Nonspecific Protein-DNA Binding Is Widespread in the Yeast Genome

Ariel Afek and David B. Lukatsky*
Department of Chemistry, Ben-Gurion University of the Negev, Be'er-Sheva, Israel

ABSTRACT   Recent genome-wide measurements of binding preferences of ~200 transcription regulators in the vicinity of transcription start sites in yeast, have provided a unique insight into the *cis*-regulatory code of a eukaryotic genome. Here, we show that nonspecific transcription factor (TF)-DNA binding significantly influences binding preferences of the majority of transcription regulators in promoter regions of the yeast genome. We show that promoters of SAGA-dominated and TFIID-dominated genes can be statistically distinguished based on the landscape of nonspecific protein-DNA binding free energy. In particular, we predict that promoters of SAGA-dominated genes possess wider regions of reduced free energy compared to promoters of TFIID-dominated genes. We also show that specific and nonspecific TF-DNA binding are functionally linked and cooperatively influence gene expression in yeast. Our results suggest that nonspecific TF-DNA binding is intrinsically encoded into the yeast genome, and it may play a more important role in transcriptional regulation than previously thought.

INTRODUCTION

High-throughput measurements of protein-DNA binding in vivo

Specific transcription factor (TF) binding to genomic DNA in promoter regions is a key mechanism regulating gene expression in both prokaryotic and eukaryotic organisms. Recent advances in high-throughput methods of measuring TF-DNA binding preferences genome-wide in vivo, such as chromatin immunoprecipitation (ChIP) followed by microarray analysis (ChIP-chip), or followed by high-throughput sequencing analysis (ChIP-seq), provide a remarkable snapshot of the physical interaction map that exists within a living cell in different organisms (1–8). These measurements have demonstrated quite generally that TFs extensively bind thousands of active and inactive regions across the genome, and strikingly, in many cases no specific TF binding sites (TFBSs) can be identified in the regions of particularly strong binding (3–6). These observations have thus challenged the classical picture of specific TF-DNA binding. In their recent, seminal work, Venters et al. (1) have measured binding preferences of 202 regulatory, DNA-binding proteins in three representative genomic regions in yeast. This work provides the most extensive view of TF-DNA binding currently in yeast, and it concludes that over 90% of yeast promoter regions are significantly occupied by more than 10 regulators, and ~10% are occupied by at least 75 regulators. The key, open question is what determines binding preferences of these regulators toward genomic DNA?

Definition and design principles of nonspecific (nonconsensus) protein-DNA binding

The existence and functional importance of nonspecific protein-DNA binding in *Escherichia coli* were demonstrated

in the early seventies of the last century in seminal experimental works of Riggs et al. (9), and Hinkle and Chamberlin (10); and in seminal theoretical works of von Hippel et al. (11–15), and Richter and Eigen (16). These early works suggested that DNA-binding proteins use different conformations in specific and nonspecific protein-DNA binding modes, respectively. Recent direct biophysical measurements performed both in vivo (17) and in vitro (18–22), unambiguously show that nonspecific protein-DNA binding is widespread in genomes of different organisms.

As presented in seminal works of von Hippel and Berg (13–15), the notion of nonspecific protein-DNA binding can be schematically described by two key-related mechanisms. The first mechanism is largely sequence-independent DNA, and it is entirely based i), on the overall electrostatic attraction between DNA-binding proteins (such as TFs) and DNA, and ii), on the overall geometry of DNA (13). The second mechanism assumes that for any sequence-specific DNA-binding protein, any DNA sequence, which is similar enough to canonical recognition motifs (consensus sequences) of this protein, possesses some residual protein-DNA binding affinity. For example, the yeast transcription factor Reb1 binds the TTACCCG motif with a relatively high affinity, and hence, any sequence similar to this consensus sequence is expected to possess a higher affinity to Reb1 than an entirely unrelated sequence (23). The fact that statistically, there is a high probability of having such sequence in many genomic locations by pure chance, might lead to nonspecific protein-DNA binding (13,24).

We have recently suggested the existence of an additional, nonconsensus nonspecific protein-DNA binding mechanism (25,26). By using the term nonconsensus nonspecific binding we mean to express the fact that the predicted binding affinity is computed without experimental knowledge of the high-affinity sites for the TFs. In what follows, we always mean such nonconsensus nonspecific TF-DNA binding. In particular, we predicted analytically

that correlation properties of genomic DNA sequences generically regulate the nonspecific TF-DNA binding affinity (25). We use the term correlation to describe statistically significant repeats of DNA sequence patterns. For example, we predicted that homooligonucleotide sequence correlations, where nucleotides of the same type are clustered together (such as poly(dA:dT) and poly(dC:dG) tracts) generically enhance the nonspecific TF-DNA binding affinity. Sequence correlations in which nucleotides of different types alternate have the opposite effect, reducing the nonspecific TF-DNA binding affinity (25). Because the predicted effect stems from the intrinsic symmetry properties of DNA sequences, we suggested that it is quite general, and qualitatively robust with respect to microscopic details of the protein-DNA interaction potential (25). We also note that the predicted effect is entropy dominated, and it assumes that TFs sample all possible binding sites along DNA (25,26).

Synopsis of obtained results

Here, having obtained experimental binding preferences of 202 DNA binding proteins (1), we thought to answer the question of what role does nonspecific (nonconsensus) protein-DNA binding play in a living yeast cell, genome-wide?

