



From DNA sequence to transcriptional behaviour: a quantitative approach

Eran Segal* and Jonathan Widom†

Abstract | Complex transcriptional behaviours are encoded in the DNA sequences of gene regulatory regions. Advances in our understanding of these behaviours have been recently gained through quantitative models that describe how molecules such as transcription factors and nucleosomes interact with genomic sequences. An emerging view is that every regulatory sequence is associated with a unique binding affinity landscape for each molecule and, consequently, with a unique set of molecule-binding configurations and transcriptional outputs. We present a quantitative framework based on existing methods that unifies these ideas. This framework explains many experimental observations regarding the binding patterns of factors and nucleosomes and the dynamics of transcriptional activation. It can also be used to model more complex phenomena such as transcriptional noise and the evolution of transcriptional regulation.

Nucleosome

The basic unit of chromatin, which contains 147 bp of DNA wrapped around a histone protein octamer.

Many cellular and organismal processes depend on the establishment of complex patterns of gene expression at precise times and spatial locations, and inaccuracies in carrying out such transcriptional programmes are often deleterious and lead to disease. The information for directing these complex expression patterns is encoded in regulatory DNA sequences; for example, reporter genes attached directly to regulatory sequences adopt the expression pattern of the endogenous gene¹⁻³, and when an entire human chromosome is transferred into mice, its DNA-binding and gene expression patterns remain almost unchanged⁴.

Given the central role of transcriptional programmes in many biological processes, a predictive and quantitative understanding of the transcriptional behaviours encoded by DNA sequences is desirable. Such an understanding would allow us to go beyond merely identifying the transcription factors and regulatory DNA elements that are involved in a biological process, and would replace the existing qualitative and phenomenological descriptions with a mechanistic view of the process that integrates the components that are involved into realistic mechanistic models. Indeed, our ability to quantitatively predict the behaviour of a regulatory system is a useful objective measure of the extent to which we understand how the system works. At a more practical level, the ability to accurately predict transcriptional behaviours from DNA sequences should allow us to predict the effect that sequence variation among individuals in the population

has on gene expression and thus on more complex phenotypes and disease. It would also allow the improved rational design of transgenes for biotechnology and gene therapy.

Recent work has substantially advanced our understanding of how genomic sequences are translated into transcriptional outputs. Progress has been made possible by the availability of large amounts of data on gene regulation, and through the development of quantitative models that explain how molecules such as transcription factors⁵⁻⁷ and nucleosomes^{8,9} bind DNA sequences and how these binding events produce expression patterns^{10,11}. In this Review, we unify these studies into a conceptual framework, based on existing methods, that quantitatively models the process of transcriptional regulation. The framework is founded on the idea that transcriptional regulation can be explained by an 'equilibrium competition' between nucleosomes and other DNA-binding proteins. The details of this competition are specified by every regulatory DNA sequence through the unique binding affinity 'landscape' that every sequence defines for each molecule. Each transcription factor or nucleosome 'views' every regulatory sequence in a unique way, depending on its recognition specificity; at any given set of concentrations of DNA-binding molecules, the range of affinities that the molecules have for any sequence (the binding affinity landscape) dictates the cooperative and competitive binding interactions between the DNA-binding

*Department of Computer Science and Applied Mathematics and Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel.

†Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, 2205 Tech Drive, Evanston, Illinois 60208-3500, USA.

e-mails:

eran.segal@weizmann.ac.il;
j-widom@northwestern.edu

doi:10.1038/nrg2591

Published online 9 June 2009

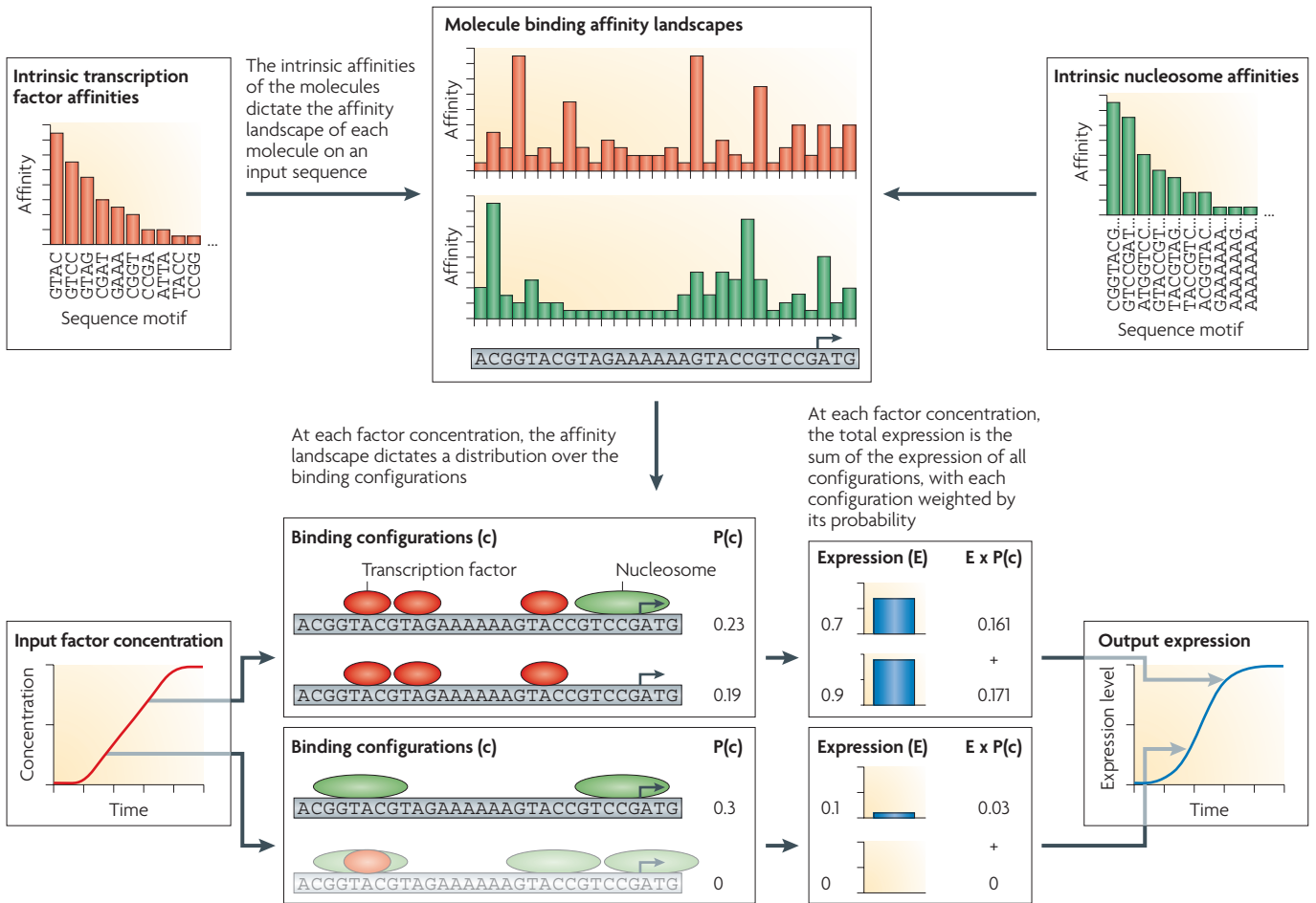


Figure 1 | Overview of quantitative models for computing expression from DNA sequences. Flow diagram of the computational approach for a simplified regulatory sequence with nucleosomes and one transcription factor as the input binding molecules. Each of the input molecules has intrinsic binding affinities for every possible sequence of length k (top panels, left and right), in which k is the number of base pairs recognized by the binding molecule. These intrinsic molecule affinities dictate how every DNA sequence is 'translated' into a unique binding affinity landscape for each molecule along the sequence (top panel, centre). For each factor concentration (c ; bottom panel, left), the model uses these binding affinity landscapes to compute a probability (P) distribution over configurations of bound molecules (see BOX 1 for details); a small subset of these configurations is illustrated (bottom panel, centre). Configurations in which two bound molecules overlap are not allowed owing to steric hindrance constraints, thereby modelling binding competition between molecules (see the bottom configuration, which has a probability of zero). Finally, each configuration results in a particular transcriptional output (bottom panel, right); the final expression is then the sum of the expression contribution of each configuration, weighted by their probability.

molecules and the DNA sequences. This unique binding affinity landscape leads to a distinct distribution of molecule binding configurations for a particular sequence and, consequently, to a distinct transcriptional behaviour for any given combination of a DNA sequence and binding molecule concentrations (FIG. 1).

