MIT PHYSICS 8.592 (2007)

Effects of nucleosome positioning on eukaryotic transcriptional regulation

T.L. To
Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139
(Received 17 May 2007)

In eukaryotes, protein binding sites on DNA are wrapped in nucleosomes. As such, there exists a thermodynamic equilibrium between a specific DNA sequence and its regulatory proteins that compete with nucleosomes for occupancy. This thermodynamic equilibrium has a strong implication in transcriptional regulation in which RNA polymerase (RNAP) and transcription factors (TF) bind to their specific DNA target sites. We propose a thermodynamic model to study the competition in occupancy of specific DNA sequences between histones, RNAP and TF. In order to parameterize our thermodynamic model with nucleosomal parameters, we utilize a PSSM based probabilistic model constructed from nucleosomal information in *S. Cerevisiae*, which shows that nucleosome interacts with DNA in a sequence specific manner. Together, our model shows that nucleosome positioning indeed has a profound effect on determining the equilibrium for gene activation.

I. INTRODUCTION

Eukaryotic genomic DNA is efficiently organized into chromatin [1]. The basic unit of chromatin, the nucleosome, is a 10×6nm sized spool composed of 147 base pairs of DNA tightly wrapped around an octamer of histone proteins. The interface between DNA and histone is extensive as 142 hydrogen bonds are formed between DNA and histone core in each nucleosome. In addition, hydrophobic interactions and salt linkages are also important in holding DNA and histone protein together. The core histones are rich in lysine and arginine and their positive charges can effectively neutralize the negatively charged DNA backbone. With all these intricate biochemical interactions, extremely long stretches of DNA with enormous information content are able to be packed into a single tiny cell and perform a huge variety of functions.

For the most part, a cell regulates its functions through transcriptional control. An important step in eukaryotic transcriptional regulation is the binding of a regulatory protein to a specific DNA target site. Even when a transcription factor is present, binding sites in wrapped nucleosomes are inaccessible and the gene remains silence. This conformation results in a thermodynamic equilibrium between the target DNA sequence and its site-specific transcription factors that compete with nucleosomes for occupancy [2]. In this view, it seems important to know if nucleosome positions are regulated in *cis* by their intrinsic sequence preferences, which would then have significant transcription regulatory roles. In fact, *in vitro* studies [3] have shown that the range of histone-DNA affinities is 1,000-fold or greater, further supporting the notion that the nucleosomes have substantial sequence preferences. From the mechanics point of view, DNA sequences differ greatly in their ability to bend sharply [4, 5]. As such, the ability of the histone octamer to wrap differing DNA sequences into nucleosome is highly dependent on the specific DNA sequence. One may question if such sequence preferences are used to determine the ability of DNA binding proteins to access a particular binding sites. For example, the nucleosome positioning may be involved in site specific recruitment of transcription factors towards a particular site, while preventing irrelevant binders from cryptically initiating transcription of a silent gene [6]. In a recently published work [7], Segal and co-workers performed a genome-wide isolation of nucleosome-bound sequences and proposed a probabilistic nucleosome-DNA interaction model. The authors determined the "grammars" *S. Cerevisiae* used in encoding nucleosomal information. Not surprisingly, it was found that the transcriptional start sites of many genes had low nucleosomal occupancy. A similar study, performed by Yuan and co-workers [8], also revealed that promoter sequences in *S. Cerevisiae* were mostly located at nucleosome-free regions of the chromatin. Together, it has become increasingly clear that the

nucleosomes are indeed major players in eukaryotic transcription.

Nevertheless, a fundamental model that describes the biophysical basis of nucleosomal positioning is still lacking. The simplest idea for constructing such model is to consider DNA binding proteins that influence nucleosome positioning by competing for DNA occupancy. At equilibrium, this competition depends on the concentration and sequence specificities of both the DNA binding proteins and nucleosomes. Apparently, the DNA binding proteins have high binding specificity but are present at low concentrations, whereas the histones have lower binding specificity but are present at high concentrations. In order to understand the effect of nucleosome positioning on transcriptional regulation, one should first seek to determine whether histones can compete thermodynamically with regulatory proteins for binding sites before and after regulatory complexes are established. In light of this, the objective of this work is to develop a simple statistical mechanical model that allows us to quantify the influence of nucleosomes in gene regulation at equilibrium.