To address this question, here we compute the nonspecific binding free energy of random protein-DNA binders. We use the term random binder to emphasize the fact that model TFs bind genomic DNA nonspecifically. We compute statistical properties of such nonspecific binding. Strikingly, we

show that nonspecific binding alone can explain statistical binding preferences observed experimentally. Our results provide further support of the hypothesis that nucleosome occupancy in yeast is significantly influenced by nonspecific TF-DNA binding (26).

We note that in the experiments that we are using for this analysis (1), TFs can be cross-linked and immunoprecipitated in association with a given DNA segment by virtue of at least four kinds of interactions. i), Binding to the local DNA. ii), Cooperative binding to a combination of local DNA and other locally bound TFs. iii), Cooperative binding only to other locally bound TFs and not to DNA. iv), Binding to nascent RNA transcripts and/or proteins bound to nascent RNA transcripts. Our theoretical analysis of protein-DNA binding affinity focuses largely on mechanism (i). Yet, because all our predictions are statistical in nature, and the number of experimentally measured TFBSs is very large, we suggest that all our conclusions are quite general, and most likely represent the statistical law, rather than the exception.

This work is organized as follows. First, we describe our method to compute the nonspecific TF-DNA binding free energy landscape. Second, we show that nonspecific TF-DNA binding significantly influences experimentally observed TF-DNA binding preferences in promoter regions of the yeast genome, Fig. 1 and Fig. 2. Third, we show that promoter regions of highly regulated (e.g., SAGA dominated) and weakly regulated (e.g., TFIID dominated) genes are characterized by distinct profiles of the nonspecific binding free energy, Fig. 3. In conclusion, we show that
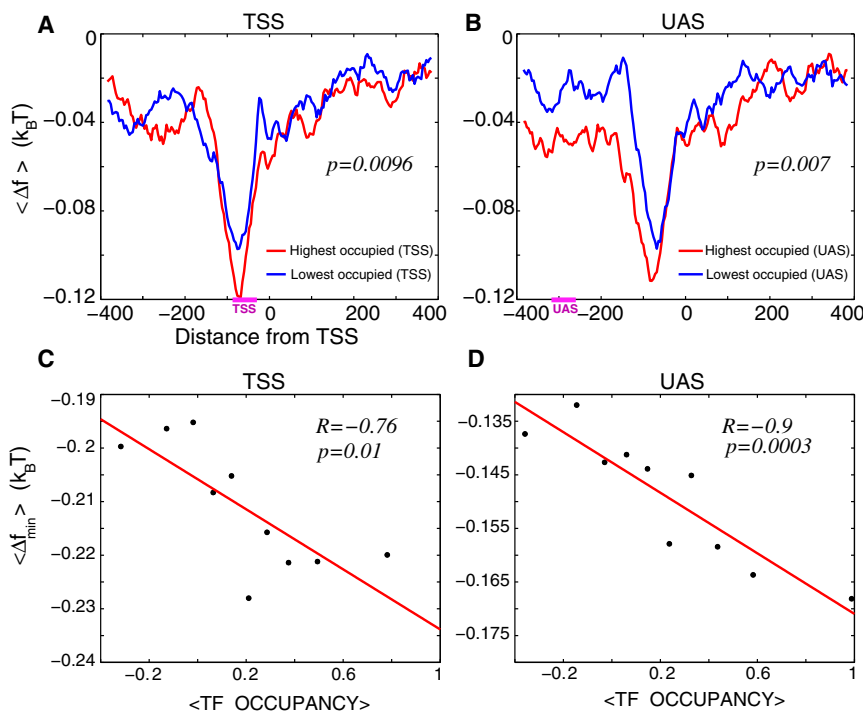

