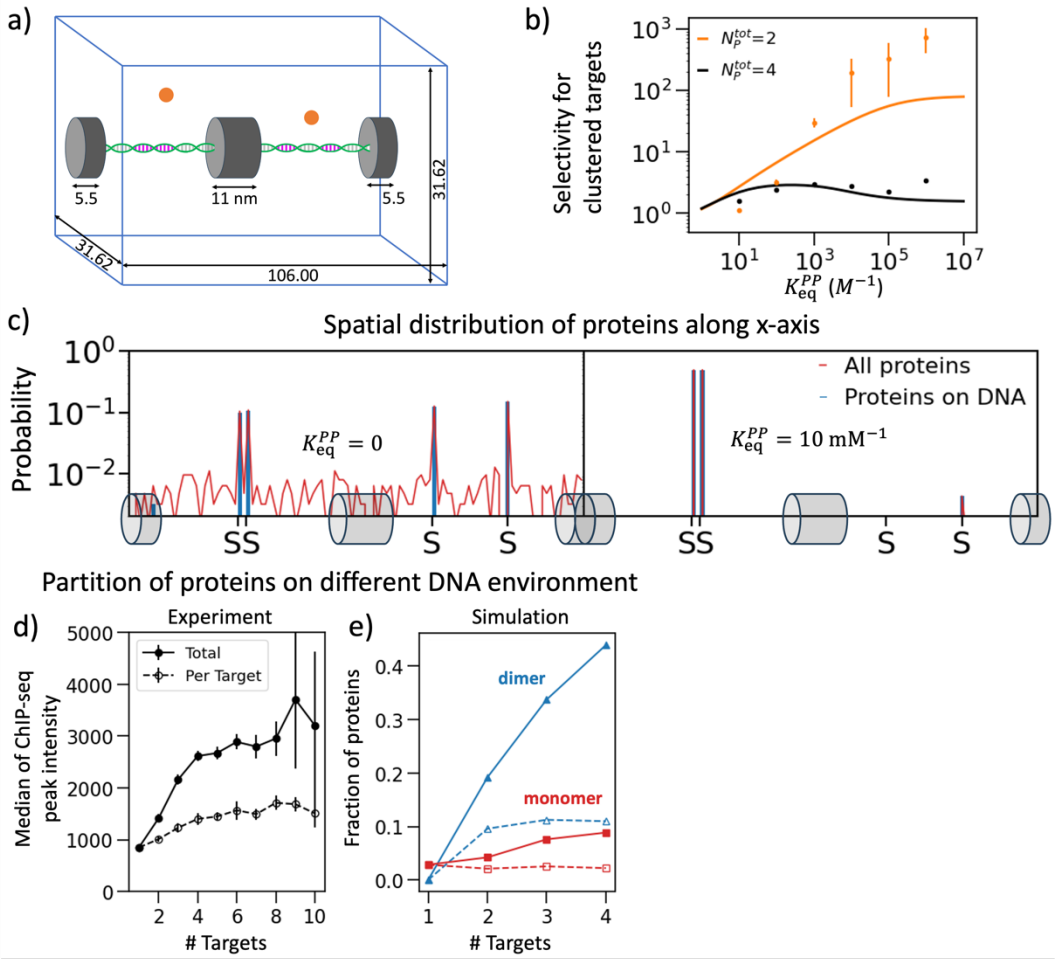

Figure 7 Selectivity between adjacent DNA segments with different target distribution. a) shows that two targets are clustered together on the left side and two

targets are separated on the right side. Two proteins are searching for targets. b) shows that DNA-bound proteins usually prefer clustered targets. The scatters are results from spatial simulations by NERDSS with totally 2 proteins ($N_P^{tot} = 2$). The results are calculated using the ODE model with $K_{eq}^{PN} = 10\ M^{-1}$ and $K_{eq}^{PS} = 10\ mM^{-1}$. c) shows the spatial distribution of proteins with different dimerization strength. The y axis is the probability that proteins locate in the region (x-0.5, x+0.5). Red curves are proteins both in solution and on DNA and the blue histograms are proteins on DNA. They are both normalized by the total number of proteins. d) Our simulation of 4 DNA segments with different number of targets show that dimerization helps proteins select DNA with a cluster of more than 2 targets. Dimeric proteins (blue) selectively bind with clustered targets and almost no proteins bind to the isolated target. Even though more monomers (red) are found on the DNA with 4 clustered targets, the proteins bound to each target are the same. (Diffusion constants of protein monomers are $D_{1D} = 0.6\ nm^2/\mu s$ and $D_{3D} = 1.5\ nm^2/\mu s$. $k_{off}^{P} = 10\ s^{-1}$, $k_{off}^{S} = 6.05\ s^{-1}$, and $k_{off}^{N} = 1210\ s^{-1}$.) e) The ChIP-seq peaks of GAGA factors [59] with more targets tend to have a higher intensity. However, the per-target intensity reaches a plateau when around 6 targets clustered together. The standard errors are generated by 100 bootstraps f) Less than 3% of the GAGA factor (GAF) targets between two consecutive nucleosome dyads exist as clusters larger than 6 and over 70% of the GAF targets are isolated targets.