As transcriptional regulation across different organisms uses the same types of molecules, which interact according to the universal laws of physical chemistry, the basic rules of this framework apply broadly. Indeed, different aspects of the approach presented here were shown in bacteria¹¹⁻¹⁴, yeast^{7,8,15-17}, flies¹⁰ and mammals^{18,19}.

We start by reviewing the substantial progress that has been made in understanding the intrinsic affinities of various molecules for DNA. This progress has been

achieved using experiments that directly measure the binding affinity landscapes of different types of molecule and computational models that identify the sequence rules that underlie and predict these affinity landscapes across several organisms. We then present different models that aim to connect these affinity landscapes to molecule-binding configurations and transcriptional outputs. Less is known about the mapping of binding affinity landscapes to transcriptional outputs, and we therefore highlight the areas in which crucial information is needed. We show that the framework presented here explains a broad range of experimental observations related to transcriptional regulation, including the binding patterns of transcription factors and nucleosomes and the dynamics of transcriptional activation.

Binding configuration
A particular arrangement of molecules along a DNA sequence, which includes specification of the precise position and orientation (or DNA strand) at which each molecule is bound.

Transcriptional noise

The variability in the transcription rate (or in steady state mRNA levels) of genes across different cells from an isogenic cell population grown in the same conditions.

Chromatin remodeller

A protein or protein complex that has the capacity to alter the structure of chromatin. Some remodellers require ATP hydrolysis for their activity.

Footprinting

A method for detecting protein–DNA interactions by using an enzyme to cut DNA, followed by analysis of the resulting cleavage pattern. The method is based on the fact that a protein bound to DNA protects that DNA from enzymatic cleavage.

Gel-shift analysis

A technique that uses native gel electrophoresis to determine whether, and how tightly, a protein of interest can bind a given DNA sequence.

Southwestern blotting

A method that involves identifying DNA-binding proteins after SDS–PAGE and transfer to a membrane using their ability to bind to specific oligonucleotide probes.

SELEX

(Systematic evolution of ligands by exponential enrichment). A combinatorial technique for producing DNAs that bind specifically and with high affinity to a DNA-binding protein of interest.

ChIP–chip

A technique (also known as ChIP-on-chip) that combines chromatin immunoprecipitation (ChIP) with microarray technology (chip). It is a high-throughput method for identifying, on a genome-wide scale, DNA regions that are bound *in vivo* by a target protein of interest.

ChIP–seq

A similar technique to ChIP–chip, but the resulting interactions are read out by high-throughput parallel sequencing and not by microarrays as in ChIP–chip.

We end by discussing how these models can be used to understand more complex features of transcriptional regulation such as transcriptional noise and expression divergence across evolution. The models presented here thus provide a concrete framework for understanding transcriptional behaviour from DNA sequence.

Binding affinity landscapes

Nucleosomes. Measurements carried out over the past two decades show that the sequence-dependent affinities of nucleosomes²⁰ for different DNA sequences can vary enormously, with the greatest affinities being at least 5,000-fold stronger than the weakest^{21,22}. These differing affinities are thought to reflect the energetic cost of sharply bending different DNA sequences around the histone octamer to make them conform to the nucleosome structure²⁰. As it is not feasible to directly measure the nucleosome affinities of all of the possible 147 bp sequences, approaches for comprehensively characterizing nucleosome affinities are based on computational models that generalize from a manageable number of nucleosome affinity measurements. Early characterizations of nucleosome sequence preferences showed that there were ~10 bp periodicities of specific dinucleotides along the length of the nucleosome. These periodicities were first observed in alignments of ~200 *in vivo* nucleosome sequences from chickens²³ and confirmed in similarly sized collections of sequences from yeast^{8,24}, worms²⁵, flies²⁶ and humans²⁶. These dinucleotide periodicities formed the basis of earlier models for predicting the binding affinity landscape of nucleosomes^{8,9}. More recently, the availability of genome-wide measurements of nucleosome occupancy has allowed researchers to identify longer sequence motifs that are generally favoured or disfavoured by nucleosomes, regardless of their position in the nucleosome. Incorporating these motifs into models that describe nucleosome sequence preferences has substantially improved predictions of nucleosome-binding affinity landscapes^{26–29}.

All of these models were based on measurements of *in vivo* nucleosomes, the positions of which are determined by multiple factors, including transcription factors³⁰, chromatin remodellers³¹, transcription³², DNA replication^{33,34} and the sequence preferences of the nucleosomes^{8,9,15,23,27,28}. A general question has therefore been the extent to which these models truly represent nucleosome sequence preferences alone, as opposed to capturing the sequence preferences of nucleosomes as well as other factors. A recent study addressed this issue by measuring the genome-wide occupancy of nucleosomes assembled on purified yeast genomic DNA¹⁵. The resulting map, in which the positions of the nucleosomes are governed only by their intrinsic sequence preferences, provides a direct experimental measurement of the binding affinity landscape of nucleosomes. A computational model constructed from these data predicted the experimentally measured affinity landscape with a high per-bp correlation of 0.89 and thus allows us to predict binding affinity landscapes of nucleosomes from DNA sequence alone. Moreover, the nucleosome organization predicted by this model matches many aspects

of the *in vivo* nucleosome organization in both yeast and worms, showing that nucleosome sequence preferences are a dominant determinant of *in vivo* nucleosome organization¹⁵.

Transcription factors. Compared with nucleosomes, transcription factors recognize and bind much shorter stretches of DNA (typically 5–15 bp). It should therefore be theoretically feasible to directly measure the binding affinity of a given factor for most, if not all, possible recognition sequences. Earlier methods based on footprinting, gel-shift analysis, southwestern blotting, SELEX (systematic evolution of ligands by exponential enrichment) and reporter constructs could only measure the affinity of factors for a small number of sequences. Substantial progress was recently achieved with the use of high-throughput technologies such as ChIP–chip¹ and ChIP–seq³⁵, which measure all of the *in vivo* bound targets of a given factor. However, the genomic regions identified by these methods are typically hundreds of base pairs, and so identifying the much shorter sequence motifs that are common to the bound regions still requires postprocessing computation³⁶.

Another limitation of using *in vivo* data is that, as in the case of nucleosomes, the derived binding specificities might also reflect the specificities of other factors. Here too, high-throughput *in vitro* methods such as protein-binding microarrays^{37,38} and microfluidic platforms³⁹, in which binding is measured across all possible ~8–10 bp sequences and is governed only by the intrinsic sequence preferences of a factor, are being used to measure the sequence specificities of DNA-binding factors. Protein-binding microarrays were recently applied to derive the binding specificities for 168 homeodomain transcription factors from mouse⁴⁰ and for 112 transcription factors from yeast^{41,42}. Although transforming the resulting microarray intensities into binding affinities is not a trivial problem, the binding affinity of many factors for any location on any DNA sequence can now be accurately characterized.

