DNA Sequence Correlations Shape Nonspecific Transcription Factor-DNA Binding Affinity

Itamar Sela and David B. Lukatsky*
Department of Chemistry, Ben-Gurion University of the Negev, Beer-Sheva, Israel

ABSTRACT Transcription factors (TFs) are regulatory proteins that bind DNA in promoter regions of the genome and either promote or repress gene expression. Here, we predict analytically that enhanced homooligonucleotide sequence correlations, such as poly(dA:dT) and poly(dC:dG) tracts, statistically enhance nonspecific TF-DNA binding affinity. This prediction is generic and qualitatively independent of microscopic parameters of the model. We show that nonspecific TF binding affinity is universally controlled by the strength and symmetry of DNA sequence correlations. We perform correlation analysis of the yeast genome and show that DNA regions highly occupied by TFs exhibit stronger homooligonucleotide sequence correlations, and thus a higher propensity for nonspecific binding, than do poorly occupied regions. We suggest that this effect plays the role of an effective localization potential that enhances quasi-one-dimensional diffusion of TFs in the vicinity of DNA, speeding up the stochastic search process for specific TF binding sites. The effect is also predicted to impose an upper bound on the size of TF-DNA binding motifs.

INTRODUCTION

Transcription factors (TFs) are proteins that regulate gene expression in both prokaryotic (e.g., bacteria) and eukaryotic (e.g., yeast or human) cells. TFs bind regulatory promoter regions of DNA in the genome. It is commonly accepted that each TF binds specifically a relatively small set of DNA sequences called TF binding motifs or TF binding sites (TFBSs). A TF binds its specific binding motifs with a higher affinity than other genomic sequences of the same length (1,2). A typical length of TF binding motif varies between 6 and 20 nucleotides. Recent high-throughput measurements of TF binding preferences on a genome-wide scale have challenged the classical picture of TF specificity (3,4). These experiments measured binding preferences of >100 TFs to tens of thousands of DNA sequences and demonstrated a high level of multispecificity in TF binding (3,4). It has been also pointed out that weak-affinity TF binding motifs are essential for gene-expression regulation (5).

A key question is how TFs find their specific binding sites in a background of $10^6 - 10^9$ nonspecific sites in a cell genome. This question was first addressed theoretically in seminal works of Berg, Winter, and von Hippel (6,7). The central idea of this approach, as expressed in recent reviews (8–10), is that the search process is a combination of three-dimensional and one-dimensional diffusion. It has been shown in different theoretical models that one-dimensional diffusion (termed sliding or hopping in different models) facilitates the search process under certain conditions (11–17). Despite the success of these phenomenological models, a complete understanding of the search process phenomena is still lacking (8). In particular, one of the key open questions is what makes a TF switch from three-dimensional diffusion to one-dimensional sliding in specific genomic locations (8). Invariably, an assumption is made about the existence of some nonspecific binding sites that bring TFs to the vicinity of DNA for one-dimensional sliding. This assumption is a key component of all theoretical models, yet the molecular origin of this effect is not understood (8,10). Recent single-molecule experimental studies undoubtedly show that different DNA-binding proteins spend the majority of their time nonspecifically bound and diffusing along DNA (18–22). The question is, what biophysical mechanism provides such nonspecific attraction toward genomic DNA and regulates the strength of this attraction at a given genomic location?

Here, we predict that DNA sequence correlations statistically regulate nonspecific TF-DNA binding preferences. Depending on the symmetry and lengthscale of sequence correlations, the nonspecific binding affinity can be either enhanced or reduced. In particular, we show that homooligonucleotide sequence correlations, where nucleotides of the same type are clustered together generically, reduce the nonspecific TF-DNA binding free energy, thus enhancing the binding affinity (Fig. 1). Sequence correlations in which nucleotides of different types alternate have the opposite effect, increasing the nonspecific TF-DNA binding free energy (Fig. 1). Correlation analysis of the yeast-genome regulatory sequences suggests that the predicted design principle is exploited at the genome-wide level to increase the strength of nonspecific binding at these regulatory genomic locations.

This article is organized as follows. First, we present a simple, analytically solvable model that describes TF-DNA binding. This model uses two-nucleotide alphabet

Submitted March 15, 2011, and accepted for publication April 19, 2011.
*Correspondence: lukatsky@bgu.ac.il
Editor: Laura Finzi.