FIGURE 1 (*A*) Average free energy of nonspecific TF-DNA binding per bp, $\langle \Delta f \rangle = \langle \langle \Delta F \rangle_{TF} \rangle_{seq}/M$, computed within the interval $(-384,384)$ for the two groups of genes selected according to the experimentally measured average TF occupancy in the TSS region: 10% highest TF occupancy in the TSS region (*red*) and 10% lowest TF occupancy in the TSS region (*blue*). Each group contains 496 genes. Horizontal bar, marked TSS, on the *x* axis, shows the corresponding region where the TF occupancy was measured. (*B*) Similar to (*A*), but the two groups of genes are selected according to the experimentally measured average TF occupancy in the UAS region. Horizontal bar, marked UAS, on the *x* axis, shows the corresponding region where the TF occupancy was measured. (*C*) Correlation between the minimal value of the free energy of nonspecific TF-DNA binding within the TSS regions, $\Delta f_{min} = \min(\langle \Delta F \rangle_{TF})/M$, and the average TF occupancy within this region. Genes were binned into 10 bins according to the value of the average TF occupancy. Each point in the graph corresponds to the average, $\langle \Delta f_{min} \rangle$, for the genes in a given bin plotted as a function of the experimentally measured average TF occupancy for the genes in this bin. (*D*) Analogous to (*C*), but for $\langle \Delta f_{min} \rangle$ computed within the UAS regions, plotted versus the average TF occupancy measured within the UAS regions, as described in *C*.
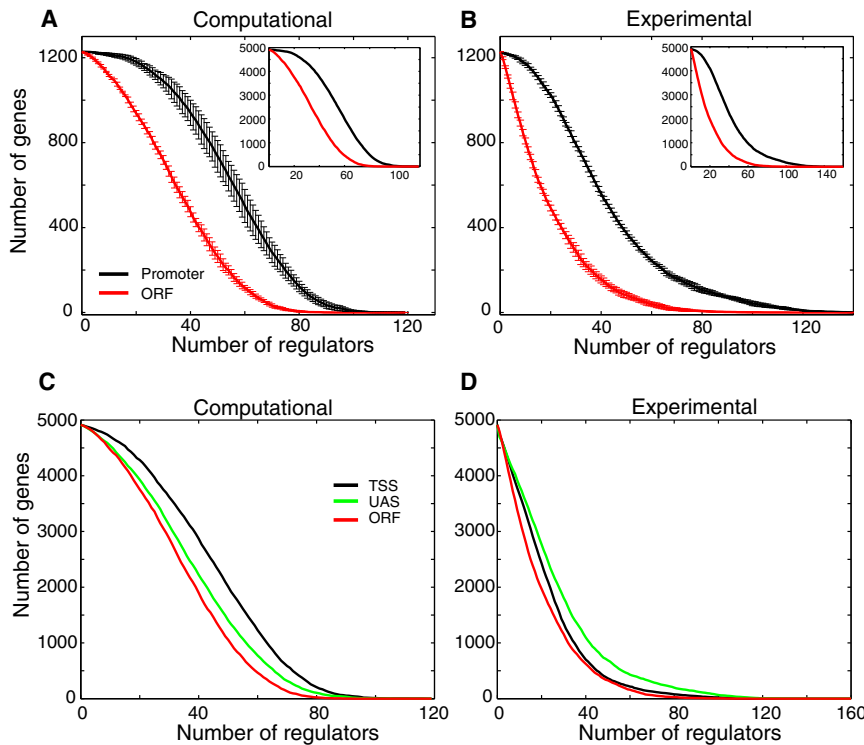
FIGURE 2 (*A* and *B*) Number of promoter regions (TSSs and UASs) (*black*) and coding regions (ORFs) (*red*) occupied by the number of regulators (i.e., TFs) indicated along the *x* axis, as computed using the model of nonspecific TF-DNA binding (*A*) and experimental data from (1). (*B*) This corresponds to Fig. 2 *A* of (1). In the computational prediction we assumed that a given genomic region is occupied by a given TF if the minimal free energy of nonspecific TF-DNA binding (within this genomic region) is less than the cutoff value of -1 k_BT, and we used 250 TFs in the computation (Materials and Methods). To compute error bars, we divided all genes into four subgroups, and computed the corresponding occupancy separately for each subgroup. The error bars are defined as one standard deviation of the occupancy between the subgroups. Inset in each panel shows the occupancy for the entire set of ~5000 genes. (*C* and *D*) Analogous to the insets in *A* and *B*, but with the cumulative TF occupancy computed separately for TSSs and UASs. We used $M = 8$ for the TF length in all our calculation of the free energy.

the level of gene expression in yeast grown in YPD medium is correlated with the landscape of the nonspecific binding free energy in promoter regions, Fig. 4 and Fig. 5.

MATERIALS AND METHODS

Gene set

In our analysis, we used a highly confident set of 4962 yeast genes from (27). We used the following terms, adopted from (1), to describe three types of genomic regions: transcription start sites (TSSs) located in the interval $(-90, -30)$, upstream activating sequences (UASs) located in the interval $(-320, -260)$, and open reading frames (ORFs) located in the 3′ end of the coding regions. The zero of the coordinate system is located in the transcription start site for each gene. We note that in (1) the ORF regions were positioned in slightly different genomic locations in different genes, downstream of the transcription start site, Table S2 of (1). In our analysis we used the precise, experimental location of the ORF for each gene.

Experimental TF occupancy

The experimental average TF occupancy in each of the three genomic locations, TSSs, UASs, and ORFs, measured in (1), is defined for each gene in the following way. For each gene, in each genomic location, we computed the average occupancy of all regulators from Table S2 of (1), measured at temperature of 25°C. Only regulators with the occupancy above 5% threshold for false discover rate reported in (1) are taken into account in the calculation of the average TF occupancy. At the end of this procedure, each genomic location, TSSs, UASs, and ORFs, respectively, in each gene is assigned a value of the average TF occupancy.

Gene expression data

The experimental gene expression data in YPD medium is taken from (28).

SAGA-dominated and TFIID-dominated genes

To compute Fig. 3 *A*, we extracted the sets of 40 SAGA-dominated, TATA-containing genes and 178 TFIID-dominated, TATA-less, nonribosomal protein genes, from Table S6 of (1). The extended list of all known SAGA-dominated and TFIID-dominated genes is taken from (29).

Transcriptional plasticity

To compute Fig. 3 *B*, we used the classification of transcriptional plasticity from (30), and refined in (31).
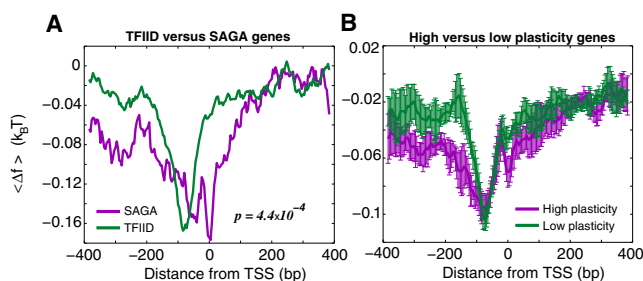