Comparing affinity landscapes: nucleosomes versus transcription factors.

The intrinsic nucleosome affinity landscape explains many aspects of the binding patterns of nucleosomes *in vivo*¹⁵. By contrast, the sequence specificity of many transcription factors is low compared with the size of the genome on which they act, such that canonical binding sites for such factors will occur multiple times across the genome by chance. For example, a factor that has a total binding specificity of 5 bp probably has over one million canonical binding sites in the human genome; however, the number of molecules of that factor present in the cell might typically be only one-tenth to one-thousandth of the number of binding sites. Thus, although a minimal level of binding affinity is a prerequisite for factor binding, most binding sites that meet such criteria occur by chance and are not bound *in vivo*. Consequently, the binding affinity landscape of most factors is a poor predictor of their *in vivo* bound locations. Nevertheless, as we discuss below, knowing the binding specificities of transcription factors, together

with other information, especially the clustering of transcription factor-binding sites and the relation of these binding sites to the nucleosome landscape, allows us to integrate factor specificities together with nucleosomes into models that accurately predict transcriptional regulation.

From affinity landscapes to configurations

Binding affinity landscapes describe how each molecule translates an input DNA sequence into a binding potential that is specific to that molecule. The next step in decoding the transcriptional behaviour of a regulatory sequence is to understand the configurations of molecules that are bound to the sequence. Several quantitative frameworks^{5,10,43} have addressed this problem. These models consider all possible configurations of molecules on the input sequence. They then associate a statistical weight with each configuration, which is computed from the concentration of the participating molecules and the strength (affinities) of the binding sites that they occupy in the configuration. The probability of each configuration can then be computed exactly by dividing the statistical weight of the configuration by the partition function, which is equal to the sum of the statistical weight of all possible configurations (BOX 1).

Such frameworks model several important aspects of the binding process. First, by allowing molecules to bind anywhere along the input sequence, the entire range of affinities is considered, thereby allowing contributions from both strong and weak binding sites^{16,44}. Second, the binding sites of any two molecules are not allowed to overlap in the same configuration, and thus the binding competition between molecules that results from steric hindrance constraints is explicitly modelled. Third, conventional cooperative binding interactions can be explicitly modelled by assigning higher statistical weights to configurations in which two molecules are bound in close proximity¹⁰. Fourth, the cooperativity that arises between factors when both nucleosomes and factors are integrated^{10,45} is captured automatically.

A noteworthy consequence of these frameworks is that, in a system comprised of both factors and nucleosomes, the locations at which nucleosomes intrinsically 'want' to bind influences the locations at which factors will be bound; conversely, the combination of factors trying to bind influences where the nucleosomes will be bound and nucleosome occupancy. The DNA sequence defines the outcome of this competition, which will change in response to the changing combinations of active transcription factors that are induced, for example, by cellular signalling during development or in response to a change in the environment.

Applications and limitations. The models above were used to identify regulatory sequences in flies⁵ and transcription factor–target interactions in humans¹⁹, and for predicting expression patterns in the fly embryo¹⁰ and in yeast¹⁶ from regulatory sequences. For example, in the segmentation gene network of the fly embryo, a model predicted that cooperative interactions and contributions from both strong and weak sites are important for

generating expression patterns¹⁰. These predictions were later supported by large-scale measurements of transcription factor binding in the fly embryo, which showed prevalent transcription factor binding to weak sites⁴⁶. Despite these successes, several aspects of the modelling framework are not well understood, such as the effects of higher-order chromatin structure, the way in which factors compete with each other and with nucleosomes, and the mechanism and quantitative magnitude of the resulting cooperative interactions.

The assumption of binding equilibrium. The models above assume that molecules bind at thermodynamic equilibrium, such that the probability of any DNA-binding configuration is simply its equilibrium probability, which is equal to the statistical weight of the configuration divided by the partition function. The success of quantitative modelling and prediction of gene regulation in prokaryotes^{13,14} has rested largely on this equilibrium hypothesis^{11,47}. However, it remains unclear how, and even whether, regulatory systems equilibrate. The equilibrium question is an especially daunting problem in eukaryotes, owing to added complexities such as nucleosomes. However, models based on an assumption of equilibrium competition also have high predictive value in eukaryotes^{10,16}, in which ATP-dependent nucleosome remodelling mechanisms might facilitate or subvert the binding equilibrium (BOX 2).

From binding configurations to transcription

The final step in modelling transcriptional behaviour from DNA sequences is to understand the transcriptional output that results from each configuration of bound molecules. This should be simple: configurations with bound activators should recruit the transcription initiation machinery and result in high transcription rates, whereas configurations with bound repressors should result in low transcription rates. Indeed, current approaches for translating binding configuration to transcriptional output are based on this rationale: they model transcriptional output as proportional to the binding probability of the transcription initiation machinery^{11,16} or proportional to the weighted sum of the bound molecules, in which activators have positive weights and repressors have negative weights¹⁰. However, these oversimplified approaches do not model the effects of architectural features of the configuration, such as DNA looping, the location and orientation of bound transcription factors relative to a nucleosome and the distance of factors from the transcription start site^{48,49}. Current models also assume that once regulatory regions are in transcriptionally active configurations their rates of transcription will be the same. This assumption ignores the additional layers of regulation that are allowed by *trans*-acting factors, which include any regulation that the transcriptional initiation complex undergoes after it is bound^{50,51}, regulation of transcriptional elongation and the effects that nucleosomes positioned in the transcript might have on the elongation process. The quantitative details of these additional effects are poorly understood and,

Protein-binding microarray

A method that allows the high-throughput characterization of the *in vitro* DNA-binding site sequence specificities of transcription factors. In this approach, a DNA-binding protein of interest is expressed, purified and then bound directly to a dsDNA microarray that contains a large number of different potential DNA-binding sites.

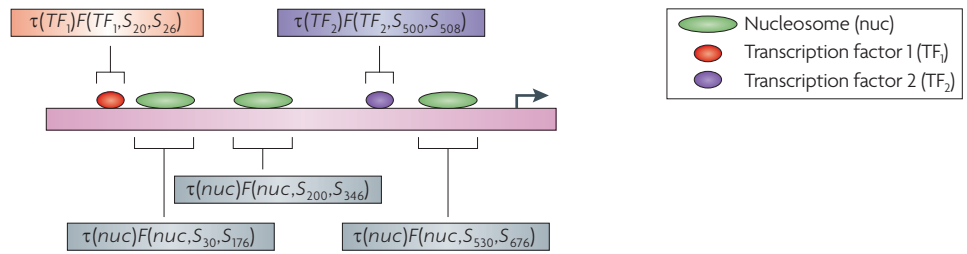
Microfluidic platform

A high-throughput platform for measuring protein–DNA affinities on the basis of mechanically induced trapping of molecular interactions.

DNA looping

A conformation of a dsDNA sequence in which two regions of the DNA that are separated along the DNA in one dimension are brought close together in three-dimensional space.