0006-3495/11/07/0160/7 $2.00
doi: 10.1016/j.bpj.2011.04.037

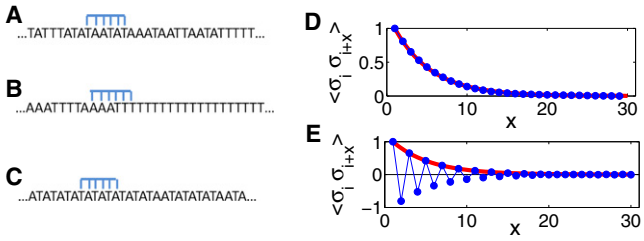

FIGURE 1 Schematic representation of the model for TF binding to DNA, and examples of DNA sequence correlation functions. (*A*) Random sequence. (*B*) Enhanced homooligonucleotide (i.e., ferromagneticlike) correlations lead to statistically enhanced nonspecific TF-DNA binding affinity. (*C*) Enhanced antiferromagneticlike correlations (alternating nucleotides of different types) lead to reduced nonspecific TF-DNA binding affinity. All examples of sequences (*A–C*) represent simulation snapshots. (*D* and *E*) Examples of the correlation function computed for sequences with enhanced ferromagneticlike correlations (*D*) and those with enhanced antiferromagneticlike correlations (*E*), where bold lines represent the exponential decay of the correlation functions.

DNA sequences. We develop a stochastic procedure allowing us to design DNA sequences with a controlled symmetry and strength of sequence correlations. We analyze the free energy of nonspecific TF-DNA binding within the framework of this model, and give an intuitive explanation for the origin of the predicted effect. Second, we generalize the model to four-letter alphabet DNA sequences and show, as well, that all key conclusions hold qualitatively true in this case. Third, we compute the free energy of nonspecific TF-DNA binding for yeast genomic sequences and show that sequences highly occupied by TFs in vivo possess a statistically higher propensity for nonspecific binding to TFs compared with sequences depleted in TFs. We conclude by proposing experiments that will allow direct testing of the predicted effect.

THEORY AND RESULTS

Free energy of nonspecific TF-DNA binding in model sequences

In this work, we use a simple variant of the Berg-von Hippel model to describe TF-DNA binding (1). For the analytical analysis, we apply the model to artificial DNA sequences containing two types of nucleotide rather than four. However, we show that all key conclusions hold qualitatively true for four-nucleotide alphabet sequences, as well.

The energy of a TF bound to DNA at a specific location i (see Fig. 1) can be expressed as

$$U(i) = -K\sum_{j=i}^{\mathscr{M}+i-1}\sigma_j, \tag{1}$$

where i and j represent individual basepairs, $\mathscr{M}$ is the effective length of the TF (i.e., the number of contacts between TF and DNA), $\sigma_j = \pm 1$ describes two possible nucleotide types at each position j, and K is the interaction strength. We therefore assume that the energy contributions of individual basepairs to the total binding energy, $U(i)$, are additive. We also assume that the energy of each contact is exclusively defined by the basepair type. The sequence of a DNA molecule of length L is uniquely defined by the set of L numbers, σ_j, where $j = 1\ldots L$.

We note that Eq. 1 provides a minimal model for TF-DNA binding. It captures the recognition specificity of TF in the simplest possible way, by assigning different contact energies, $+K$ and $-K$, with two possible nucleotide types. In reality, a TF recognizes DNA motifs forming a complex, cooperative network of hydrogen and electrostatic bonds (1,2). Yet we suggest that the design principle for enhanced nonspecific TF-DNA binding predicted using such a simplified model is likely to be quite general and robust with respect to microscopic details of TF-DNA interactions.

The free energy of binding of an individual TF to DNA is given by $\mathscr{F} = -k_BT\ln Z$, with the partition function

$$Z = \sum_{i=1}^{L}\exp(-U(i)/k_BT), \tag{2}$$

where k_B is the Boltzmann constant, T is the absolute temperature, and we imply periodic boundary conditions. We ask the question, what are the statistical properties of $\mathscr{F}$ as a function of the symmetry and strength of DNA sequence correlations?

To answer this question, we first design a DNA sequence using a stochastic design procedure. This procedure allows nucleotides within the DNA sequence to anneal, with each configuration being accepted with the Boltzmann probability

$$p(E_d) = \frac{1}{Z_d}e^{-E_d/k_BT_d}, \tag{3}$$

where T_d is the design temperature controlling the strength of correlations (this is different from the thermodynamic temperature, T), E_d is the design intra-DNA energy. For simplicity, we take into account only the nearest-neighbor interactions in the design energy:

$$E_d = -J\sum_{i=1}^{L}\sigma_i\sigma_{i+1}, \tag{4}$$

where J is the design intrasequence interaction strength, and Z_d is the corresponding Ising model partition function (23),

$$Z_d = 2^L\left(\cosh^L(\beta_dJ) + \sinh^L(\beta_dJ)\right), \tag{5}$$

where $\beta_d = 1/k_BT_d$.