FIGURE 3 (*A*) Average free energy of nonspecific TF-DNA binding per bp, $\langle \Delta f \rangle$, computed within the interval $(-384, 384)$ for the highly confident SAGA-dominated and TFIID-dominated groups of genes, respectively. There are 40 SAGA-dominated, TATA-containing genes and 178 TFIID-dominated, TATA-less, nonribosomal protein genes, respectively (these highly confident groups are taken from (1)). (*B*) Average free energy of nonspecific TF-DNA binding per bp, $\langle \Delta f \rangle$, computed within the interval $(-384, 384)$ for the high and low transcriptional plasticity genes, respectively. There are 732 genes in each group. To compute error bars, we divided each group of genes into five arbitrary subgroups, computed $\langle \Delta f \rangle$ in each of the subgroups, and computed the standard deviation of $\langle \Delta f \rangle$ between the subgroups. Error bars correspond to one standard deviation.
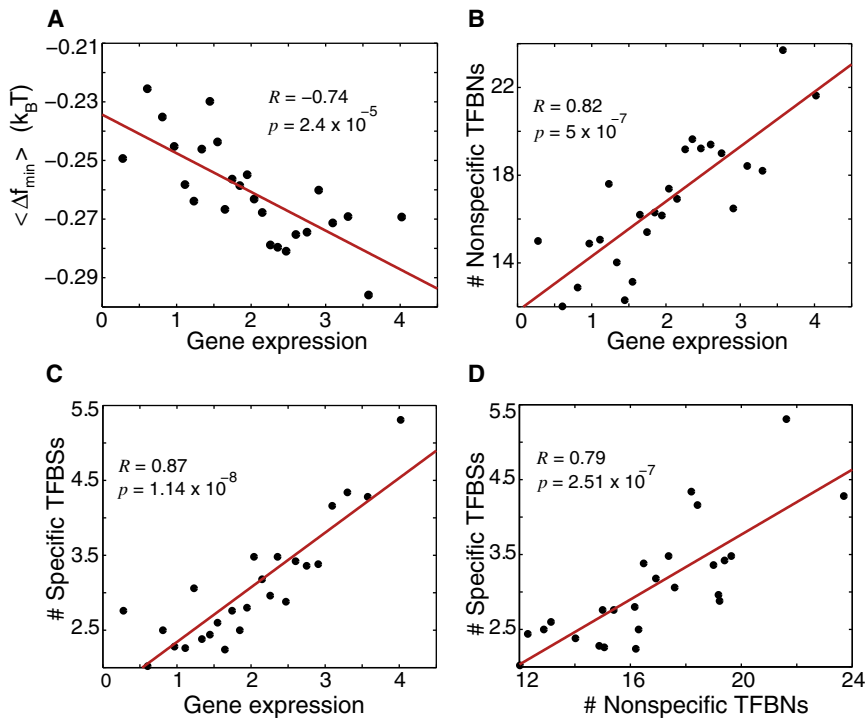
FIGURE 4 (*A*) Correlation between the minimal value of the free energy of nonspecific TF-DNA binding in the promoter region, within the interval $(-150,0)$, $\Delta f_{min} = \Delta F_{min}/M$, and the average value of gene expression within this region. All ~5000 genes were binned into 25 bins according to the level of gene expression. Each point in the graph corresponds to the average, $\langle \Delta f_{min} \rangle$, for the genes in a given bin plotted as a function of the experimentally measured average level of gene expression for the genes in this bin. (*B*) Correlation between the computed number of nonspecific TFBNs within the interval $(-150,0)$, and the level of gene expression. A given genomic coordinate is assigned to belong to nonspecific TFBN, if the average free energy of nonspecific TF-DNA binding per nucleotide is smaller than a given cutoff value, $\Delta f < -0.25 \ k_B T$. (*C*) Correlation between the number of specific TFBSs and the gene expression. The information about specific TFBSs is taken from (33). (*D*) Correlation between the number of specific TFBSs and the number nonspecific TFBNs. The binning in *B–D* is preformed as in *A*.

p-value calculations

Fig. 1, *A* and *B*: To compute the *p*-values, first, we selected 10^5 pairs of groups of randomly chosen 496 genes. Each pair of groups represents randomized analogs of the highest occupancy and the lowest occupancy genes, respectively. Second, for each of these pairs of random groups we computed the free energy of nonspecific binding, as described previously. Third, within each region of interest (TSS or UAS), we computed the difference between the minima of the average free energy of nonspecific binding, $\langle \Delta f \rangle_{min}$, for the corresponding pairs of groups. Finally, we computed the probability that this difference is equal or larger than the actual value of the difference. The latter probability was taken as the *p*-value.
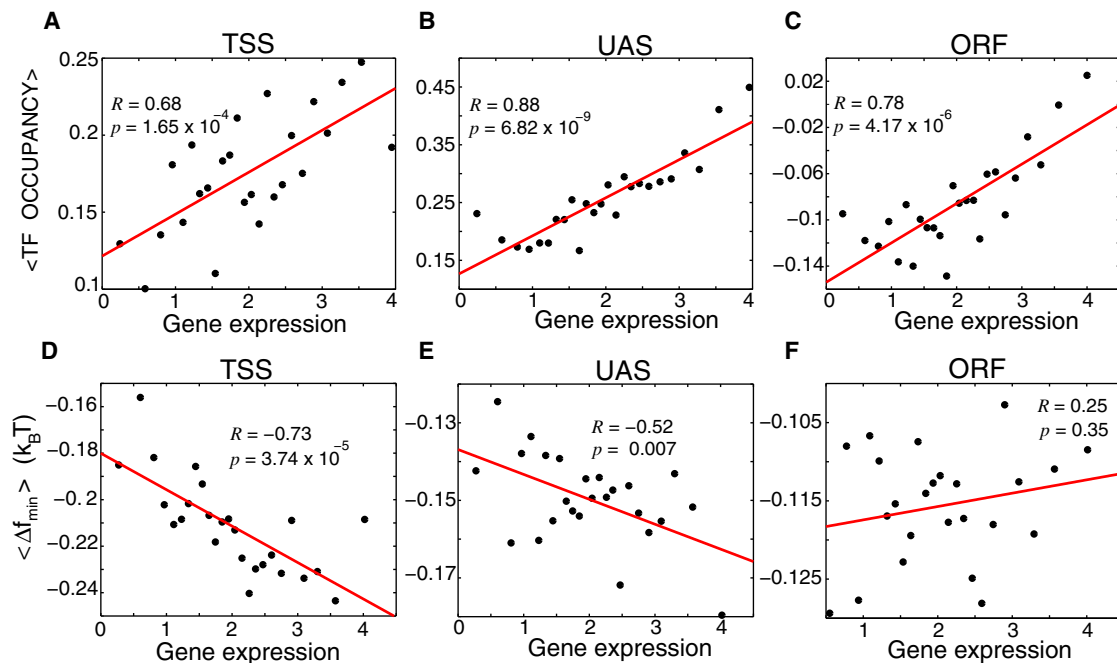