Box 1 | Computing gene expression from DNA sequence



$$W(c) = F(0)\tau(TF_1)F(TF_1, S_{20}, S_{26})\tau(nuc)F(nuc, S_{30}, S_{176})\tau(nuc)F(nuc, S_{200}, S_{346})\tau(TF_2)F(TF_2, S_{500}, S_{508})\tau(nuc)F(nuc, S_{530}, S_{676})$$

Several models have been proposed^{10,11,16} for translating DNA sequences into transcriptional behaviour. These models are all based on an assumption of thermodynamic equilibrium (BOX 2). They use the intrinsic equilibrium affinities and concentrations of the various DNA-binding molecules (for example, activators and repressors) to compute the probability of RNA polymerase occupancy and then assume that the gene expression level is proportional to polymerase occupancy. The computation is divided into two steps: one that computes the occupancy distribution of the molecules on the target DNA sequence and another that translates this occupancy distribution into a level of gene expression. Denoting each possible configuration of molecules on the DNA by c and the probability of RNA polymerase binding by $P(E)$, the overall probability of polymerase binding is the sum, over all legal configurations, of the probability of polymerase binding given a particular configuration c , $P(E|c)$ weighted by the probability of the configuration itself, $P(c)$:

$$P(E) = \sum_{c \in C} P(c)P(E|c)$$

Step 1: occupancy distribution of molecules on target DNA

Under the equilibrium assumption, the probability of a configuration is:

$$P(c) = \frac{W(c)}{\sum_{c' \in C} W(c')}$$

in which $W(c)$ represents the statistical weight of c . The simplest model assumes that molecules bind independently, and so the statistical weight of a configuration is the product of the contribution of each molecule bound in the configuration. In turn, the contribution of a molecule is determined by its concentration and by its affinity to the sequence at the bound position. Thus, for a configuration with k molecules m_1, \dots, m_k bound at positions p_1, \dots, p_k , the statistical weight $W(c)$ of the configuration is:

$$W(c) = F(0) \prod_{i=1}^k \tau(m_i)F(m_i, p_i, p_{i+L_i})$$

in which $\tau(m_i)$ is the concentration of the molecule bound at position p_i , F_0 is the statistical weight of the empty configuration, and the energetic contribution from the binding of molecule m_i from position p_i to position p_{i+L_i} with L_i being the binding site length of molecule m_i , is given by:

$$F(m_i, p_i, p_{i+L_i}) = e^{-\frac{E_i}{K_B T}}$$

The normalizing term, also known as the partition function, sums over all possible legal configurations of molecules and is given by:

$$\sum_{c' \in C} W(c')$$

The figure illustrates the computation of $W(c)$ for one example configuration. S_i indicates sequence position.

Step 2: from occupancy distribution to expression level

The second model component, $P(E|c)$, which translates configurations into expression levels, is less well understood. One simple model¹⁰ assumes that each factor bound in the configuration contributes independently to the expression outcome, with activators contributing positively and repressors contributing negatively. This model uses the logistic function to translate these contributions into expression. From a mechanistic perspective, this expression component represents the total attractive force that recruits the RNA polymerase to the sequence. Thus, for a configuration with k molecules m_1, \dots, m_k bound at positions p_1, \dots, p_k , the probability of expression is:

$$P(E|c) = \text{logit} \left(w_0 + \sum_{i=1}^k w_{m_i} \right) = 1 / \left(1 + \exp \left(- \left(w_0 + \sum_{i=1}^k w_{m_i} \right) \right) \right)$$

in which w_0 represents the basal expression level and w_i represents the expression contribution of transcription factor i .

The models above compute the expression of one sequence at the concentration of a particular binding molecule. To compute the expression pattern of a sequence across a spatial or temporal axis along which molecule concentrations change, these models are applied separately to every point along the axis and the expression at each point is then combined to produce the entire expression pattern along the axis.

Legal configuration

An arrangement of molecules along a DNA sequence in which there is no steric overlap between any two molecules on the DNA.

Box 2 | The equilibrium assumption and the role of nucleosome remodellers

The models we review assume that molecules bind at thermodynamic equilibrium. Although models based on this assumption predict gene regulation in both prokaryotes^{13,14} and eukaryotes^{10,16}, the validity of this assumption has not been shown. This assumption is particularly challenging to justify (or imagine) in eukaryotes, owing to the presence of nucleosomes and many ATP-dependent nucleosome-remodelling factors. *In vitro*, these nucleosome-remodelling factors can control the spacing between nucleosomes on long stretches of DNA⁸⁸ and drive nucleosomes to unfavourable locations on shorter DNA fragments; for example, from favoured positioning sequences to DNA ends⁸⁹. *In vivo*, inactivation of the ATP-dependent chromatin remodelling complex Isw2 leads to an average shift of ~15 bp in the location of nucleosomes relative to wild-type cells at some loci, suggesting that Isw2 might determine the positions of nucleosomes at these loci⁹⁰. In principle, such remodelling activities could invalidate the equilibrium assumption on which the models discussed here are based, and could instead require detailed and unique kinetic models for every sequence.

However, other evidence suggests that the equilibrium hypothesis might be a good approximation *in vivo*. The high similarity of *in vivo* nucleosome organizations to those obtained in a purified *in vitro* reconstitution system¹⁵, which is believed to achieve and then freeze in a true thermodynamic equilibrium²², directly shows that many aspects of the *in vivo* nucleosome organization are similar to an equilibrium distribution. Moreover, even when chromatin remodellers drive nucleosomes to different locations on short DNA fragments *in vitro*, the positions adopted by the nucleosomes remain the same as those that are favoured intrinsically by the nucleosomes, and only the degree to which the nucleosomes occupy the different favoured positions changes⁹¹.

One way of understanding these facts collectively is if the remodelling factors themselves do not determine the destinations of the nucleosomes that they mobilize but instead catalyse nucleosome mobility, allowing nucleosomes to rapidly sample alternative positions. In this view, ATP hydrolysis does not force nucleosomes to unfavourable locations; rather, ATP hydrolysis is required to provide sufficient energy for a nucleosome to cross the transition state free energy barriers that separate occupancy at thermodynamically favoured locations. The same logic is presently used to explain the requirement of ATP for kinesin movement along a microtubule and of helicases along DNA⁹²; indeed, the ATP-dependent motor domains of all of the known nucleosome-remodelling factors are members of helicase protein superfamilies.

Thus, in this view, the result of remodeller action is a thermodynamic equilibrium between the nucleosomes and the transcription factors that compete with nucleosomes for occupancy along the genome. When the combination of transcription factors changes, for example, during development or following an environmental fluctuation, the action of remodellers allows the system to rapidly re-equilibrate to a new distribution of bound molecule configurations. In this view, the changed nucleosome positions that result from Isw2 inactivation⁹⁰ would be interpreted as a failure to establish an equilibrium distribution in the absence of remodeller activity recruited specifically to the affected regions.

therefore, the simplicity with which existing models translate binding configurations to transcriptional output mainly reflects gaps in our knowledge of this process.

To summarize, the intrinsic binding specificities of each molecule determine the particular binding affinity landscape that the molecule experiences on an input DNA sequence. At a given concentration of binding molecules, these landscapes and the competitive and cooperative interactions between the molecules dictate the probabilities of all possible configurations of bound molecules. Finally, the transcriptional output of a regulatory sequence is simply the sum of the transcriptional output of all binding configurations, and each binding configuration is weighted by its probability. Having presented the general modelling framework, we now review the experimental observations that it explains, starting with observations regarding nucleosome organization.

Determinants of nucleosome organization *in vivo*

As mentioned in an earlier section, it is difficult to estimate the relative contributions of multiple factors to nucleosome organization *in vivo* using *in vivo* measurements. Advances in this direction were made possible by comparing the organization of nucleosomes *in vivo* with the genome-wide organization of nucleosomes assembled *in vitro* on purified yeast genomic DNA¹⁵.