Simulation shows similar plateau at the size of protein multimer as observed in experiment.

Through spatial simulations, we challenged the dimer-forming proteins to now select between targets organized as a single target, pair of targets, three-in-a-row, or four-in-a-row, as more representative of what faces a protein like GAF in the nucleus (Fig 7d). We inevitably expect more proteins at the 4-site cluster compared to the 1-site cluster, and indeed if the proteins do

not dimerize, their occupancy is increased by exactly 4-fold, or n-fold at each of the n-sites clusters. This means that the proteins-per-site is independent of the cluster sizes (Fig 7f). When we allow dimerization, we now see no proteins binding to the 1-site cluster, and an increasing fraction from 2 to 4 sites. Normalizing by *n,* we note a plateau in enhanced occupancy for $n \geq 2$. In other words, a dimer is maximally selective for a cluster of size 2, and larger clusters only increase recruited via their higher density. We can therefore show that if we allow the proteins to form a tetrameric oligomer, its selectivity increases up to a cluster of 4-sites, and then similarly plateaus (SI).

This phenomenon is also observed in *in-vivo* experiments. The GAF protein of *D. melanogaster* forms higher-order oligomers up to 6-mers (or possibly larger) [14, 30, 67]. From ChIP-seq, the partitioning of GAF throughout the drosophila genome shows increasing intensity for proteins observed at clusters with increasing *n* sites (Fig 7), indicating higher occupancy at those sites [59]. Normalizing the intensity by the number of sites per clusters (the number of non-overlapping targets found within peaks), we retain an increase in occupancy for larger clusters, with a plateau that is approximately at 6 targets (Figure 7e). We note that the selectivity in the *in-silico* experiment appears to be more enhanced compared to GAF. However, a key difference is that all of the simulated clusters (n=1,2,3 or 4 sites per DNA segment) were present at equal amounts, to simplify the simulation set-up. In contrast, for GAF targets, cluster abundances are highly non-uniform; the majority of target sequences appear alone (n=1), and more than 90% are present at n<3 [66]. Less than 3% of GAF targets are clustered in a size larger than 6. Hence the GAF shows remarkable selectivity for these large but infrequent clusters over single-target sites.

DISCUSSION

Our results show that dimerization between DNA-binding proteins will only universally enhance DNA target occupancy and selectivity when the binding sites are clustered to the same degree (as dimers) to the protein dimer. A dimer that binds an isolated single target will enhance its own residence time on DNA via nonspecific interactions, but the population of proteins can overall suffer in target binding via sequestering proteins away from an open target. In this regime, where proteins are comparable to or less abundant than their targets, we can thus measure the mean dwell time of the population is reduced, while the target occupancy is increased. This surprising result occurs because dimers can increase flux onto target DNA (higher effective on-rate), despite shortening the mean lifetime of the population at targets (higher effective off-rate). Our model further illustrates how even without targets present, the residence time of dimers is distinguishable from that of monomers, producing a different distribution of survival times.

This model represents an important advance over previous quantitative models for predicting DNA dwell time and occupancy by explicitly adding in dimerization and thus significantly expanding the state-space (up to 15 states) beyond the monomer bound to specific or nonspecific DNA (3-4 states). Nonetheless, it fails to capture much of the additional complexity of the in vivo environment, including the spatial organization of the genome (rather than linear segments), the mechanical contributions of DNA to localization, the condensation of proteins into liquid-like droplets, and perhaps most critically, the interaction of protein dimers or oligomers with other proteins, including RNA polymerase or the transcription initiation complex. Many other proteins (or nucleic acids) can thus compete to bind each protein or the same DNA sequence, potentially reducing dwell times or occupancies below our predictions. Additionally, we treated all nonspecific association as equally strong, while it has been shown that repeat motifs in DNA often lead to more stable nonspecific association. Lastly, we note that the actions

of DNA polymerase and various enzymes in the nucleus drive the system out-of-equilibrium, whereas we assumed an equilibrium steady-state.