FIGURE 5 Analysis of experimental results from (1): Correlation between the average TF occupancy and the level of gene expression for TSS (*A*), UAS (*B*), and ORF (*C*) regions, respectively. Genes were binned into 25 bins according to the level of gene expression. Each point in the graph corresponds to the average, experimental TF occupancy for the genes in a given bin plotted as a function of the experimentally measured, average level of gene expression for the genes in this bin. Correlation between the computed, average value of the minimal free energy of nonspecific TF-DNA binding, $\langle \Delta f_{min} \rangle$, and the level of gene expression for TSS (*D*), UAS (*E*), and ORF (*F*) regions, respectively. The binning is performed as explained previously.

: To compute the *p*-value, we first compiled 3×10^5 pairs of groups of randomly chosen 178 and 40 genes, respectively. These groups represent the randomized analogs for TFIID and SAGA genes, respectively. Second, for each of these pairs of random groups we computed the average free energies, $\langle \Delta f \rangle$, of nonspecific binding separately for randomized TFIID and randomized SAGA groups, as described previously. Third, for each pair of randomized groups we computed the difference of the integrated free energy within the interval (−384,100) between randomized TFIID and SAGA groups. Finally, we computed the probability that this difference is equal or larger than the actual value of the difference. The latter probability was taken as the *p*-value.

RESULTS

Model free energy of nonspecific TF-DNA binding

We begin by computing the free energy of nonspecific TF-DNA binding in three genomic locations surrounding the transcription start sites of 4962 highly confident yeast transcripts (Materials and Methods). We use the following terms, adopted from (1), to describe these three types of locations: transcription start sites (TSSs) located in the interval (−90,−30), upstream activating sequences (UASs) located in the interval (−320,−260), and open reading frames (ORFs) located in the 3′ end of the coding regions of genes (Materials and Methods). The occupancy of 202 transcription regulators (we use the term, TFs, to describe the regulators) were experimentally determined in these three locations in (1). We note that we use a conventional abbreviation, TSS, to describe both the transcription start site, where zero of our coordinate system is positioned in each gene, and the region in the upstream vicinity of the TSS site, located in the interval (−90,−30), as defined in (1). This coincidence should not lead to confusion, because the precise meaning of TSS will be clear from the context in each case.

To compute the free energy of nonspecific TF-DNA binding in each genomic location specified previously, we use a simple variant of the Berg-von Hippel model (14,15), developed recently (25,26). In particular, we can assign the free energy of nonspecific TF-DNA binding to each DNA basepair along the genome in the following way. First, we position a midpoint of the sliding window of width $L = 50$ bp at a given genomic coordinate.

Second, we compute the partition function of the model TF sliding along the sliding window:

$$Z = \sum_{i=1}^{L} \exp\left(\frac{-U(i)}{k_B T}\right), \qquad (1)$$

where K_B is the Boltzmann constant, T is the temperature, and $U(i)$ is the TF-DNA binding energy at the position i within the sliding window. The TF-DNA binding energy of the TF forming M contacts with DNA basepairs, at a given position i within the sliding window:

$$U(i) = -\sum_{j=i}^{M+i-1} \sum_{\alpha=1}^{4} K_\alpha s_\alpha(j), \qquad (2)$$

where $s_\alpha(j)$ is a four-component vector of the type $(\delta_{\alpha A}, \delta_{\alpha T}, \delta_{\alpha C}, \delta_{\alpha G})$, specifying the identity of the basepair at each DNA position j, with $\delta_{\alpha\beta} = 1$, if $\alpha = \beta$, and $\delta_{\alpha\beta} = 0$, if $\alpha \neq \beta$. For example, if a given DNA site, j, is occupied by the *A* nucleotide, this vector takes the form: (1,0,0,0); if the site j is occupied by the *C* nucleotide, this vector is (0,0,1,0). Within the framework of our model, each TF is fully described by four energy parameters, K_A, K_T, K_C, and K_G (25). To model nonspecific TF-DNA binding, we generate an ensemble of 250 TFs, and for each TF we draw the energies K_A, K_T, K_C, and K_G from the Gaussian probability distributions, $P(K_\alpha)$, with zero mean and standard deviations, $\sigma_\alpha = 2$ $k_B T$, where $\alpha =$ A, T, C, G. Therefore, each random realization of $P(K_\alpha)$ describes one TF.