II. THERMODYNAMIC MODEL

The central modeling concept employed here, pioneered by Von Hippel and Berg [9], is that the expression level of the gene of interest can be deduced solely by considering the equilibrium probability that its associated DNA is occupied by regulatory proteins. This modeling methodology has been further extended to the studies of simple repressor-activator motifs [10, 11] and parameter space analysis [12]. In these previous studies, however, the models only describe gene regulation in prokaryotes, which do not possess nucleosomes. In other words, transcription factor (TF) binds to naked DNA, and is competed for between specific binding sites and non-specific genomic DNA. Apparently, the same analysis seems inappropriate for eukaryote. The eukaryotic genome size is of orders of magnitude greater than the genome size of prokaryotes. Using prokaryotic TF-DNA model, one would predict the occupancy of promoter to be unreasonably low due to the dominance of non-specific sequence binding. Yet in eukaryotes, DNA is wrapped around and interacting with the histone proteins, such that the non-specific sites are no longer as readily available as in prokaryotes. As a result, non-specific sites become less potent competitors for the regulatory proteins. However, it should be noted that histone proteins now compete with the DNA binding protein for the specific sites. In other words, histone proteins act as both activator and repressor for TF-DNA interactions.

We now introduce a thermodynamic model, based on equilibrium statistical mechanics, to capture the three-way interactions between specific DNA binding sites, DNA binding proteins, and histones. We follow the approach developed by Von Hippel and Berg [9], by making the following key assumptions:

I. The system is at equilibrium.

II. The level of gene expression is proportional to the probability that RNA polymerase (RNAP) occupies the promoter of interest.

III. The molecular players (RNAP, TF and histones) are either bound to their specific sites or bound non-specifically to genomic DNA. The contributions of free cytoplasmic proteins are ignored.

The overall statistical weight of states is based on both number of possible arrangement (entropic contribution) and Boltzmann weight (enthalpic contribution). The number of ways to distribution R molecules of RNA polymerase, T molecules of transcription factor, and H molecules of histone among S non-specific sites on the DNA is:

$$\Omega = \frac{S!}{R!T!H!(S-R-T-H)!} \tag{1}$$

We assume that all the aforementioned molecular players bind to DNA non-specifically with finite binding energies. Therefore, the overall statistical weights of these non-specific binding states are given by:

$$Z(R,T,H) = \frac{S!}{R!T!H!(S-R-T-H)!} e^{-\beta\left(R\varepsilon_{R,s}+T\varepsilon_{T,s}+H\varepsilon_{H,s}\right)} \tag{2}$$

where $\varepsilon_{R,s}$, $\varepsilon_{T,s}$ and $\varepsilon_{H,s}$ are the binding energies to non-specific DNA sites for RNAP, TF, and histone, respectively. Next, we consider all nine possible states of occupancy as list in Figure 1, and the total statistical weight is:

$$\begin{aligned} Z_{total}(R,T,H) = {} & Z(R,T,H) + Z(R,T,H-2)e^{-\beta\left(\varepsilon_{H,u}+\varepsilon_{H,p}\right)} \\ & + Z(R,T,H-1)e^{-\beta\varepsilon_{H,u}} + Z(R,T,H-1)e^{-\beta\varepsilon_{H,p}} \\ & + Z(R,T-1,H)e^{-\beta\varepsilon_{T,u}} + Z(R-1,T,H)e^{-\beta\varepsilon_{R,p}} \\ & + Z(R,T-1,H-1)e^{-\beta\left(\varepsilon_{T,u}+\varepsilon_{H,p}+\varepsilon_{TH}\right)} \\ & + Z(R-1,T,H-1)e^{-\beta\left(\varepsilon_{R,p}+\varepsilon_{H,u}+\varepsilon_{RH}\right)} \\ & + Z(R-1,T-1,H)e^{-\beta\left(\varepsilon_{R,p}+\varepsilon_{T,u}+\varepsilon_{RT}\right)} \end{aligned} \tag{3}$$

where $\varepsilon_{i,j}$ denotes the energy of interaction between protein i and protein binding site j, and ε_{ij} denotes the energy of interaction between protein i and protein j.