Direct consequences of the nucleosome landscape: distinct nucleosome positioning. Comparison of *in vitro* and *in vivo* nucleosome organization in yeast showed that the large regions of nucleosome depletion around the ends of genes^{26,52–54} and around transcription factor-binding sites^{24,28,55} that are observed *in vivo* are largely encoded by nucleosome sequence preferences. The nucleosome affinity landscape might therefore assist in directing transcription factors to their appropriate sites in the genome^{8,56,57} (FIG. 2a). For *Abf1* and *Reb1*, which are two abundant transcription factors known to influence chromatin structure, the nucleosome affinity landscape encodes only a small part of the nucleosome depletion; the large depletion observed around the binding sites of these factors *in vivo* is therefore probably due to the ability of these factors to outcompete nucleosomes¹⁵. The nucleosome depletion around the starts of genes was also found to be encoded by the intrinsic nucleosome affinity landscape but, in this case, the action of chromatin remodellers and the binding of transcription factors and the transcription initiation machinery also make considerable contributions to nucleosome depletion¹⁵ (FIG. 2a).

Another notable feature of the nucleosome organization *in vivo* is that some regions of the genome have a small number of well-positioned nucleosomes, whereas others have ‘fuzzy’ nucleosomes, in which many nucleosome positions are observed^{55,58}. The existence of many regions with fuzzy nucleosomes in the worm could indicate that much of the nucleosome organization is not dictated by DNA sequence⁵⁸. However, in principle, regions with well-positioned nucleosomes and regions with fuzzy nucleosomes can both be encoded by the genomic sequence, if we assume that well-positioned regions have a peaked nucleosome affinity landscape and that fuzzy regions have a relatively flat landscape. Indeed, both types of regions exist in the map of nucleosomes that was assembled on purified yeast DNA, and a model of nucleosome sequence preferences constructed from these yeast data is significantly correlated with the *in vivo* nucleosome organization in the worm¹⁵ (FIG. 2b,c).

Indirect consequences of the nucleosome landscape: long-range ordering of nucleosomes. The examples above are cases in which the nucleosome landscape directly accounts for the experimental observations. Other observations might be explained by using the part of the framework that converts binding landscapes to binding configurations. For example, several studies observed a long-range ordering of nucleosomes downstream of the starts of genes, which decays with the distance from the start of the gene. There are strong

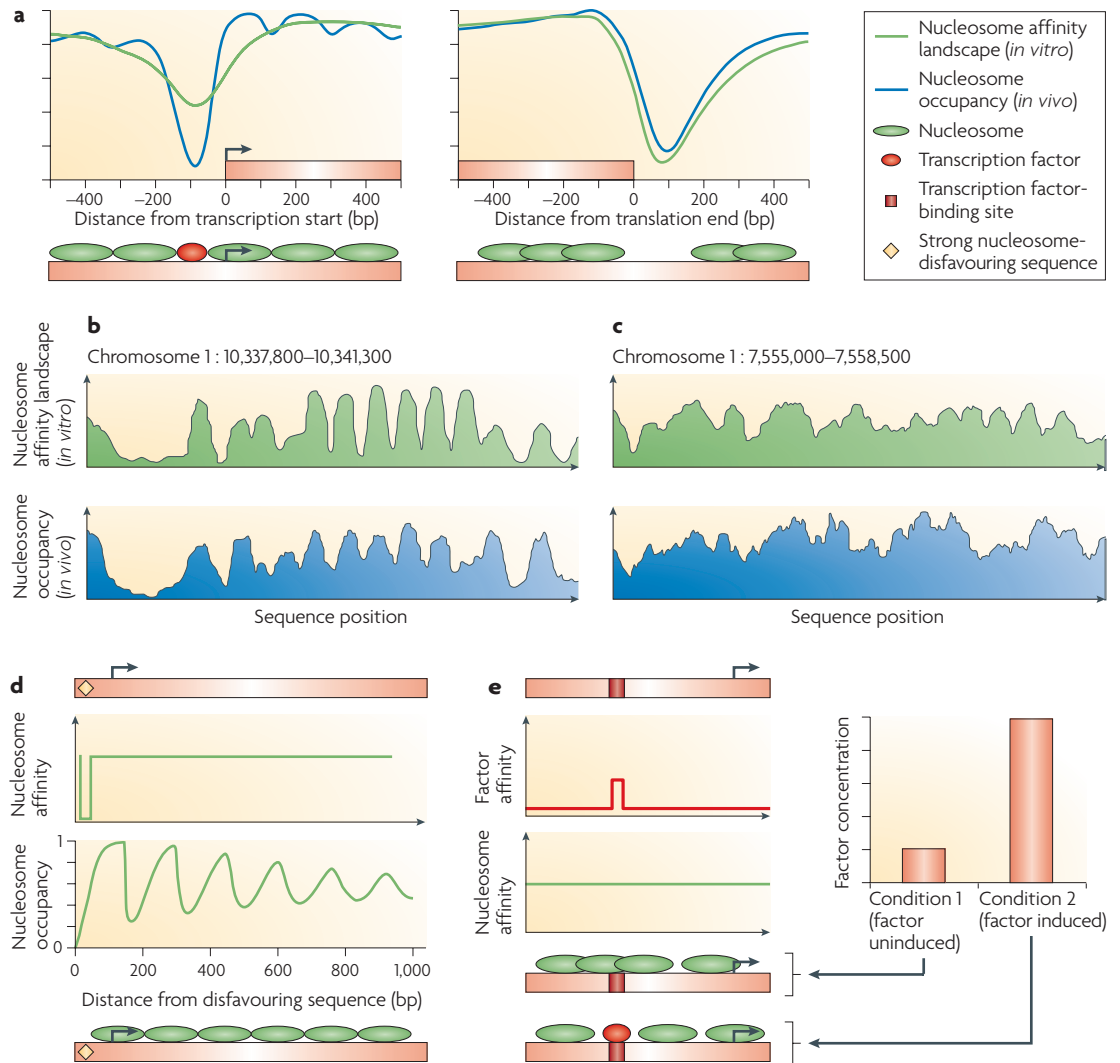


Figure 2 | Main determinants of *in vivo* nucleosome organization. **a** | The nucleosome occupancy *in vivo* in yeast (blue) and the nucleosome affinity landscape measured *in vitro* by assembling purified histones on purified yeast genomic DNA¹⁵ (green), averaged across all genes. The occupancy around gene transcription start sites is shown on the left and around gene translation end sites on the right. A schematic illustration of the key components that contribute to the *in vivo* nucleosome occupancy is also shown below each graph. Nucleosome depletion around the ends of genes is largely encoded by the nucleosome affinity landscape, and nucleosome depletion around the starts of genes results both from the encoded nucleosome affinity landscape and from the binding action of transcription factors. **b** | The average nucleosome occupancy for a genomic region from the worm with well-positioned nucleosomes *in vivo*⁵⁸ (blue) and the average nucleosome affinity landscape for that region as predicted by a model constructed from *in vitro* data in yeast¹⁵ (green) are shown. **c** | As in part **b**, the average nucleosome occupancy *in vivo* and the average nucleosome affinity landscape across a genomic region from the worm with less well-defined nucleosome locations ('fuzzy nucleosomes') are shown. The agreement between the predictions of a model based on nucleosome sequence preferences and the experimental measurements in parts **b** and **c** suggests that both types of regions might be encoded by the genomic sequence, through peaked nucleosome affinity landscapes (in the case of well-positioned nucleosomes, **b**) or relatively flat landscapes (in the case of fuzzy nucleosomes, **c**). **d** | Nucleosome-disfavouring sequences can have a long-range effect on the nucleosome organization. This example sequence contains a strong nucleosome-disfavouring sequence (yellow diamond) — these sequences are abundant in eukaryotic genomes⁹³. When such a nucleosome-affinity landscape is combined with a high nucleosome concentration, as *in vivo*, the bound nucleosomes automatically organize into ordered arrays, the order of which decays with the distance from the original disfavouring sequence (bottom graph and schematic bottom sequence). This phenomenon is termed 'statistical positioning'⁵⁹. **e** | A single sequence might potentially encode different nucleosome organizations in different cell types or biological conditions by encoding different outcomes for nucleosome-factor competition at different factor concentrations. This sequence has a uniform landscape for nucleosomes and a landscape for one factor that includes a single strong binding site. In condition 1, in which the hypothetical factor is expressed at low levels, the most likely configurations have nucleosomes covering the factor-binding site, whereas in condition 2, in which the factor is expressed at high levels, the most likely configurations have the factor binding to its site, causing a displacement of nucleosomes from their cognate sites.