This model will offer a foundation to build additional complexity into the dynamics of DNA-binding proteins during transcription. More immediately, the formulas can be used to predict how perturbations to individual domains in these proteins, affecting DNA-binding or protein-protein association, would reflect on dwell times, occupancy, or spatial distributions on heterogeneous DNA.

Acknowledgements:

MEJ gratefully acknowledges support from an NIH U01DK127432 Award. We thank Profs Carl Wu, Taekjip Ha, and Yaojun Zhang, along with Dr. Xiaona Tang for helpful discussions on the project.

References:

1. Lu, F. and T. Lionnet, *Transcription Factor Dynamics*. Cold Spring Harbor Perspectives in Biology, 2021. **13**(11): p. a040949.
2. Takahashi, K. and S. Yamanaka, *Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors*. Cell, 2006. **126**(4): p. 663-76.
3. Iwafuchi-Doi, M. and K.S. Zaret, *Cell fate control by pioneer transcription factors*. Development, 2016. **143**(11): p. 1833-1837.
4. Magnani, L., J. Eeckhoute, and M. Lupien, *Pioneer factors: directing transcriptional regulators within the chromatin environment*. Trends in Genetics, 2011. **27**(11): p. 465-474.
5. Darzacq, X., et al., *In vivo dynamics of RNA polymerase II transcription*. Nat Struct Mol Biol, 2007. **14**(9): p. 796-806.
6. Chen, J., et al., *Single-molecule dynamics of enhanceosome assembly in embryonic stem cells*. Cell, 2014. **156**(6): p. 1274-1285.
7. Mueller, F., et al., *Quantifying transcription factor kinetics: At work or at play?* Critical Reviews in Biochemistry and Molecular Biology, 2013. **48**(5): p. 492-514.
8. Perlmann, T., P. Eriksson, and O. Wrange, *Quantitative analysis of the glucocorticoid receptor-DNA interaction at the mouse mammary tumor virus glucocorticoid response element*. Journal of Biological Chemistry, 1990. **265**(28): p. 17222-17229.
9. Berg, O.G. and C. Blomberg, *Association kinetics with coupled diffusional flows*. Biophysical Chemistry, 1976. **4**(4): p. 367-381.

10. DeHaseth, P.L., et al., *Nonspecific interactions of Escherichia coli RNA polymerase with native and denatured DNA: differences in the binding behavior of core and holoenzyme*. Biochemistry, 1978. **17**(9): p. 1612-1622.
11. Leven, I. and Y. Levy, *Quantifying the two-state facilitated diffusion model of protein–DNA interactions*. Nucleic Acids Research, 2019. **47**(11): p. 5530-5538.
12. Suter, D.M., *Transcription Factors and DNA Play Hide and Seek*. Trends Cell Biol, 2020. **30**(6): p. 491-500.
13. Chen, X., M.-Y. Tsai, and P.G. Wolynes, *The role of charge density coupled DNA bending in transcription factor sequence binding specificity: a generic mechanism for indirect readout*. Journal of the American Chemical Society, 2022. **144**(4): p. 1835-1845.
14. Tang, X., et al., *Kinetic principles underlying pioneer function of GAGA transcription factor in live cells*. Nat Struct Mol Biol, 2022. **29**(7): p. 665-676.
15. Kim, S., et al., *DNA-guided transcription factor cooperativity shapes face and limb mesenchyme*. Cell, 2024. **187**(3): p. 692-711 e26.
16. Morgunova, E. and J. Taipale, *Structural perspective of cooperative transcription factor binding*. Current opinion in structural biology, 2017. **47**: p. 1-8.
17. Amoutzias, G.D., et al., *Choose your partners: dimerization in eukaryotic transcription factors*. Trends in Biochemical Sciences, 2008. **33**(5): p. 220-229.
18. Jolma, A., et al., *DNA-dependent formation of transcription factor pairs alters their binding specificity*. Nature, 2015. **527**(7578): p. 384-8.
19. Rodríguez-Martínez, J.A., et al., *Combinatorial bZIP dimers display complex DNA-binding specificity landscapes*. elife, 2017. **6**: p. e19272.
20. Amoutzias, G., et al., *One Billion Years of bZIP Transcription Factor Evolution: Conservation and Change in Dimerization and DNA-Binding Site Specificity*. Molecular Biology and Evolution, 2006. **24**(3): p. 827-835.
21. Judd, J., F.M. Duarte, and J.T. Lis, *Pioneer-like factor GAF cooperates with PBAP (SWI/SNF) and NURF (ISWI) to regulate transcription*. Genes & Development, 2021. **35**(1-2): p. 147-156.
22. Ackers, G.K., A.D. Johnson, and M.A. Shea, *Quantitative model for gene regulation by lambda phage repressor*. Proceedings of the National Academy of Sciences, 1982. **79**(4): p. 1129-1133.
23. Berg, O.G., R.B. Winter, and P.H. von Hippel, *Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory*. Biochemistry, 1981. **20**(24): p. 6929-48.
24. Berg, O.G. and C. Blomberg, *Association kinetics with coupled diffusion. An extension to coiled-chain macromolecules applied to the lac repressor-operator system*. Biophysical Chemistry, 1977. **7**(1): p. 33-39.
25. Kalodimos, C.G., et al., *Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes*. Science, 2004. **305**(5682): p. 386-9.
26. Winter, R.B., O.G. Berg, and P.H. von Hippel, *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): p. 6961-77.