The ferromagneticlike case, $J>0$, produces sequences with homooligonucleotide stretches. The correlation length, $\xi = -1/\ln(\tanh\beta_d|J|)$, is the characteristic lengthscale of the correlations decay, $\langle\sigma_i\sigma_{i+x}\rangle = \exp(-x/\xi)$ (23). The

antiferromagneticlike case, $J<0$, produces sequences with a different symmetry of alternating nucleotides (Fig. 1). We define the average free energy of TF binding to DNA as the annealed average,

$$\langle \mathscr{F} \rangle = -\frac{1}{\beta} \ln \langle Z \rangle, \tag{6}$$

where the averaging is performed with probability $p(E_d)$ (Eq. 3) and $\beta = 1/k_B T$. The quenched averaging, $\langle \mathscr{F} \rangle_q = -\langle \ln Z \rangle / \beta$, is analyzed numerically below, and it gives qualitatively similar results (Fig. 2). The averaging in Eq. 6 gives

$$\langle Z \rangle = \frac{2^{L-M-1} L}{Z_d} \Bigg[\left(\lambda_+^{\mathscr{M}} + \lambda_-^{\mathscr{M}}\right) \left(\cosh^{L-\mathscr{M}}(\beta_d J) + \sinh^{L-\mathscr{M}}(\beta_d J)\right) + \left(\lambda_+^{\mathscr{M}} - \lambda_-^{\mathscr{M}}\right) \left(\cosh^{L-\mathscr{M}}(\beta_d J) - \sinh^{L-\mathscr{M}}(\beta_d J)\right) \times \frac{e^{-\beta_d J}}{\sqrt{e^{2\beta_d J} \sinh^2(\beta K) + e^{-2\beta_d J}}} \Bigg], \tag{7}$$

where Z_d is given by Eq. 5, and

$$\lambda_{\pm} = e^{\beta_d J} \cosh(\beta K) \pm \sqrt{e^{2\beta_D J} \sinh^2(\beta K) + e^{-2\beta_d J}}. \tag{8}$$

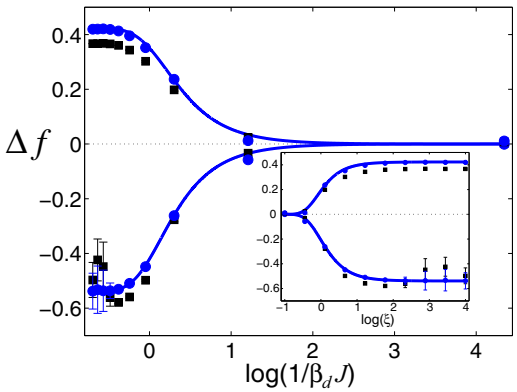

FIGURE 2 TF-DNA binding free-energy difference normalized per base-pair, $\Delta f = \beta \langle \Delta \mathscr{F} \rangle / \mathscr{M}$, computed using Eq. 7 as a function of the reduced design temperature, $1/\beta_d J$ (*solid curves*). The upper and lower branches of the graph correspond to $J<0$ (antiferromagneticlike DNA sequence correlations) and $J>0$ (ferromagneticlike correlations), respectively. The results of MC simulations of the system are in excellent agreement with the analytical results (*solid circles*). We used the parameters $\beta K = 1$, $\mathscr{M} = 18$, and $L = 1000$. In Monte Carlo simulations, we used 7.5×10^6 MC moves to design each DNA sequence at each value of T_d. To generate each point in the plot, we used a set of 100 sequences. To compute error bars, we divided each set of 100 sequences randomly into 10 subsets, and then calculated the SD of the subset averages for Δf. The error bars correspond to 1 SD. The numerically computed quenched average, $-\langle \ln (Z/Z_\infty) \rangle / \mathscr{M}$, is also shown (*solid squares*). In the computations, we used the same parameters and definitions as specified above. (*Inset*) Same data for Δf as in the main figure, but plotted as a function of ξ.