Third, we compute the free energy of nonspecific TF-DNA binding, $F = -k_B T \ln Z$, for each randomly generated TF in this sliding window. We always consider the difference, $\Delta F = F - F_\infty$, where F_∞ is the free energy computed for a randomized sequence of the same width, L, and averaged over 50 random realizations of this sequence, for a given TF. This normalization procedure removes the effect of the compositional bias, and allows us to compare the free energies of nonspecific TF-DNA binding in different genomic regions, despite the variation of the average nucleotide composition along the genome. We perform this calculation for all 250 randomly generated TFs. We note that the results are very weakly dependent of the sliding window width, L (data not shown).

Fourth, we move the sliding window along the genome, assigning the free energy of nonspecific TF-DNA binding for each randomly generated TF, to each genomic coordinate in steps of 4 bp, within the three regions described previously: TSSs, UASs, and ORFs, respectively. This procedure allows us to perform a direct comparison of the TF occupancy in these genomic regions between the model and experiment (1).

Nonspecific binding significantly influences experimentally observed TF-DNA binding preferences

We now seek to answer the question to what extent does nonspecific TF-DNA binding influence experimentally observed TF binding preferences within the TSS, UAS, and ORF regions, respectively? To answer this question, we first select 10% highest and 10% lowest average TF occupancy genes (see Materials and Methods for the definition of the experimentally measured, average TF occupancy). We perform such selection separately with respect to TF occupancy in the TSS, UAS, and ORF regions, respectively.

Next, we compute the profile of nonspecific TF-DNA binding free energy, within the range $(-384, 384)$, for the highest and the lowest TF occupancy genes, selected in both the TSS regions, Fig. 1 A, and UAS regions, Fig. 1 B. For each gene, we compute the free energy (normalized per bp), averaged with respect to 250 model TFs, $\Delta f = \langle \Delta F \rangle_{TF}/M$. After that, we compute the average of Δf with respect to the selected 10% highest and 10% lowest average TF occupancy genes, aligned with respect to their transcription start sites, $\langle \Delta f \rangle = \langle \langle \Delta F \rangle_{TF} \rangle_{seq}/M$, where the second average, $\langle \ldots \rangle_{seq}$, describes the averaging with respect to the aligned sequences. In both the TSS and UAS regions we observe that the highest TF occupancy genes exhibit a lower free energy of nonspecific TF-DNA binding compared with the lowest TF occupancy genes that exhibit a higher free energy. This result is statistically significant with the p-values, $p \simeq 0.01$ and $p \simeq 0.007$, for the TSS and UAS regions, respectively (Materials and Methods). A different definition of the average TF occupancy leads to similar results, Fig. S1 in the Supporting Material.

The free energy of nonspecific TF-DNA binding significantly correlates with the experimentally observed average TF occupancy within the entire dynamic range of the occupancy values in both TSS and UAS regions, Fig. 1, C and D. Here, we ordered genes in bins with respect to the value of their average TF occupancy, and computed the minimal free energy, $\min(\Delta f)$, for each sequence, in each bin. It is remarkable that in both TSS and UAS regions the linear fits exhibit identical slopes, Fig. 1, C and D. We conclude therefore that nonspecific TF-DNA binding significantly influences binding preferences in promoter regions of the majority of transcription regulators in yeast.

We note that theoretical analysis of the nonspecific TF-DNA binding free energy in the ORF regions does not show statistically significant correlation with the experimentally measured average TF occupancy in these regions, unlike the trend described previously in the TSS and UAS regions (data not shown). Overall, the magnitude of the nonspecific TF-DNA binding free energy in the ORFs regions is weak compared to the TSS and UAS regions, as Fig. 1 clearly demonstrates.

We now demonstrate that the experimentally measured cumulative TF occupancy in the promoter regions as compared to coding regions (1) is also accurately predicted within the framework of our model. In particular, to define the cumulative TF occupancy theoretically, we assume in the computational procedure that if the minimal binding free energy, $\min(\Delta f)$, of a given TF within a given genomic region for a particular gene is less than a certain cutoff value, then it binds to this region. Fig. 2, A and B, show the result of such comparison for promoter regions (combined binding to TSSs and UASs) and coding regions (ORFs), based on the theoretical calculation (Fig. 2 A) and experimental measurements (Fig. 2 B). These results are highly statistically significant, as the biological error bars

demonstrate, Fig. 2, A and B. The agreement between the theory and experiment holds quantitatively significant for a wide, physically relevant range of the free energy cutoff values (data not shown). Notably, when we compute the cumulative TF occupancy, separating promoter regions into TSSs and UASs, we observe a disagreement with the experimental data for TSSs and UASs. In particular, our model predicts that TSS regions possess a higher propensity for nonspecific TF-DNA binding than UAS regions, whereas the experimental data show an opposite trend. The reason for this disagreement is currently not understood. At least two additional factors may be responsible for the observed discrepancy. First, as we mentioned in the Introduction, several types of interactions determine the measured TF occupancy in vivo (1). These are i), direct binding to the local DNA; ii), cooperative binding to a combination of local DNA and other locally bound TFs; iii), cooperative binding only to other locally bound TFs and not to DNA; and iv), binding to nascent RNA transcripts and/or proteins bound to nascent RNA transcripts. Our current model takes into account only mechanism (i). Second, our theoretical approach is purely equilibrium, whereas kinetic barriers might significantly influence TF binding preferences in vivo (17).

We conclude, therefore, that nonspecific TF-DNA binding alone can accurately account for the experimentally observed differences in the cumulative TF occupancy of promoter regions as compared with coding regions, yet our model fails to predict the experimentally measured, absolute differences between TSSs and UASs within the promoter regions (1).