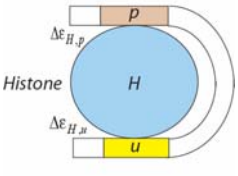
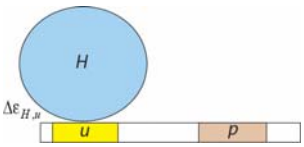
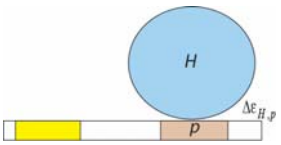
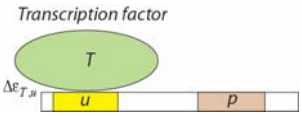
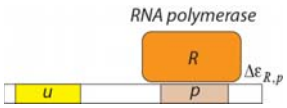
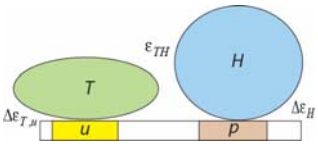
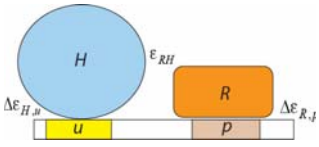
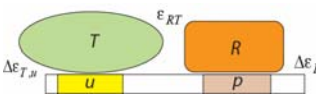
State	Normalized weight
	1
	$h^2 e^{-\beta(\Delta\varepsilon_{H,u}+\Delta\varepsilon_{H,p})}$
	$he^{-\beta\Delta\varepsilon_{H,u}}$
	$he^{-\beta\Delta\varepsilon_{H,p}}$
	$te^{-\beta\Delta\varepsilon_{T,u}}$
	$re^{-\beta\Delta\varepsilon_{R,p}}$
	$the^{-\beta(\Delta\varepsilon_{T,u}+\Delta\varepsilon_{H,p}+\varepsilon_{TH})}$
	$rhe^{-\beta(\Delta\varepsilon_{R,p}+\Delta\varepsilon_{H,u}+\varepsilon_{RH})}$
	$rte^{-\beta(\Delta\varepsilon_{R,p}+\Delta\varepsilon_{T,u}+\varepsilon_{RT})}$

FIG. 1. The 9 different states of occupancy and their corresponding statistical weights normalized by the weight of the state with unoccupied promoter and upstream regulatory sequence.

We consider the case of having one promoter and one corresponding upstream regulatory sequence for this particular promoter. In particular, we want to find the probability of RNA polymerase being bound to a specific promoter. Using the statistical weights developed above, we obtain:

$$p_{R,p} = \frac{re^{-\beta\Delta\varepsilon_{R,p}} + rhe^{-\beta(\Delta\varepsilon_{R,p}+\Delta\varepsilon_{H,u}+\varepsilon_{RH})} + rte^{-\beta(\Delta\varepsilon_{R,p}+\Delta\varepsilon_{T,u}+\varepsilon_{RT})}}{\left[\begin{array}{l} 1+\left(h^2+h\right)e^{-\beta(\Delta\varepsilon_{H,u}+\Delta\varepsilon_{H,p})} + te^{-\beta\Delta\varepsilon_{T,u}} \\ +re^{-\beta\Delta\varepsilon_{R,p}} + the^{-\beta(\Delta\varepsilon_{T,u}+\Delta\varepsilon_{H,p}+\varepsilon_{TH})} \\ +rhe^{-\beta(\Delta\varepsilon_{R,p}+\Delta\varepsilon_{H,u}+\varepsilon_{RH})} + rte^{-\beta(\Delta\varepsilon_{R,p}+\Delta\varepsilon_{T,u}+\varepsilon_{RT})} \end{array}\right]} \quad (4)$$

where $\Delta\varepsilon_{i,j} = \varepsilon_{i,j} - \varepsilon_{i,s}$, $h = H/(S\text{-}H)$, $t = T/(S\text{-}H)$ and $r = R/(S\text{-}H)$. To obtain the above expression we have assumed that $S >> R$ and T. This assumption is valid in the budding yeast *S. Cerevisiae*, in which the genome size is 10^7 base pairs, and the numbers of RNAP and TF vary from 10 − 1000.

The model has to be parameterized with biologically relevant values in interaction energies and protein numbers. The number of non-specific DNA binding site is equal to the genome size, as the discriminative energy between protein and DNA has the resolution of a single base pair. As such, we set S = 10,000,000 as we consider *S. Cerevisiae* as our *in silico* model system. It should be noted that H does not represent the number of histone molecules. Instead, H is the total number of DNA interacting sites on all available histone molecules. Each histone has 147 DNA interacting sites, as it is wrapped around by 147 bp of DNA. By assuming the reservoir of histones is the non-specifically bound molecules, we set H = 8,000,000 given that 70-90% of the genomic DNA are associated with nucleosome [1, 6]. We set the number of RNA polymerase R = 1000, in accordance with experimental evidence [2]. For the interaction energies, we fix the site-specific binding energies of RNAP and TF, so that $\Delta\varepsilon_{R,p} = -1\ k_BT$ and $\Delta\varepsilon_{T,u} = -10\ k_BT$. We assume an adhesion energy between the RNAP and TF, i.e. $\Delta\varepsilon_{RT} = -5\ k_BT$. In addition, we assume no interaction energy between RNAP and histone. As such, the key parameters to vary are the interaction energies between histone and the specific binding sites on DNA (promoter and upstream regulatory sequence), denoted by $\Delta\varepsilon_{H,p}$ and $\Delta\varepsilon_{H,u}$. Given the well known fact that biology is a k_BT science [1], we start with varying $\Delta\varepsilon_{H,p}$ and $\Delta\varepsilon_{H,u}$ between − 2 to 2 k_BT. A more heuristic approach for determining the histone-DNA interaction energy is presented in the next section.