We argue that the DNA correlations symmetry affects statistically the interaction free energy. It is natural therefore to analyze the free-energy difference between designed sequences and their randomized analogs, which lack symmetry:

$$\langle \Delta \mathscr{F} \rangle = \langle \mathscr{F} \rangle - \langle F_\infty \rangle, \tag{9}$$

where $\langle F_\infty \rangle$ is the free energy computed for entirely random sequences (i.e., for sequences designed using a very high value of T_d or, equivalently, $1/\beta_d J \gg 1$). The first key property of $\langle \Delta \mathscr{F} \rangle$ is that it is invariant with respect to the sign of the TF-DNA binding-affinity constant, K. Second, it is always satisfied that $\langle \Delta \mathscr{F} \rangle < 0$ if $J>0$ (ferromagneticlike correlations within designed DNA sequences (see Eq. 4)), and $\langle \Delta \mathscr{F} \rangle > 0$ if $J<0$ (antiferromagneticlike correlations). Fig. 2 shows the behavior of $\langle \Delta \mathscr{F} \rangle$ at different magnitudes of the design strength. The central observation here is that the behavior of $\langle \Delta \mathscr{F} \rangle$ critically depends on the symmetry and the lengthscale of DNA sequence correlations. The presence of homooligonucleotide stretches along DNA sequences statistically increases the propensity of such sequences for nonspecific binding to TFs. The DNA stretches with nucleotides of different types alternating produce the opposite effect: such sequences will have a reduced propensity for nonspecific binding. We note that the quenched average, $\langle \Delta \mathscr{F} \rangle_q = -\langle \ln (Z/Z_\infty) \rangle / \beta$, computed numerically, is in good agreement with the annealed average (Fig. 2)

The reduction in TF-DNA binding free energy in the presence of homooligonucleotide sequence correlations can be understood intuitively in the following way. Homooligonucleotide sequence correlations generically enhance fluctuations of the TF-DNA binding energy, $\sigma_U^2 = \langle U^2 \rangle - \langle U \rangle^2$. This effect has to do with the symmetry: a TF sliding along correlated DNA sequences where nucleotides of the same type have the tendency to cluster, will experience homogeneous DNA islands, such as poly(dA:dT) and poly(dC:dG) tracts. Statistically, this leads to the dominant contribution of either very strong or very weak energies to the TF-DNA binding energy spectrum. This symmetry effect leads therefore to the widening of the TF-DNA binding energy spectrum, $P(U)$. Such widening generically leads to the reduction of the TF-DNA binding free energy, due to the fact that the dominant contribution to the partition function, Z, comes from the low-energy tail of $P(U)$ (24). Alternatively, a DNA sequence with enhanced antiferromagneticlike correlations (i.e., with alternating nucleotides of different types) will have the opposite effect: a TF sliding along such a sequence will experience very heterogeneous binding sites. This leads to the narrowing down of $P(U)$ and consequently to the increase of the nonspecific TF-DNA binding free energy.

We note that the predicted effect is not restricted to TFs; it is operational for any other kind of DNA-binding protein.

Extension of the model to four-letter-alphabet DNA sequences

We show in this section that four-letter alphabet DNA sequences demonstrate statistical binding properties qualitatively similar to those of the two-letter alphabet sequences analyzed above. This will allow us to extend all our insights gained from the analytical model directly to genomic DNA sequences. We argue that the same underlying physical mechanism controls the nonspecific binding propensity in both cases.

Contrary to the two-letter alphabet DNA sequences, where within our modeling framework a TF is fully described by the single parameter K, in the four-letter alphabet DNA case, a TF is characterized by four energy parameters, K_A, K_T, K_C, and K_G. Although those energy constants are generally unknown, their order of magnitude can be roughly estimated as $1k_BT$, and in addition, we allow the TF-DNA contact energies to fluctuate. We therefore draw these energies from the Gaussian probability distributions, $P(K_\alpha)$, with zero mean and standard deviations (SD), σ_α, where $\alpha = A, T, C, G$; and we average the free energy over many TF realizations.

The binding energy of TF at a given site i is expressed by

$$U(i) = -\sum_{j=i}^{\mathscr{M}+i-1}\sum_{\alpha=1}^{4} K_\alpha \sigma_j^\alpha, \qquad (10)$$

where σ_j^α is a four-component vector of the type $(\delta_{\alpha A}, \delta_{\alpha T}, \delta_{\alpha C}, \delta_{\alpha G})$ at each DNA position j, with the position of 1 specifying one of four possible identities (A, T, C, or G) of the basepair at position j, with $\delta_{\alpha\beta}$ being the Kronecker delta. The sequence design procedure is analogous to the one introduced above, Eq. 4, with the 4×4 symmetric matrix of the design potentials entering the sum, $-J_{\alpha\beta}\sigma_i^\alpha\sigma_{i+1}^\beta$. The results for the average TF-DNA binding free energy in the ensemble of different TFs are shown in Fig. 3. The key conclusion here is that, provided that in the design procedure nucleotides of the same type attract, the lower the design temperature (and thus the longer the correlation length of homooligonucleotide stretches), T_d, the lower the TF-DNA binding free energy.