DISCUSSION AND CONCLUSION

Nonspecific binding distinguishes between SAGA-dominated and TFIID-dominated genes

Genome-wide studies found that ~90% of the yeast genome is TFIID dominated, whereas the remaining ~10% of genes are SAGA dominated (1,29,32). SAGA-dominated genes typically contain the TATA box, and they are highly regulated compared to TFIID-dominated genes, which are typically TATA-less (29,32). The majority of the known stress-response genes in yeast tend to belong to the SAGA-dominated class (28,32). It is also known that the high transcriptional plasticity genes are enriched in SAGA-dominated genes compared to the low transcriptional plasticity genes (30,31). It was concluded in a recent study by Venters et al. (1) that SAGA-dominated/TATA-containing genes were occupied by a larger variety of regulators compared to TFIID-dominated/TATA-less genes. Here, we show that the nonspecific TF-DNA binding free energy can qualitatively explain the observed difference in the TF occupancy between these two classes of genes.

In particular, we computed the profile of nonspecific TF-DNA binding free energy for 40 highly confident

SAGA-dominated (TATA-containing) genes, and 178 TFIID-dominated (TATA-less, nonribosomal protein) genes, respectively (1). The key conclusion here is that SAGA-dominated genes exhibit a wider region of the reduced free energy (within the interval $(-384,384)$ around the TSS site) compared to TFIID-dominated genes, Fig. 3 A, with the p-value, $p \simeq 4.4 \times 10^{-4}$. We suggest, therefore, that the reduced free energy of nonspecific TF-DNA binding plays the role of an effective, attractive potential that facilitates nonspecific binding to promoters of both SAGA-dominated and TFIID-dominated genes, however, the predicted effect is stronger for the former group, leading to a higher average TF occupancy of SAGA-dominated genes. To test the functional robustness of our conclusion, we selected two larger groups of 15% highest and 15% lowest transcriptional plasticity genes, respectively (732 genes in each group) (30,31) (Materials and Methods). The high-plasticity genes are enriched in SAGA-dominated genes (272 SAGA genes out of 723 high-plasticity genes) compared to the low-plasticity genes (3 SAGA genes out of 732 low-plasticity genes), with $p < 10^{-6}$. The free energy calculation performed for these two groups shows that the high-plasticity genes possess a wider region of the reduced free energy compared to the low-plasticity genes, Fig. 3 B. Therefore, we conclude quite generally that nonspecific TF-DNA binding significantly influences functional properties of yeast genes, and presumably, it facilitates the search of specific TF binding sites in promoter regions. Promoters of highly regulated genes appear to possess a wider region of the reduced nonspecific TF-DNA binding free energy (on average), compared to weakly regulated genes.

Gene expression is correlated with nonspecific TF-DNA free energy landscape

We ask now the question: How is the gene expression in yeast influenced by nonspecific TF-DNA binding? To answer this question, first, for each gene we computed the average free energy profile, $\Delta f = \langle \Delta F \rangle_{\mathrm{TF}}/M$, in the promoter region within the interval $(-150,0)$. Second, for each gene we found the minimum of Δf within this interval, $\Delta f_{\min} = \min(\Delta f)$. As a result, we observe a statistically significant correlation of $\Delta f_{\min}$ with the level of gene expression (28), Fig. 4 A. This result suggests that nonspecific TF-DNA binding influences gene expression in yeast. To obtain a deeper insight into a relationship between nonspecific TF-DNA binding and gene expression, we introduce the notion of nonspecific transcription factor binding nucleotides (TFBNs). In particular, we define a given position within the genome as being the nonspecific TFBN, if the computed average free energy of nonspecific TF-DNA binding in this genomic location is less than a certain cutoff value. The correlation between the number of nonspecific TFBNs within the interval $(-150,0)$ and the gene expression level is shown in

Fig. 4 B. Statistically significant correlation persists for a wide range of the cutoff values (data not shown). To understand how specific and nonspecific TF-DNA binding is related, as far as gene expression is concerned, we also present the correlation between the number of specific TFBS and the level of gene expression, Fig. 4, C and D, where the information about specific TFBSs is extracted from (33). We conclude, therefore, that first, the propensity of promoter regions toward nonspecific TF binding statistically significantly influences gene expression, and second, specific and nonspecific binding are functionally linked.

Next, we seek to understand whether the obtained relationship between nonspecific TF-DNA binding within the interval (-150,0) and the level of gene expression persists within the TSS, UAS, and ORF regions, respectively. To answer this question, we first present the correlation between the experimentally measured average TF occupancy in each region and the level of gene expression, Fig. 5, A, B, and C. Remarkably, a statistically significant correlation is observed in all three regions. Fig. 5, D, E, and F, present the correlation between the computed minimal free energy of nonspecific binding and the level of gene expression, within each region, TSS, UAS, and ORF, respectively. The strongest correlation is observed for the TSS regions, the correlation in the UAS regions is also significant, but contrary to experimental results, we do not observe correlation between $\Delta f_{\min}$ and the level of gene expression in the ORF regions, Fig. 5 F. We conclude, therefore, that in yeast, the strength of nonspecific TF-DNA binding is encoded and fine-tuned within a wide interval (of ~300 bp) in promoter regions, and it influences the level of gene expression. Our results suggest that in the coding regions the effect of nonspecific TF-DNA binding on gene expression is insignificant, and it is likely that other factors, such as specific ATP-dependent chromatin modifying factors, might play a dominant role there.