We first examine the effect of including histones in gene regulation model. The adjustable parameters

are the histone-DNA and histone-TF interaction energies. When $\Delta\varepsilon_{H,p} = \Delta\varepsilon_{H,u} = 0$, the response curves for histone and non-histone cases are very similar, although not identical. In that case, existence of histone can only impose marginal entropic effects on the system. However, when we slightly increase or decrease the histone-DNA interaction energies, we observe strong effects on promoter occupancy by histone. Figure 2 shows the response curve in TF titration at $\Delta\varepsilon_{H,p} = \Delta\varepsilon_{H,u} = 0.2\ k_BT$. Although the saturation levels with or without histones are similar, we can see a more sensitive induction of promoter occupancy with histones in the model. It is not surprising, as the histone-DNA interaction is replaced by the more energetically favorable TF-DNA binding at equilibrium.

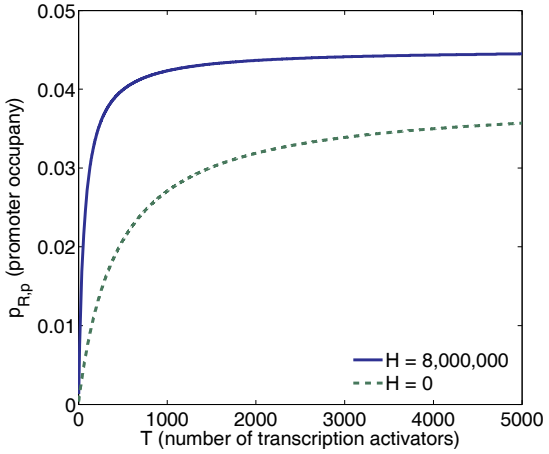


FIG 2. Probability of promoter occupancy as a function of the transcription activator level T in the presence of histones (solid line) and the absence of them (dotted line). The histone-DNA interaction energies are $\Delta\varepsilon_{H,p} = \Delta\varepsilon_{H,u} = 0.2\ k_BT$.

Next, we examine the effect of histone concentrations. Figure 3 shows the promoter occupancy as a function of histone concentrations under three histone-DNA interaction energies: $\Delta\varepsilon_{H,p} = \Delta\varepsilon_{H,u} = 0$, 0.1, or 0.2 k_BT. The TF level is kept constant at $T = 100$. In all 3 cases, we observe a maximum in promoter occupancy at a certain histone level. Maxima are also observed for negative histone-DNA energies (not shown). In general, the maximum for histone number decreases with histone-DNA interaction energies. The maximum point vanishes when both energies become highly negative ($\Delta\varepsilon_{H,p} = \Delta\varepsilon_{H,u} = -0.2\ k_BT$), and the promoter occupancy is monotonically decreasing with increasing number of histones. Apparently, the existence of maxima indicates the countering effect between enthalpy loss by TF binding (favors high histone numbers) and entropy gain by histone occupancy on non-specific sites (favors low histone numbers).

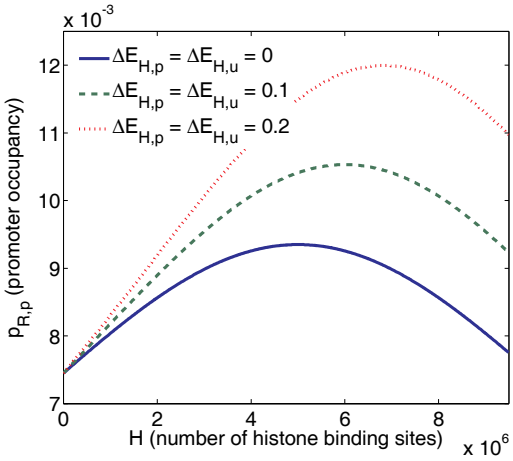


FIG 3. Probability of promoter occupancy as a function of the histone level H. The transcription factor level $T = 100$.