Free energy of nonspecific TF-DNA binding in yeast genome

We ask further the key question, is the predicted design principle for nonspecific TF-DNA binding operating in a living cell? To answer this question, we computed TF-DNA binding free energies using yeast-genome DNA. Our working hypothesis here is that if the predicted effect is operational, genomic regions that need to be highly accessible to TFs should possess a higher propensity for nonspecific TF-DNA binding than regions that need not be highly accessible to TFs. To test this hypothesis, we compiled two datasets of genomic DNA. First, we collected ~1600 high-confidence yeast DNA regulatory promoter sequences (for organelle organization and biogenesis genes), each sequence 100 nucleotides long. We describe this dataset as upstream. These upstream sequences are experimentally known to be highly accessible to TFs. The second dataset involves a comparable number of weakly accessible genomic sequences. For this purpose, we chose the first 100 nucleotide stretches of the mRNA coding regions of those organelle organization and biogenesis genes. We describe the second dataset as downstream. The datasets were compiled by Lee et al. (25).

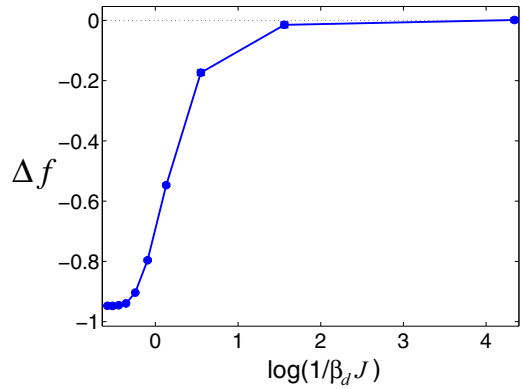

FIGURE 3 Average TF-DNA binding free energy, Δf, numerically computed at different values of the design temperature, where $\Delta f = -\langle \ln (Z/Z_\infty)\rangle/\mathscr{M}$, with Z_∞ the partition function for an entirely random DNA sequence. We designed 200 sequences with length $L = 400$ at each T_d. We performed 5×10^6 MC steps to design each sequence, attempting in each step to exchange two basepairs chosen at random. The overall nucleotide composition for each sequence was uniform and fixed. The design potential was $+J$ (attraction) for identical nearest-neighbor basepairs and $-J$ (repulsion) for different nearest-neighbor basepairs, with $J = 1k_BT$. The contact energies, K_α, were drawn from a Gaussian distribution, $P(K_\alpha)$, with zero mean, $\langle K_\alpha\rangle = 0$, and standard deviation $\sigma_\alpha = 2k_BT$ for each nucleotide type α. We computed Δf as an average over 250 TFs and 200 sequences at each T_d and used $\mathscr{M} = 8$. The error bars are calculated as specified in Fig. 2, and they are smaller than the marker size.

It turns out that upstream sequences demonstrate statistically stronger homooligonucleotide correlations in A and T compared to downstream sequences, and the difference in correlations of C and G is not significant between the datasets. The normalized correlation function, $C_{AA}(x)$, computed for the sets of upstream and downstream sequences, is shown in Fig. 4 (*blue* and *red*, respectively). This function is defined as $C_{\alpha\alpha}(x) = s_{\alpha\alpha}(x)/\langle s^r_{\alpha\alpha}(x)\rangle$, where $s_{\alpha\alpha}(x) = \langle \sigma_\alpha(i)\sigma_\alpha(i+x)\rangle$, and $\langle s^r_{\alpha\alpha}(x)\rangle$ is obtained analogously, using the set of randomly permuted sequences averaged with respect to different random realizations. $C_{TT}(x)$ shows qualitatively similar behavior (data not shown).