CONCLUSION

In summary, we first showed that nonspecific TF-DNA binding significantly influences binding preferences of ~200 transcription regulators in promoter regions of the yeast genome. Second, our analysis suggests that specific and nonspecific binding are functionally linked. Third, we observed quite generally, that promoter regions of highly regulated genes, such as SAGA-dominated genes, possess a wider region of the reduced nonspecific binding free energy compared to promoter regions of weakly regulated genes, such as TFIID-dominated genes. This qualitatively explains the experimental observation in (1) that promoters of SAGA-dominated genes are more highly occupied (on average) than promoters of TFIID-dominated genes. Fourth, we showed that the landscape of nonspecific binding free energy in promoter regions correlates with the level of gene expression.

We emphasize that to compute the nonspecific TF-DNA binding free energy genome-wide; we used a highly simplified biophysical model. Despite the simplicity of this model, we suggest that our conclusions are quite general, and most likely they represent the statistical rule, rather than the exception. The generality of our conclusions stems from the fact that the computed, location-dependent affinity of the genome for nonspecific TF-DNA binding is dominated exclusively by the symmetry of DNA sequence correlations, and this affinity is expected to be weakly dependent of microscopic details of the model.

SUPPORTING MATERIAL

A figure and a reference are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(12)00385-2.

REFERENCES

1. Venters, B. J., S. Wachi, …, B. F. Pugh. 2011. A comprehensive genomic binding map of gene and chromatin regulatory proteins in *Saccharomyces*. *Mol. Cell.* 41:480–492.

2. Ren, B., F. Robert, …, R. A. Young. 2000. Genome-wide location and function of DNA binding proteins. *Science.* 290:2306–2309.

3. Zheng, W., H. Zhao, …, M. Snyder. 2010. Genetic analysis of variation in transcription factor binding in yeast. *Nature.* 464:1187–1191.

4. Gerstein, M. B., Z. J. Lu, …, R. H. Waterston. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science.* 330:1775–1787.

5. Niu, W., Z. J. Lu, …, V. Reinke. 2011. Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans. *Genome Res.* 21:245–254.

6. Nègre, N., C. D. Brown, …, K. P. White. 2011. A cis-regulatory map of the *Drosophila* genome. *Nature.* 471:527–531.

7. Li, X. Y., S. MacArthur, …, M. D. Biggin. 2008. Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol.* 6:e27.

8. Ernst, J., P. Kheradpour, …, B. E. Bernstein. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature.* 473:43–49.

9. Riggs, A. D., S. Bourgeois, and M. Cohn. 1970. The lac repressor-operator interaction. 3. Kinetic studies. *J. Mol. Biol.* 53:401–417.

10. Hinkle, D. C., and M. J. Chamberlin. 1972. Studies of the binding of *Escherichia coli* RNA polymerase to DNA. I. The role of sigma subunit in site selection. *J. Mol. Biol.* 70:157–185.

11. von Hippel, P. H., A. Revzin, …, A. C. Wang. 1974. Non-specific DNA binding of genome regulating proteins as a biological control mechanism: I. The lac operon: equilibrium aspects. *Proc. Natl. Acad. Sci. USA.* 71:4808–4812.

12. 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.

13. von Hippel, P. H., and O. G. Berg. 1986. On the specificity of DNA-protein interactions. *Proc. Natl. Acad. Sci. USA.* 83:1608–1612.

14. von Hippel, P. H., and O. G. Berg. 1989. Facilitated target location in biological systems. *J. Biol. Chem.* 264:675–678.

15. 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.

16. Richter, P. H., and M. Eigen. 1974. Diffusion controlled reaction rates in spheroidal geometry. Application to repressor—operator association and membrane bound enzymes. *Biophys. Chem.* 2:255–263.

17. 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.

18. Liebesny, P., S. Goyal, …, L. Finzi. 2010. Determination of the number of proteins bound non-specifically to DNA. *J. Phys. Condens. Matter.* 22:414104.

19. Zurla, C., C. Manzo, …, L. Finzi. 2009. Direct demonstration and quantification of long-range DNA looping by the lambda bacteriophage repressor. *Nucleic Acids Res.* 37:2789–2795.

20. 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.

21. Blainey, P. C., G. Luo, …, X. S. Xie. 2009. Nonspecifically bound proteins spin while diffusing along DNA. *Nat. Struct. Mol. Biol.* 16:1224–1229.

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

23. Rhee, H. S., and B. F. Pugh. 2012. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature.* 483:295–301.

24. Wunderlich, Z., and L. A. Mirny. 2009. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet.* 25:434–440.

25. Sela, I., and D. B. Lukatsky. 2011. DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. *Biophys. J.* 101:160–166.

26. Afek, A., I. Sela, …, D. B. Lukatsky. 2011. Nonspecific transcription-factor-DNA binding influences nucleosome occupancy in yeast. *Biophys. J.* 101:2465–2475.

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

28. David, L., W. Huber, …, L. M. Steinmetz. 2006. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. USA.* 103:5320–5325.

29. Huisinga, K. L., and B. F. Pugh. 2004. A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in *Saccharomyces cerevisiae*. *Mol. Cell.* 13:573–585.

30. Ihmels, J., G. Friedlander, …, N. Barkai. 2002. Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* 31:370–377.

31. Tirosh, I., and N. Barkai. 2008. Two strategies for gene regulation by promoter nucleosomes. *Genome Res.* 18:1084–1091.

32. Basehoar, A. D., S. J. Zanton, and B. F. Pugh. 2004. Identification and distinct regulation of yeast TATA box-containing genes. *Cell.* 116:699–709.

33. MacIsaac, K. D., T. Wang, …, E. Fraenkel. 2006. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics.* 7:113.