We examine further the effects on promoter occupancy by varying histone-DNA interacting energies. We observe a wide range of response curve behaviors even for a small range of energy values ($|\Delta\varepsilon_H| \leq 2\ k_BT$), as shown in Figure 4. In order to compare the effects on response curve more faithfully, we define the fold-change, as the ratio of promoter occupancy in the presence of TF to the occupancy in the absence of TF, i.e. $f(T) = p_{R,p}(T)/p_{R,p}(0)$. Based on our previous definition, $\Delta\varepsilon_H = 0$ when the binding site of interest has the average (over all genomic binding sites) affinity to histone. The case in which both $\Delta\varepsilon_H$ values are highly positive ($\Delta\varepsilon_H = 2\ k_BT$) may correspond to nucleosome-free promoter and upstream regulatory sequence. In that case, the dynamic range is seen at lower TF levels as compared to the base case ($\Delta\varepsilon_H = 0$). This corresponds to the experimental finding [8] that highly active genes are often found in nucleosome-free regions. A similar dynamic range can be observed when only the upstream regulatory sequence is 'naked' ($\Delta\varepsilon_{H,u} = 2$), even if the promoter is tightly wrapped ($\Delta\varepsilon_{H,p} = -2$). This corresponds to the experimentally confirmed cooperativity [13, 14], in which two proteins binding to the sites on the same nucleosome facilitate the binding of each other. However, the reverse case with naked promoter and tightly wrapped URS results in a right-shifted dynamic range (not shown in Figure 4), since promoter has a low affinity to RNAP and thus there is asymmetry in switching roles of binding between TF and RNAP. The case in which both $\Delta\varepsilon_H$ values are highly negative ($\Delta\varepsilon_H = -2\ k_BT$) may correspond

to tightly regulated genes. In such case, a significantly higher level of TF is required to induce response, and yet the fold change at saturation is considerably higher than the base case. We speculate that this type of tight control is biologically relevant, and is used in the activation of genes in response to catastrophes.

We also consider the scenario in which the bound transcription factor destabilizes histone and causes it to unbind DNA. In essence, we try to account for the fact that histone acetylase is often recruited to the upstream regulatory sequence, and that acetylated histone becomes unstable and un-wraps DNA [6, 15]. Figure 5 shows to fold-change in promoter occupancy as a function of the acetylation-driven destabilizing energy, which is represented as repulsive interaction energy between the TF and histone. Surprisingly, the fold-change is relatively small and soon reaches saturation.

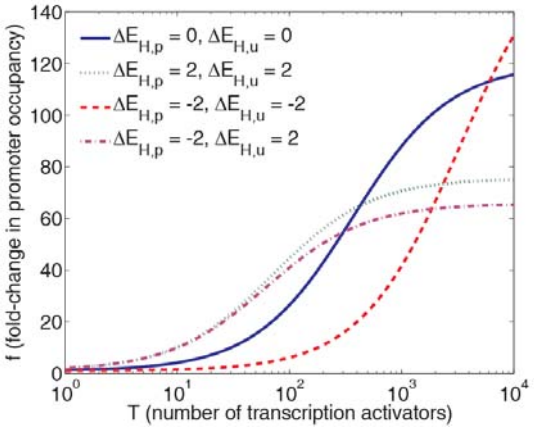


FIG 4. Fold-change of promoter occupancy (relative to the occupancy in the absence of TF) as a function of the transcription factor level T. The histone level H = 8,000,000.

Apparently, our model is robust against changes in the parameter governing TF-histone interaction ($\Delta\varepsilon_{HT}$). In other words, our current model suggests that nucleosomal positioning (through changes in histone-DNA energies) has a much more profound effect than histone destabilization, which is unlikely to be the case in reality. In fact, histone acetylation significantly changes the conformation of histone, which results in drastic changes in histone-DNA interaction energies. Currently, our model fails to capture these changes.

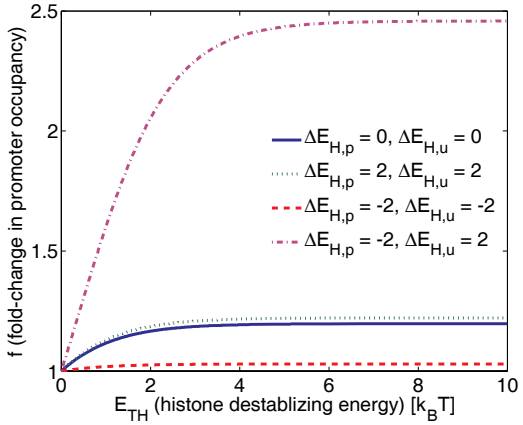


FIG 5. Fold-change of promoter occupancy (relative to no interaction between TF and histone) as a function of the TF-histone repulsive energy. The transcription factor level T = 100 and the histone level H = 8,000,000.