nucleosome-disfavouring sequences upstream of gene start sites and nucleosome-positioning sequences over the starts of genes²⁶. These sequences lead to a high probability that a nucleosome-depleted region occurs upstream of the start of the gene together with a well-positioned nucleosome over the gene start. Introducing boundary constraints such as nucleosome-disfavouring and nucleosome-positioning sequences into the framework presented here automatically results in a long-range periodic ordering of nearby nucleosomes. This ordering is simply a consequence of the high concentrations of nucleosomes along the DNA and the steric hindrance between them⁵⁹ (FIG. 2d). This ordering or 'statistical positioning' is greatest immediately adjacent to the boundary constraint and decays with the distance away from it. Thus, the intrinsic nucleosome landscape is likely to contribute to the long-range ordering of nucleosomes near the starts of genes through the indirect long-range effects that the sequences surrounding the starts of genes exert on nucleosome configurations at the high nucleosome concentrations that exist *in vivo*. An additional substantial contribution to the long-range ordering *in vivo* is likely to come from transcription factors and the transcription initiation machinery, the binding of which further increases the boundary constraint⁶⁰.

Nucleosomes might also have different organizations in different conditions^{54,60} or cell types. The framework presented here can in principle explain such observations because the concentrations of either the nucleosomes or the transcription factors (or both) change in different conditions or cell types, resulting in a different distribution of binding configurations (FIG. 2e). Thus, a single binding affinity landscape can encode many different distributions of binding configurations, depending on the different concentrations of molecules.

Nucleosome landscapes and transcription factors Explaining the repressive function of nucleosomes.

Aside from explaining the DNA-binding patterns of molecules such as nucleosomes, we need to understand the dynamic transcriptional behaviour that genes show in response to changes in the concentration of the regulating factors. The quantitative framework presented here can be used to directly read DNA sequences and predict these responses; this is achieved by computing the probability of a factor binding at increasing factor concentrations from the encoded affinity landscape. For example, consider a hypothetical DNA sequence that has a landscape for only one transcription factor, which in turn recognizes a single binding site. The activation dynamics of such a target gene are determined only by the affinity of the single site. When nucleosomes are added to the equation with a uniform energy landscape such that there are no intrinsic favoured locations, they compete with the factor for binding. This results in a lower probability of factor binding, and a given level of gene activation then requires a higher factor concentration⁶¹ (FIG. 3a,b). This model thus provides a simple explanation for why nucleosomes are considered to be general repressors^{62–64}.

Competition between nucleosome and transcription factor binding. In the more realistic setting in which a non-uniform nucleosome landscape dictates a non-uniform nucleosome occupancy distribution, the activation dynamics depend on the nucleosome occupancy around the factor site, and activation occurs at a lower factor concentration for sequences in which the binding site for the factor is located in a region of low intrinsic nucleosome occupancy⁶¹. Indeed, a recent study showed that lower nucleosome occupancy at sites for the yeast transcription factor *Pho4* in conditions in which *Pho4* concentration is low is predictive of an earlier onset (in terms of both the *Pho4* concentration and the period of time) of activation^{65,66}. The importance of nucleosome occupancy for activation⁶⁵ combined with the reliability with which nucleosome occupancy can be predicted from sequence¹⁵ means that the dependence of activation on factor concentrations can be predicted directly from sequence, although the accuracy of such predictions remains to be shown. For example, nucleosome-disfavouring sequences generate low nucleosome occupancy in their immediate vicinity and, for nearly all yeast transcription factors, the subset of their sites that is near nucleosome-disfavouring sequences has lower nucleosome occupancy²⁶. This context-dependent accessibility of factor sites thus provides a mechanism predicted directly from the DNA sequence by which the same factor might regulate its different targets with different activation dynamics by positioning some of its sites near nucleosome-disfavouring sequences^{26,61} (FIG. 3c–e).

Many regulatory sequences contain multiple sites for multiple transcription factors. In sequences in which multiple sites are close to each other, each of the corresponding binding factors separately competes with nucleosomes, resulting in indirect binding cooperativity between factors⁶⁷ and, again, gene activation at a lower factor concentration. Indeed, such cooperativity was shown in yeast between an endogenous yeast transcription factor and two foreign transcription factors from *Escherichia coli*⁴⁵. Note that this obligate cooperativity is predicted by the framework directly from the affinity landscape of the sequence and without invoking specific cooperative mechanisms such as protein–protein interactions⁶¹ (FIG. 3f–h). This competition between transcription factors and nucleosomes might be part of the mechanism of cooperative binding that was shown in regulatory sequences from yeast^{16,68,69} and flies^{10,70} that contain clusters of factor-binding sites. Intriguingly, as this obligate cooperativity occurs between any two factors, it might also explain why some transcription factors have both activatory and inhibitory roles^{71,72}; for example, a transcriptional repressor can apparently act as an activator if its competition with nucleosomes promotes the binding of a nearby activator (FIG. 3f–h).

Encoding distinct modes of regulation

Chromatin remodellers are important in transcriptional regulation⁷³ and target specific sets of genes⁷⁴. Although the mechanism by which chromatin remodellers are recruited to specific loci is not well understood, a recent study²⁶ suggested that the differential requirements for remodellers at different loci might be partly explained