27. Elf, J., G.W. Li, and X.S. Xie, *Probing transcription factor dynamics at the single-molecule level in a living cell*. Science, 2007. **316**(5828): p. 1191-4.
28. Marklund, E., et al., *Sequence specificity in DNA binding is mainly governed by association*. Science, 2022. **375**(6579): p. 442-445.
29. Marklund, E., et al., *DNA surface exploration and operator bypassing during target search*. Nature, 2020. **583**(7818): p. 858-861.
30. Feng, X.A., et al., *GAGA Factor Overcomes 1D Diffusion Barrier by 3D Diffusion in Search of Nucleosomal Targets*. bioRxiv, 2023.
31. Normanno, D., et al., *Probing the target search of DNA-binding proteins in mammalian cells using TetR as model searcher*. Nat Commun, 2015. **6**: p. 7357.
32. Horton, C.A., et al., *Short tandem repeats bind transcription factors to tune eukaryotic gene expression*. Science, 2023. **381**(6664): p. eadd1250.
33. Rastogi, C., et al., *Accurate and sensitive quantification of protein-DNA binding affinity*. Proc Natl Acad Sci U S A, 2018. **115**(16): p. E3692-E3701.
34. Le, D.D., et al., *Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding*. Proc Natl Acad Sci U S A, 2018. **115**(16): p. E3702-E3711.
35. Adam, G. and M. Delbruck, *Reduction of Dimensionality in Biological Diffusion Processes*, in *Structural Chemistry and Molecular Biology*. 1968, Freeman: San Francisco. p. 198-215.
36. Yogurtcu, O.N. and M.E. Johnson, *Cytosolic proteins can exploit membrane localization to trigger functional assembly*. PLoS Comput Biol, 2018. **14**(3): p. e1006031.
37. Mishra, B. and M.E. Johnson, *Speed limits of protein assembly with reversible membrane localization*. J Chem Phys, 2021. **154**: p. 194101.
38. Read, J.T., et al., *Receptor-DNA Interactions: EMSA and Footprinting*. 2009, Humana Press. p. 97-122.
39. Gebhardt, J.C., et al., *Single-molecule imaging of transcription factor binding to DNA in live mammalian cells*. Nat Methods, 2013. **10**(5): p. 421-6.
40. Riley, T.R., et al., *SELEX-seq: A Method for Characterizing the Complete Repertoire of Binding Site Preferences for Transcription Factor Complexes*. 2014, Springer New York. p. 255-278.
41. Aughey, G.N., S.W. Cheetham, and T.D. Southall, *DamID as a versatile tool for understanding gene regulation*. Development, 2019. **146**(6): p. dev173666.
42. Lionnet, T. and C. Wu, *Single-molecule tracking of transcription protein dynamics in living cells: seeing is believing, but what are we seeing?* Curr Opin Genet Dev, 2021. **67**: p. 94-102.
43. Li, M., et al., *Double DAP-seq uncovered synergistic DNA binding of interacting bZIP transcription factors*. Nature Communications, 2023. **14**(1).
44. Al Masri, C., B. Wan, and J. Yu, *Nonspecific vs. specific DNA binding free energetics of a transcription factor domain protein*. Biophys J, 2023. **122**(22): p. 4476-4487.
45. Dai, L., et al., *Revealing atomic-scale molecular diffusion of a plant-transcription factor WRKY domain protein along DNA*. Proc Natl Acad Sci U S A, 2021. **118**(23).