III. PSSM OF HISTONE-DNA INTERACTIONS

The parameters that govern histone-DNA interaction prove to be highly sensitive in our model. As such, we want parameterize our model with a reasonable prediction of histone-DNA energies. It is desirable to predict histone's affinity to a specific sequence, be it promoter or upstream regulatory sequence. The position specific scoring matrices (PSSMs) approach, pioneered by Berg and von Hippel [16], has been widely used for predicting binding affinities between DNA binding sites and regulatory proteins [12]. Given the usefulness of PSSM in studying equilibrium binding, our goal here is to construct a probabilistic model that represents the DNA sequence preferences of nucleosomes.

In fact, a probabilistic model of histone's preference on DNA sequence has been proposed by Segal and co-workers [7] from a set of 199 mononucleosome DNA sequences in *S. Cerevisiae*. The sequences are center aligned to give 147 base pair positions. A dinucleotide distribution, based on counts at positions [*i-2, i-1*], [*i-1, i*], [*i, i+1*], is associated to each position *i*. The use of dinucleotide distribution is motivated by the fact that dinucleotides are the simplest elements that can capture the sequence-dependence of DNA bending [5]. For any sequence $\boldsymbol{S}$ of length 147 bp, a probability is assigned:

$$P(\boldsymbol{S}) = P_1(S_1)\prod_{i=2}^{147} P_i(S_i \mid S_{i-1}) \qquad (5)$$

Next, these probabilities are converted into an apparent free energy landscape by taking their log

ratio to a predefined background model P_B. The apparent free energy score for a sequence $\boldsymbol{S}$ is then:

$$Score(\boldsymbol{S}) = \log\frac{P(\boldsymbol{S})}{P_B(\boldsymbol{S})} = \log\frac{P_1(S_1)}{P_B(S_1)} + \sum_{i=2}^{147}\log\frac{P_i(S_i \mid S_{i-1})}{P_B(S_i \mid S_{i-1})} \qquad (6)$$

The free energy score for a sequence $\boldsymbol{S}$ can be further extended to a thermodynamic model that can be used to predict nucleosome positions genome-wide. The central idea is that the probability of placing a nucleosome starting at a particular base pair *i* in the genome is assumed to be equal to the sum of the statistical weights of all legal configurations in which a nucleosome start at position *i*, divided by the sum of the statistical weights of all legal configurations. Calculating this probability is not straight forward and it requires the use of a dynamic programming method. The algorithm for such calculation can be found in the supplemental materials of [7].

Here, we consider a simpler case, as we focus on the local alignment and affinity of a short binding site to a histone protein. We assume the sequences flanking this binding site to have indistinguishable background affinity. The question we are interested in is how to convert the free energy of histone-DNA interaction score into a corresponding equilibrium binding energy. Fortunately, the relevant statistics and data, obtained from *S. Cerevisiae*, are available on Eran Segal research group's website (http://genie.weizmann.ac.il/). From the website we obtain the dinucleotide distributions for the central 141bp of the center alignment on histone. Using these data, we can predict the affinity of a protein-binding site, such as an upstream regulatory sequence, to the histone protein. Here, we consider a protein binding site $\boldsymbol{X}$ of $M \leq 15$ base pairs. Given 141 DNA interaction positions on the PSSM, there are 141 possible alignments between the histone and the the short sequence $\boldsymbol{X}$. For example, when the first base of the sequence is aligned with the i^{th} DNA interacting domain on the histone, the free energy score of such alignment is:

$$Score^{[i]}(\boldsymbol{X}) = \log\frac{P_i(X_1)}{P_B(X_1)} + \sum_{j=2}^{M}\log\frac{P_{i+j-1}(X_j \mid X_{j-1})}{P_B(X_j \mid X_{j-1})} \qquad (7)$$

Within the summation term on the right hand side, if *i+j-1* exceeds 141, the subscript of *P* becomes *i+j-142*, assuming periodicity. The discrimination energy E for a certain sequence is proportional to the average free energy score of the sequence over all 141 possible alignments:

$$-\lambda E(\boldsymbol{X}) = \frac{\left[\sum_{i=1}^{141}\log\frac{P_i(X_1)}{P_B(X_1)} + \sum_{i=1}^{141}\sum_{j=2}^{M}\log\frac{P_{i+j-1}(X_j \mid X_{j-1})}{P_B(X_j \mid X_{j-1})}\right]}{141} \qquad (8)$$

The parameter λ is related effective evolutionary temperature, defined as $1/k_BT$ [16]. Therefore, the binding sequence energy for a particular base sequence $\boldsymbol{X}$ can be readily found using the above formula.