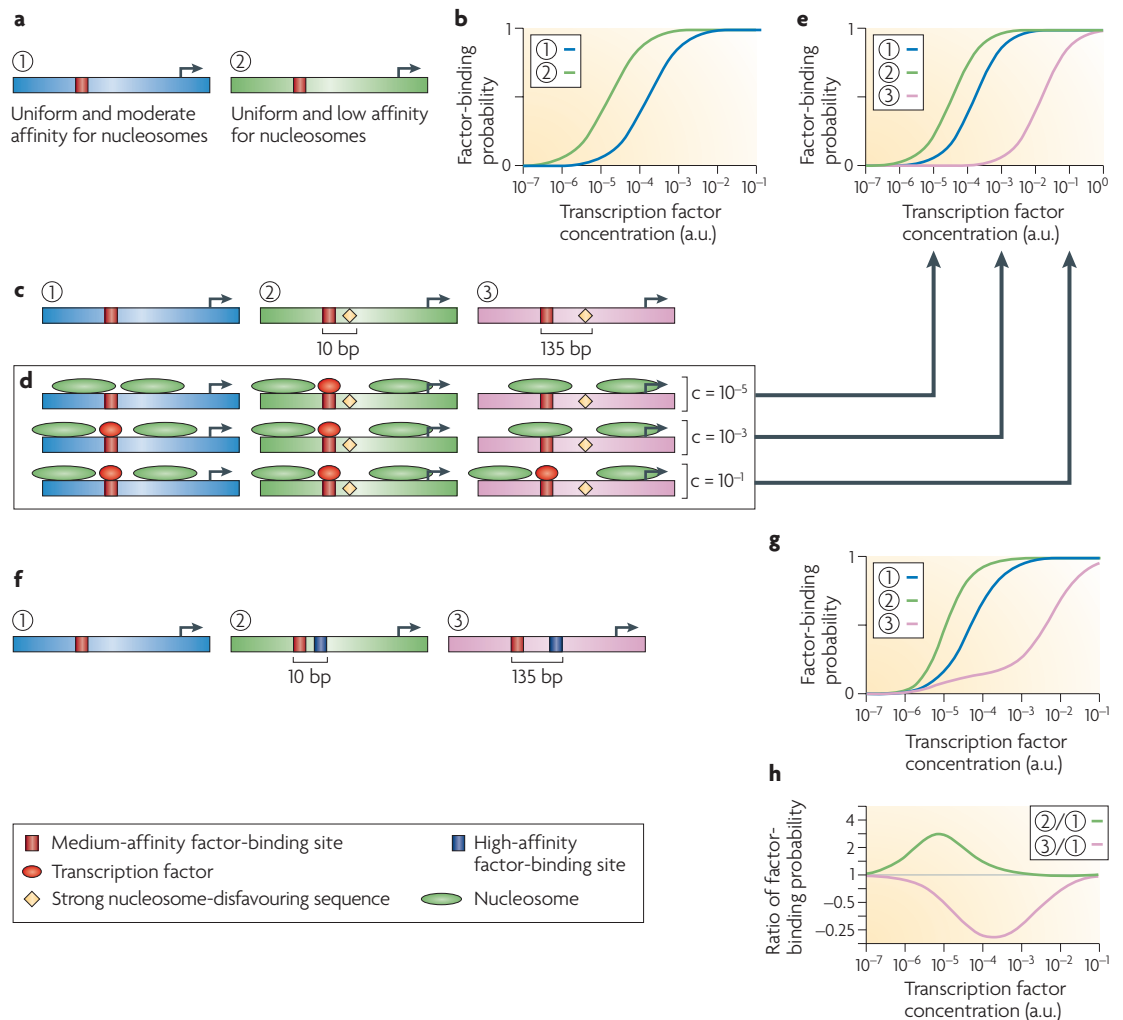


Figure 3 | Reading gene expression dynamics from DNA sequence. **a** | Nucleosomes act as general repressors. The two example sequences have a transcription factor landscape containing a single binding site and have either a uniform and moderate-affinity landscape for nucleosomes (sequence 1) or a uniform and low-affinity landscape for nucleosomes (sequence 2). **b** | For the two sequences from part **a**, the probability of transcription factor binding at different factor concentrations is computed by applying the framework presented here to the binding landscapes of those two sequences. **c** | Nucleosome-disfavouring sequences determine the threshold of activation. The three example sequences have differing nucleosome and factor landscapes: sequence 1 has a uniform nucleosome landscape; sequence 2 has a landscape with a sequence that strongly disfavours nucleosome formation, which is located 10 bp from the single transcription factor site; and sequence 3 is the same as sequence 2, but the disfavoured sequence is located 135 bp from the transcription factor site. **d** | For each of the three sequences from part **c**, the most likely molecule-binding configurations at three different factor concentrations (abbreviated as ‘*c*’) are shown. **e** | The probability of transcription factor binding at each of the three sequences from part **c**. **f** | Proximal factor sites show cooperative or destructive binding. The three example sequences have a uniform nucleosome affinity landscape and differing factor landscapes: sequence 1 has a single factor site; sequence 2 has two factor sites separated by 10 bp; and sequence 3 has two factor sites separated by 135 bp. **g** | The probability of transcription factor binding to the left (red) site at each of the three sequences from part **f**. **h** | The cooperative and destructive binding effects in sequences 2 and 3, respectively, displayed as the ratio between the factor-binding probability at sequence 2 or 3 compared with sequence 1. a.u., arbitrary units.

by DNA sequence. This study defined two categories of yeast genes based only on sequence information. The differences in the affinity landscapes of these two categories of genes suggest that they undergo different modes of regulation according to whether transcription factors compete with nucleosomes for access to the DNA. Indeed, genes in the category in which competition is predicted to occur have higher rates of histone turnover⁷⁵

and transcriptional noise⁷⁶, consistent with an ongoing dynamic competition between nucleosome assembly and factor binding. Correspondingly, these genes contain more targets of chromatin-remodelling complexes⁷⁴ (FIG. 4).

Thus, by partitioning genes on the basis of affinity landscapes, two modes of transcriptional regulation can be identified that provide a partial explanation for the differential requirement for chromatin remodellers