46. Slutsky, M. and L.A. Mirny, *Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential*. Biophysical Journal, 2004. **87**(6): p. 4021-4035.
47. Blainey, P.C., et al., *Nonspecifically bound proteins spin while diffusing along DNA*. Nature structural & molecular biology, 2009. **16**(12): p. 1224-1229.
48. Hu, J., R. Lipowsky, and T.R. Weikl, *Binding constants of membrane-anchored receptors and ligands depend strongly on the nanoscale roughness of membranes*. Proc Natl Acad Sci U S A, 2013. **110**(38): p. 15283-8.
49. Wu, Y., et al., *Transforming binding affinities from three dimensions to two with application to cadherin clustering*. Nature, 2011. **475**(7357): p. 510-3.
50. Gavutis, M., S. Lata, and J. Piehler, *Probing 2-dimensional protein-protein interactions on model membranes*. Nat Protoc, 2006. **1**(4): p. 2091-103.
51. Gavutis, M., et al., *Determination of the two-dimensional interaction rate constants of a cytokine receptor complex*. Biophys J, 2006. **90**(9): p. 3345-55.
52. Xu, G.K., et al., *Binding constants of membrane-anchored receptors and ligands: A general theory corroborated by Monte Carlo simulations*. J Chem Phys, 2015. **143**(24): p. 243136.
53. Lambert, S.A., et al., *The Human Transcription Factors*. Cell, 2018. **172**(4): p. 650-665.
54. Tronche, F., et al., *Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome*. Journal of molecular biology, 1997. **266**(2): p. 231-245.
55. Ezer, D., N.R. Zabet, and B. Adryan, *Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression*. Computational and structural biotechnology journal, 2014. **10**(17): p. 63-69.
56. Gotea, V., et al., *Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers*. Genome Research, 2010. **20**(5): p. 565-577.
57. Smith, R.P., et al., *Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model*. Nat Genet, 2013. **45**(9): p. 1021-1028.
58. Cox, R.S., 3rd, M.G. Surette, and M.B. Elowitz, *Programming gene expression with combinatorial promoters*. Mol Syst Biol, 2007. **3**: p. 145.
59. Fuda, N.J., et al., *GAGA Factor Maintains Nucleosome-Free Regions and Has a Role in RNA Polymerase II Recruitment to Promoters*. PLOS Genetics, 2015. **11**(3): p. e1005108.
60. Bhuiyan, T. and H.T.M. Timmers, *Promoter Recognition: Putting TFIID on the Spot*. Trends in Cell Biology, 2019. **29**(9): p. 752-763.
61. Lewis, M., *The lac repressor*. Comptes rendus biologies, 2005. **328**(6): p. 521-548.
62. Kmiecik, S.W. and M.P. Mayer, *Molecular mechanisms of heat shock factor 1 regulation*. Trends Biochem Sci, 2022. **47**(3): p. 218-234.
63. Varga, M.J., et al., *NERDSS: A Nonequilibrium Simulator for Multibody Self-Assembly at the Cellular Scale*. Biophysical Journal, 2020. **118**(12): p. 3026-3040.

64. Rice, S.A., *Diffusion Limited Reactions*. Comprehensive Chemical Kinetics. Vol. 25. 1985, Netherlands: Elsevier Science and Technology.
65. Collins, F.C. and G.E. Kimball, *Diffusion-Controlled Reaction Rates*. Journal of Colloid Science, 1949. **4**(4): p. 425-437.
66. Vierstra, J., et al., *Global reference mapping of human transcription factor footprints*. Nature, 2020. **583**(7818): p. 729-736.
67. Li, X., et al., *GAGA-associated factor fosters loop formation in the Drosophila genome*. Mol Cell, 2023. **83**(9): p. 1519-1526 e4.