We want to determine how well this formulation works on randomly generated sequences. As such, we generate 100,000 sequences with lengths varying from 8-18 base pairs using the background nucleotide distribution of *S. Cerevisiae*, such that $P_B(\mathrm{A})=P_B(\mathrm{T})=0.308512$, and $P_B(\mathrm{C})=P_B(\mathrm{G})=0.191488$. We also assume the background dinucleotide distribution to be independent, such that $P_B(\mathrm{S_i} \mid \mathrm{S_{i-1}}) = P_B(\mathrm{S_i})P_B(\mathrm{S_{i-1}})$. We notice that the mean binding energy do not change with sequence lengths, although the variance increases slightly with increasing length. Figure 6 shows the spectrum of binding energies for the random sequences on the histone protein. The distribution closely resembles a Gaussian distribution. The mean binding energies over all sequences is $\langle E \rangle = 0.2627 k_BT$ and the variance $\sigma^2 = 0.1937\ (k_BT)^2$. Using these parameters, we can describe histone-DNA interactions with a Random Energy Model (REM). The probability of having a binding energy of *E* on histone is given by:

$$p(E) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(E-\langle E\rangle\right)^2}{2\sigma^2}\right] \qquad (9)$$

(9)

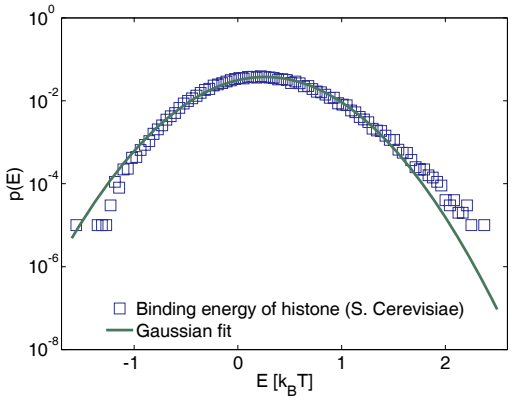


FIG. 6. Predicted binding energy of histone to random *S. Cerevisiae* DNA sequences of 8-18 base pairs using PSSM (squares). The predicted values are well fit by a Gaussian distribution (solid line).

The mean histone-DNA interaction energy is positive, suggesting the histone has sequence preference for DNA sites. In fact, natural nucleosomes have an abundance of distinctive sequence motifs [7, 8]. These motifs, such as periodic AA/TT/TA − GC dinucleotides, are known to facilitate the sharp bending of DNA around the nucleosome [5]. Our PSSM prediction for the sequence 'AAGCAACG' is −0.7357 k_BT, which is substantially lower than that a random sequence. On the other hand, the sequence 'TATA' has a high PSSM predicted discriminative energy of 0.6276 k_BT. As eukaryotic RNA polymerases are known to bind the 'TATA' box region of the promoter, the high positive discriminative energy of 'TATA' suggests that promoters have the tendency to be placed at the nucleosome-free regions in the chromatin. In fact, experimental evidence [8] supports this notion.

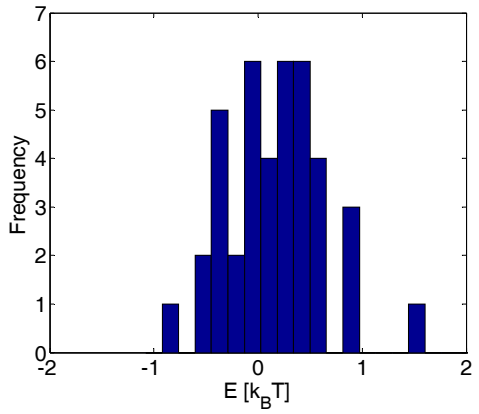


FIG. 7. Histogram showing of distribution of histone binding energies for 40 actual protein binding sites in *S. Cerevisiae*. The mean binding energy is 0.1655k_BT.

To further validate our PSSM model, we obtain the sequences of 40 actual *S. Cerevisiae* protein binding sites, ranging from 7 − 15 bp in length. The majority of these sites are specific binding sites for transcription factors. Figure 7 shows the distribution of PSSM prediction on the 40 binding sites. The average binding energy of histone to these protein binding sites is 0.1655k_BT, which is 40% lower than the average binding energy of the random sequences. Although we must analyze more actual protein binding sites before making any further statistics-based arguments, it is clear that TF binding sites exhibit a broad range of affinities to histone. This broad range of affinities may have strong implication in the control of eukaryotic transcription. Unlike its counterpart in prokaryote, eukaryotic RNA polymerase is not substantially stable on promoter, and thus the basal transcription rate is often low. Therefore, even though the 'TATA' region of the promoter tends to be nucleosome-free, this feat alone does not guarantee efficient transcription. As such, binding of transcription factor to the upstream regulatory sequence seems to determine the equilibrium of gene activation, although it may not be the rate determining step from the kinetic point of view.