We now compare the TF-DNA binding free energies for those two datasets. To get rid of the compositional bias, for a given TF interacting with a given DNA sequence, we always compare the difference, $\Delta\mathscr{F}$, between the actual free energy, $\mathscr{F}$, and the free energy computed for the randomized sequence (preserving the nucleotide composition of

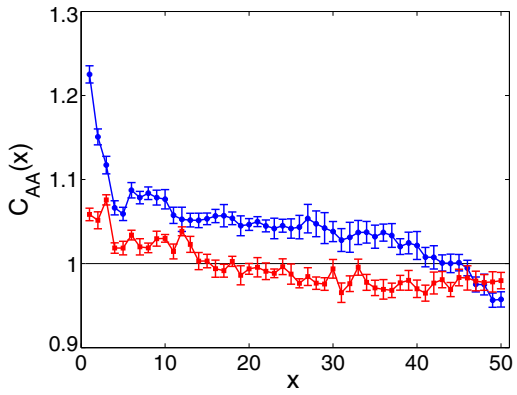

FIGURE 4 Normalized correlation function, $C_{AA}(x)$ (see the text for the definition), computed for upstream (*blue circles*) and downstream (*red squares*) sequence sets. Each set consists of 1663 sequences, with each sequence 100 nucleotides long.

each sequence), averaged over several random realizations, $\mathscr{F}_\infty$: $\Delta\mathscr{F} = \mathscr{F} - \mathscr{F}_\infty$. We therefore compute numerically the probability distribution, $P(\Delta\mathscr{F})$, for these two datasets of sequences interacting with a model set of TFs. The TF-DNA binding contact energies, K_α, are drawn from the Gaussian distributions, $P(K_\alpha)$, as described above. We stress that the only external parameters entering the model are the SDs, σ_α, of $P(K_\alpha)$. In our calculations, we set $\sigma_\alpha = 2k_BT$ for all α. The computed $P(\Delta\mathscr{F})$ values for upstream and downstream DNA sequences are shown in Fig. 5 *A*. We also show the cumulative probability at different values of the selectivity cutoff (Fig. 5 *B*). The central conclusion here is that due to the presence of enhanced homooligonucleotide (i.e., ferromagneticlike) sequence correlations, nonspecific TF-DNA binding is statistically enhanced. At

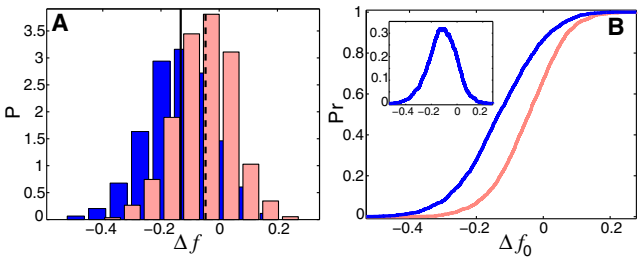

FIGURE 5 (*A*) Computed $P(\Delta f)$ for 1663 upstream (*blue*) and downstream (*red*) yeast genomic sequences, where $\Delta f = \beta\langle\Delta\mathscr{F}\rangle_{TF}/\mathscr{M}$, and $\Delta\mathscr{F} = \mathscr{F} - \mathscr{F}_\infty$. For each given TF, $\mathscr{F}_\infty$ is computed as an average over 50 randomized sequence replicas (randomization preserves the nucleotide composition of each sequence). For each sequence, we computed $\Delta\mathscr{F}$ for 250 TFs and then took the average of these 250 values, $\langle\Delta\mathscr{F}\rangle_{TF}$. We used $\mathscr{M} = 8$. The TF-DNA contact energies, K_α, are drawn from a Gaussian probability distribution, $P(K_\alpha)$, with zero mean and standard deviation $\sigma_\alpha = 2k_BT$, where α represents four possible nucleotides. Vertical lines show the mean of Δf. (*B*) The cumulative probability, $Pr(\Delta f \le \Delta f_0) = \int_0^{\Delta f_0} P(\Delta f)d\Delta f$, computed using $P(\Delta f)$ from *A*. (*Inset*) Difference between upstream and downstream $Pr(\Delta f \le \Delta f_0)$.