IV. CONCLUSION AND FUTURE DIRECTIONS

Regulating TF-DNA interaction and thus gene activation by nucleosome positioning appears to be a clever idea in biology, as this regulatory mechanism does not require much additional material and energy. The mechanism, roles, and evolution of nucleosome positioning are certainly intriguing problems in biology. Realization of nucleosome positioning may also provide a great forward leap in genetic engineering and medical research. Nevertheless, a quantitative understanding of how nucleosome positioning affects transcription is still lacking. Motivated by this reason, we have developed a simple thermodynamic model on eukaryotic transcription regulation, and have shown that nucleosome indeed plays a profound role in setting to equilibrium for the occupation of transcription machinery on the target gene. In addition, by applying a probabilistic model constructed from nucleosomal information in *S. Cerevisiae*, we have shown that nucleosome interacts with DNA in a sequence specific manner. This sequence-specific energy model allows us to parameterize our thermodynamic model with "nucleosomal parameters", further facilitating our modeling effort.

Nevertheless, the chromatin is highly dynamic [6, 15], and it is not completely known how regulatory proteins gain access to their DNA target sequences in chromatin. It is generally believed that eukaryotic cells contain chromatin remodeling complexes, protein machines that use the energy of ATP hydrolysis to change the structure of nucleosomes temporarily so that DNA becomes less tightly bound to the histone core. Previous experimental results [17, 18] suggest that nucleosomes transiently expose stretches of their DNA alternative positions rapidly, which can significantly change the equilibrium of protein-DNA interaction. In addition, dynamic behavior of nucleosomes has the additional consequence of providing a new mechanism for the cooperative or synergistic binding of protein [13, 14, 19]. More

precisely, it is speculated that the binding of one or more regulatory proteins to target sites within a nucleosome *in vivo* may displace the histone octamer through sliding process or even completely displace the histone from the DNA [15, 20]. In other words, regulatory proteins occupy nucleosome-free region so as to provide the initial step in nucleosome sliding. Together, the dynamic equilibrium conformational transition in nucleosomes is an interesting topic for further investigation. Undoubtedly, statistical mechanics will remain as the central analytical tool for attacking these arising problems.

---

[1] B. Alberts *et al, Molecular Biology of the Cell* (Garland, New York, 2002).

[2] M. Ptashne and A. Gann, Nature 386, 569 (1997).

[3] Thastrom, P. T. Lowary, and J. Widom, Methods 33, 33 (2004).

[4] S. Mihardja, A. J. Spakowitz, Y. Zhang, and C. Bustamante, Proc. Natl. Acad. Sci. U. S. A. 103, 15871 (2006).

[5] H. G. Garcia, P. Grayson, L. Han, M. Inamdar, J. Kondev, P. C. Nelson, R. Phillips, J. Widom, and P. A. Wiggins, Biopolymers 85, 115 (2007).

[6] B. Li, M. Carey, and J. L. Workman, Cell 128, 707 (2007).

[7] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. K. Moore, J. P. Wang, and J. Widom, Nature 442, 772 (2006).

[8] G. C. Yuan, Y. J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando, Science 309, 626 (2005).

[9] P. H. von Hippel and O. G. Berg, Proc. Natl. Acad. Sci. U. S. A. 83, 1608 (1986).

[10] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips, Curr. Opin. Genet. Dev. 15, 125 (2005).

[11] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, Curr. Opin. Genet. Dev. 15, 116 (2005).

[12] U. Gerland, J. D. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. U. S. A. 99, 12015 (2002).

[13] K. J. Polach and J. Widom, J. Mol. Biol. 258, 800 (1996).

[14] H. Y. Fan, X. He, R. E. Kingston, and G. J. Narlikar, Mol. Cell 11, 1311 (2003).

[15] G. J. Narlikar, H. Y. Fan, and R. E. Kingston, Cell 108, 475 (2002).

[16] O. G. Berg and P. H. von Hippel, J. Mol. Biol. 193, 723 (1987).

[17] G. Li, M. Levitus, C. Bustamante, and J. Widom, Nat. Struct. Mol. Biol. 12, 46 (2005).

[18] G. Li and J. Widom, Nat. Struct. Mol. Biol. 11, 763 (2004).

[19] K. J. Polach and J. Widom, J. Mol. Biol. 254, 130 (1995).

[20] Y. Zhang, C. L. Smith, A. Saha, S. W. Grill, S. Mihardja, S. B. Smith, B. R. Cairns, C. L. Peterson, and C. Bustamante, Mol. Cell 24, 559 (2006).