the maximal selectivity cutoff, where $\Delta\mathscr{F}_c \simeq -0.1k_BT$ per basepair, the probability of TF binding with the free energy below $\Delta\mathscr{F}_c$ is >30% higher to upstream DNA regions than to downstream regions (Fig. 5 *B*). This effect leads to a shift of the thermodynamic equilibrium toward enhanced occupancy of TFs binding upstream regions rather than downstream regions. The average strength of the effect on TF occupancy can be estimated from the difference of the average TF-DNA binding free energies, $\langle\Delta\Delta\mathscr{F}\rangle = \langle\Delta\mathscr{F}^{up}\rangle - \langle\Delta\mathscr{F}^{down}\rangle \simeq -0.1k_BT$ per basepair, between upstream and downstream DNA regions (difference between the peak positions in Fig. 5 *A*). For a TF forming $\mathscr{M}$ contacts within the TF-DNA binding site, this difference will produce $n_{up}/n_{down} \simeq \exp(0.1 \times \mathscr{M})$ shift in the relative binding occupancy, where n_{up} and n_{down} are the numbers of bound TFs in the upstream and downstream regions, respectively. For a typical TF forming contacts with 10 DNA basepairs, this leads to $n_{up}/n_{down} \simeq 2.7$. We emphasize that the latter estimate provides only a lower-bound limit for the strength of the predicted correlational effect. We suggest, therefore, that the predicted mechanism for enhanced nonspecific TF-DNA binding is operational in promoter regions of a significant fraction of yeast genes.

Finally, we note that our findings suggest the existence of an upper bound for the TF-DNA binding motif size, imposed by the maximal possible strength of nonspecific binding. It is predicted (11) that if the free energy of TF-DNA nonspecific binding falls below $-2k_BT$, this significantly slows down the sliding diffusion of TFs along DNA. Our estimates therefore suggest that such slowing down is likely when the binding motif approaches the size of 20 basepairs.

DISCUSSION AND CONCLUSION

Here, we predict a generic biophysical mechanism that statistically regulates the strength of nonspecific TF-DNA binding in a genome. We showed analytically and numerically, using both artificially designed and genomic DNA sequences, that homooligonucleotide correlations statistically enhance nonspecific TF-DNA binding affinity. We described the symmetry of such correlations as ferromagneticlike. Alternatively, DNA sequences possessing enhanced correlations of alternating nucleotides of different types (referred to here as antiferromagneticlike) have a reduced propensity for nonspecific binding to TFs.

Our model description of TF-DNA binding is highly simplified. Yet we suggest that the design principle for enhanced nonspecific TF-DNA binding predicted in this work is likely to be quite general, it is operational in genomic locations highly occupied by TFs, and it is likely to be the rule rather than the exception. The robustness of our conclusions with respect to the details of the model stems from the fact that the predicted effect arises exclusively due to DNA sequence symmetry and its strength (which is determined by the lengthscale of the correlations

decay). Computational analysis of the TF-DNA binding free energy in ~1600 yeast genomic DNA regions highly occupied by TFs shows that those regions possess much higher propensity for nonspecific binding compared with regions depleted in TFs. In our analysis, we used a simple procedure to get rid of the DNA compositional bias, allowing us to fairly compare the relative free energies of nonspecific binding in different genomic locations.

We estimated that in yeast, the predicted effect leads to a free energy reduction per DNA basepair in contact with a TF of at least $\sim 0.1 k_B T \simeq 60$ cal/mol (on average) in DNA regions with enhanced propensity for nonspecific binding. This leads to at least a threefold concentration enrichment in TFs (on average) of such highly promiscuous DNA regions in yeast. We suggest, therefore, that in addition to all known signals, genomic DNA might also encode its intrinsic propensity for nonspecific binding to TFs. The predicted effect plays the role of an effective, nonspecific localization potential, enhancing the level of one-dimensional diffusion of TFs along genomic DNA at the genome-wide level and thus speeding up the search process for specific TF binding sites (6–11). We stress that all our conclusions are obtained assuming a quasiequilibrium nature of TF-DNA binding. It would be important to investigate the dynamic aspects of the predicted phenomena.

It is important to note that too high a level of nonspecific TF-DNA binding impairs the overall search efficiency (11,12). This suggests that the strength of the predicted effect in vivo might be subject to both positive and negative regulation. It has been pointed out in a seminal work of Iyer and Struhl (26) that activity of poly(dA:dT) tracts increases with their length. We suggest that this observation is a direct consequence of the effect of enhanced nonspecific TF-DNA binding by poly(dA:dT), predicted here. Another key observation of Iyer and Struhl (26), that poly(dC:dG) functions in a similar manner to poly(dA:dT), further strengthens our prediction.

Extensive correlation analysis of different organismal genomes and direct, large-scale measurements of TF-DNA binding preferences using DNA sequences with controlled strength and symmetry of correlations, should provide an ultimate test of the phenomenon predicted here. Protein-DNA binding arrays (4) and high-throughput microfluidics technology (3) allow a direct experimental test of our predictions in vitro. A key experiment would measure the TF-DNA binding affinity in different sets of DNA, each set containing DNA sequences with a specific TF-DNA binding motif embedded in a background of nonspecific sequences with varying symmetry and strength of correlations between DNA sets. We expect that DNA sequences with enhanced homooligonucleotide correlations in background sequences will generically possess a higher binding affinity to different TFs compared with background sequences either lacking such correlations or having correlations with alternating nucleotides of different types.

We thank Noa Musa-Lempel for help in compiling yeast genomic sequences. D.B.L. acknowledges financial support from Israel Science Foundation grant 1014/09.

REFERENCES

1. Berg, O. G., and P. H. von Hippel. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* 193:723–750.
2. Stormo, G. D., and D. S. Fields. 1998. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* 23:109–113.
3. Fordyce, P. M., D. Gerber, …, S. R. Quake. 2010. De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nat. Biotechnol.* 28:970–975.
4. Badis, G., M. F. Berger, …, M. L. Bulyk. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science.* 324:1720–1723.
5. Segal, E., T. Raveh-Sadka, …, U. Gaul. 2008. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature.* 451:535–540.
6. Berg, O. G., R. B. Winter, and P. H. von Hippel. 1981. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry.* 20:6929–6948.
7. von Hippel, P. H., and O. G. Berg. 1989. Facilitated target location in biological systems. *J. Biol. Chem.* 264:675–678.
8. Kolomeisky, A. B. 2011. Physics of protein-DNA interactions: mechanisms of facilitated target search. *Phys. Chem. Chem. Phys.* 13:2088–2095.
9. Halford, S. E., and J. F. Marko. 2004. How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.* 32:3040–3052.
10. Mirny, L., M. Slutsky, …, A. Kosmrlj. 2009. How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J. Phys. A.* 42:434013.
11. Slutsky, M., and L. A. Mirny. 2004. Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophys. J.* 87:4021–4035.
12. Slutsky, M., M. Kardar, and L. A. Mirny. 2004. Diffusion in correlated random potentials, with applications to DNA. *Phys. Rev. E.* 69: 061903.
13. Mirny, L. A. 2010. Nucleosome-mediated cooperativity between transcription factors. *Proc. Natl. Acad. Sci. USA.* 107:22534–22539.
14. Cherstvy, A. G., A. B. Kolomeisky, and A. A. Kornyshev. 2008. Protein-DNA interactions: reaching and recognizing the targets. *J. Phys. Chem. B.* 112:4741–4750.
15. Das, R. K., and A. B. Kolomeisky. 2010. Facilitated search of proteins on DNA: correlations are important. *Phys. Chem. Chem. Phys.* 12:2999–3004.
16. Hu, T., A. Yu. Grosberg, and B. I. Shklovskii. 2006. How proteins search for their specific sites on DNA: the role of DNA conformation. *Biophys. J.* 90:2731–2744.
17. Sheinman, M., and Y. Kafri. 2009. The effects of intersegmental transfers on target location by proteins. *Phys. Biol.* 6:016003.
18. Wang, Y. M., R. H. Austin, and E. C. Cox. 2006. Single molecule measurements of repressor protein 1D diffusion on DNA. *Phys. Rev. Lett.* 97:048302.
19. Blainey, P. C., G. Luo, …, X. S. Xie. 2009. Nonspecifically bound proteins spin while diffusing along DNA. *Nat. Struct. Mol. Biol.* 16:1224–1229.
20. Elf, J., G.-W. Li, and X. S. Xie. 2007. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science.* 316:1191–1194.

21. Tafvizi, A., F. Huang, J. S. Leith, A. R. Fersht, …, A. M. van Oijen. 2008. Tumor suppressor p53 slides on DNA with low friction and high stability. *Biophys. J.* 95:L01–L03.

22. Liu, S., E. A. Abbondanzieri, …, X. Zhuang. 2008. Slide into action: dynamic shuttling of HIV reverse transcriptase on nucleic acid substrates. *Science.* 322:1092–1097.

23. Plischke, M., and B. Bergersen. 1994. Equilibrium Statistical Physics. World Scientific, Singapore.

24. Lukatsky, D. B., and M. Elkin. 2011. Energy fluctuations shape free energy of biomolecular interactions. arXiv:1101.4529v1.

25. Lee, W., D. Tillo, …, C. Nislow. 2007. A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* 39:1235–1244.

26. Iyer, V., and K. Struhl. 1995. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* 14:2570–2579.