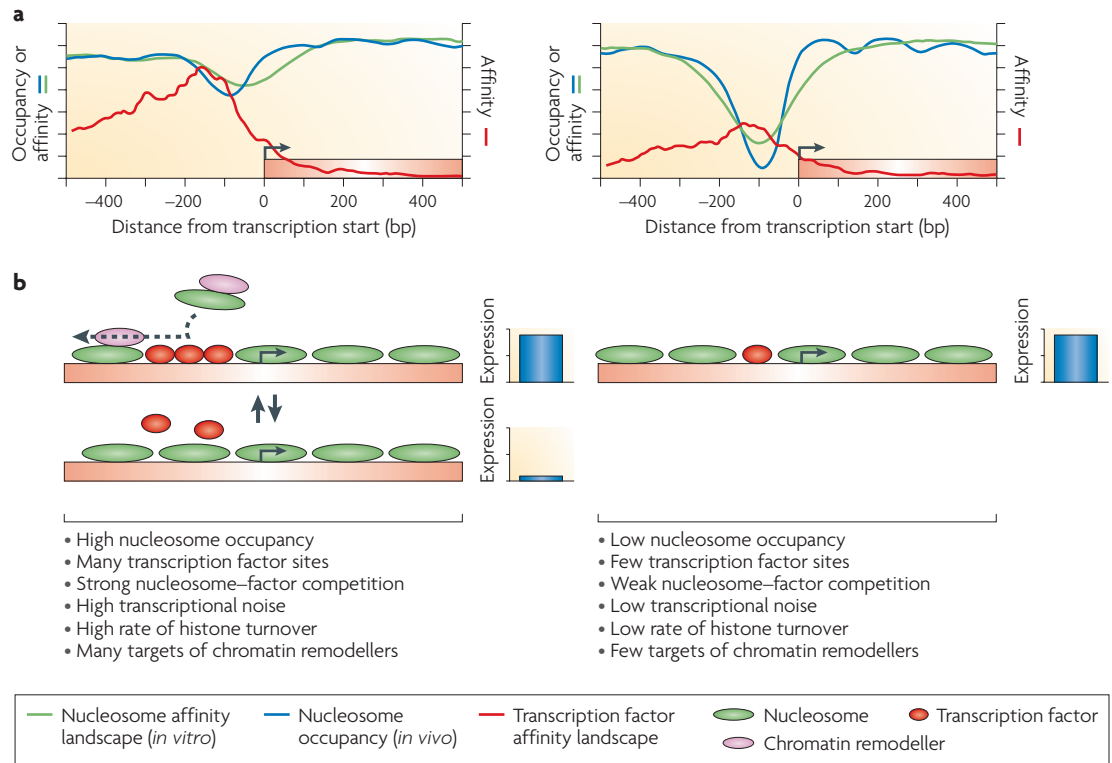


Figure 4 | Distinct modes of transcriptional regulation encoded by DNA sequence. **a** | Two sets of yeast genes were defined based on their DNA sequence²⁶: one set was defined by the absence of strong nucleosome-disfavouring sequences and the presence of TATA sequences (left); and one set was defined by the presence of strong nucleosome-disfavouring sequences and the absence of TATA sequences (right). The nucleosome occupancy *in vivo* (blue) and the nucleosome affinity landscape measured *in vitro* by assembling purified histones on purified yeast genomic DNA¹⁵ (green), averaged across all genes of each gene set, are shown. Also shown is the approximate affinity landscape for all transcription factors across all genes of each of the two gene sets, using the spatial distribution of factor-binding site occurrences as a proxy for the spatial distribution of affinity (red). **b** | The most likely configurations of each gene set. In the gene set on the left, the nucleosome landscape shows high nucleosome occupancy and the transcription factor landscape has a large number of binding sites spread across the regulatory region, suggesting that nucleosomes and factors compete for access to the DNA. In support of this suggestion are the high transcriptional noise, high rate of histone turnover and enrichment for chromatin remodeller activity that were found for this gene set²⁶. By contrast, in the gene set on the right, the nucleosome landscape shows strong nucleosome depletion around the transcription start site and the factor landscape has fewer binding sites, but with a preference for these sites to be located at the nucleosome-depleted region. These landscapes suggest that there is little competition between factors and nucleosomes, and supporting this are the low noise, low histone turnover and absence of enrichment for chromatin remodeller targets that were found for this gene set²⁶.

at different genomic loci. A corollary of this model is that genomes can encode different dynamic responses — even to the same transcription factor — at two different regulatory sequences by embedding the cognate binding sites of the transcription factor in two different nucleosome affinity landscapes.

Evolution of binding affinity landscapes

A genetic mechanism for achieving phenotypic diversity. The examples above show that many diverse aspects of transcriptional regulation can be understood directly from the affinity landscapes encoded in DNA sequences. As changes in transcriptional regulation are important for generating phenotypic diversity among species, an intriguing possibility is that the genetic mechanisms that underlie these regulatory changes involve changes in affinity landscapes.

Overall, changes in factor-binding site content alone account for only a small fraction of the observed expression divergence in both yeast^{77,78} and mammals⁷⁸. However, in the case of the yeast mating system, which is regulated by a single transcription factor, variations in the predicted binding sites explain much of the expression divergence across yeast species⁷⁸.

A recent study found that a major change in the transcriptional programme of yeast species, which is connected with the capacity for rapid anaerobic growth, is accompanied by corresponding changes in the nucleosome affinity landscape encoded in the DNA of the orthologous regulatory sequences⁷⁹. In aerobic yeast species, in which cellular respiration genes are active under typical growth conditions, the regulatory sequences encode a landscape that contains a nucleosome-depleted region, whereas in anaerobic yeast species, in which

cellular respiration genes are inactive under typical growth conditions, the orthologous regulatory sequences encode a landscape with relatively high nucleosome occupancy. This suggests that DNA sequence changes that directly alter the encoded nucleosome affinity landscape of regulatory sequences might be a general genetic mechanism for achieving phenotypic diversity across evolution.

Explaining transcriptional noise

Noise in gene expression levels. Levels of cell-to-cell expression variability vary across genes^{76,80,81}, and it is therefore interesting to address whether these noise levels can be predicted directly from the regulatory sequences of a gene. TATA sequences predict high levels of noise, presumably by amplifying fluctuations in gene activation through reinitiation of transcription^{81–84}, whereas nucleosome-disfavouring sequences predict low levels of noise²⁶. This latter observation can be understood using the framework presented here. As discussed above, regulatory target sites that are located close to nucleosome-disfavouring elements will be occupied at lower factor concentrations than target sites that are far from such elements. Thus, a noisy regime in which the probability of a factor binding is ~ 0.5 is reached at lower factor concentrations in target sequences in which the factor-binding site is near nucleosome-disfavouring elements⁶¹ (FIG. 5a–c). If we assume that the physiological concentration of the regulating factor is such that it is likely to bind both types of target sequence then, at this concentration, targets with nucleosome-disfavouring sequences have already ‘escaped’ the noisy regime, providing a plausible explanation for the low noise that has been observed for these targets²⁶. In addition, kinetic models show that regulatory sequences with a higher frequency of transition from the transcriptionally inactive state to the transcriptionally active state are less noisy⁸¹. As nucleosome-disfavouring elements create nucleosome-depleted regions, sequences that contain such elements have a lower requirement for and are less targeted by chromatin remodellers²⁶, which might result in more rapid transitions between the active and inactive states, providing another plausible explanation for the lower noise levels of sequences with nucleosome-disfavouring elements.

Noise in replication initiation. The ideas above are based on modelling the binding of molecules to DNA and, as such, they might have applications beyond the context of transcriptional regulation. For example, DNA replication origins also exhibit cell-to-cell variability: some origins initiate replication in most cell divisions and others initiate only occasionally. Thus, analogous to transcriptional noise, the framework predicts that origins that are close to nucleosome-disfavouring elements have lower nucleosome occupancy, and would therefore be more accessible to the replication initiation complex and initiate replication with higher efficiency. Indeed, when replication efficiency was measured in fission yeast⁸⁵, the origins that were close to nucleosome-disfavouring sequences initiated with higher efficiency²⁶

(and so higher probability), and a systematic sequence deletion study around one replication origin found that the deletion of a strong nucleosome-disfavouring element resulted in the largest reduction in replication efficiency⁸⁶.

In summary, by examining the activation dynamics that the framework predicts directly from DNA sequence, measurements of cell-to-cell expression and replication variability can be partly explained from sequence alone. More generally, the same approach can be used to predict the noise of sequences with more complex affinity landscapes, thereby generating testable hypotheses. For example, the framework predicts lower noise in sequences in which multiple sites are clustered in close proximity because, as discussed above, site clustering leads to cooperative factor binding and a sharper activation curve, and a noisy regime therefore spans a smaller range of factor concentrations⁶¹ (FIG. 5d,e).

Summary and future directions

This Review presents a unifying quantitative and conceptual framework for translating DNA sequences into transcriptional behaviours. The key idea behind this translation process is that DNA-binding molecules have specific intrinsic affinities for DNA sequences, and thus every sequence defines a unique affinity landscape with respect to each molecule. The molecules that interact with and bind to DNA result in a unique distribution of molecule-binding configurations at each sequence and lead to a transcriptional output.

Recent studies have determined the intrinsic affinities of nucleosomes and of many transcription factors for DNA, allowing us to accurately translate DNA sequences into affinity landscapes. These landscapes, either directly or through the application of the framework, partly explain many experimental observations regarding the binding patterns of nucleosomes, the dynamics of transcriptional activation and transcriptional noise, indicating that diverse sets of transcriptional behaviours can be read directly from the DNA sequence.

Many challenges remain. A large number of predictions generated by current models still need to be validated experimentally. Several aspects of the framework regarding the translation of binding landscapes into binding configurations, and especially the translation of binding configurations into transcriptional output, are still poorly understood and require targeted experiments for determining them at a quantitative level. In particular, the treatment of DNA sequences as being one-dimensional should ultimately be replaced by modelling binding in three dimensions, taking into account both the long-range DNA looping that allows distant enhancers to interact with promoters and the short-range looping that allows factors in the same regulatory module to interact with each other⁸⁷. Regulatory modules might have as yet unrecognized highly specific three-dimensional architectures, which could depend on the detailed locations of nucleosomes and other factors that bend or twist DNA. Any such specific architectures will influence the interactions that are mediated by short- and long-range DNA looping. The measurement

TATA sequence

A DNA sequence with a core sequence of 5′-TATA-3′ found in the promoter region of many genes. It is typically bound by a corresponding TATA-binding protein during the process of recruiting RNA polymerase to a promoter.

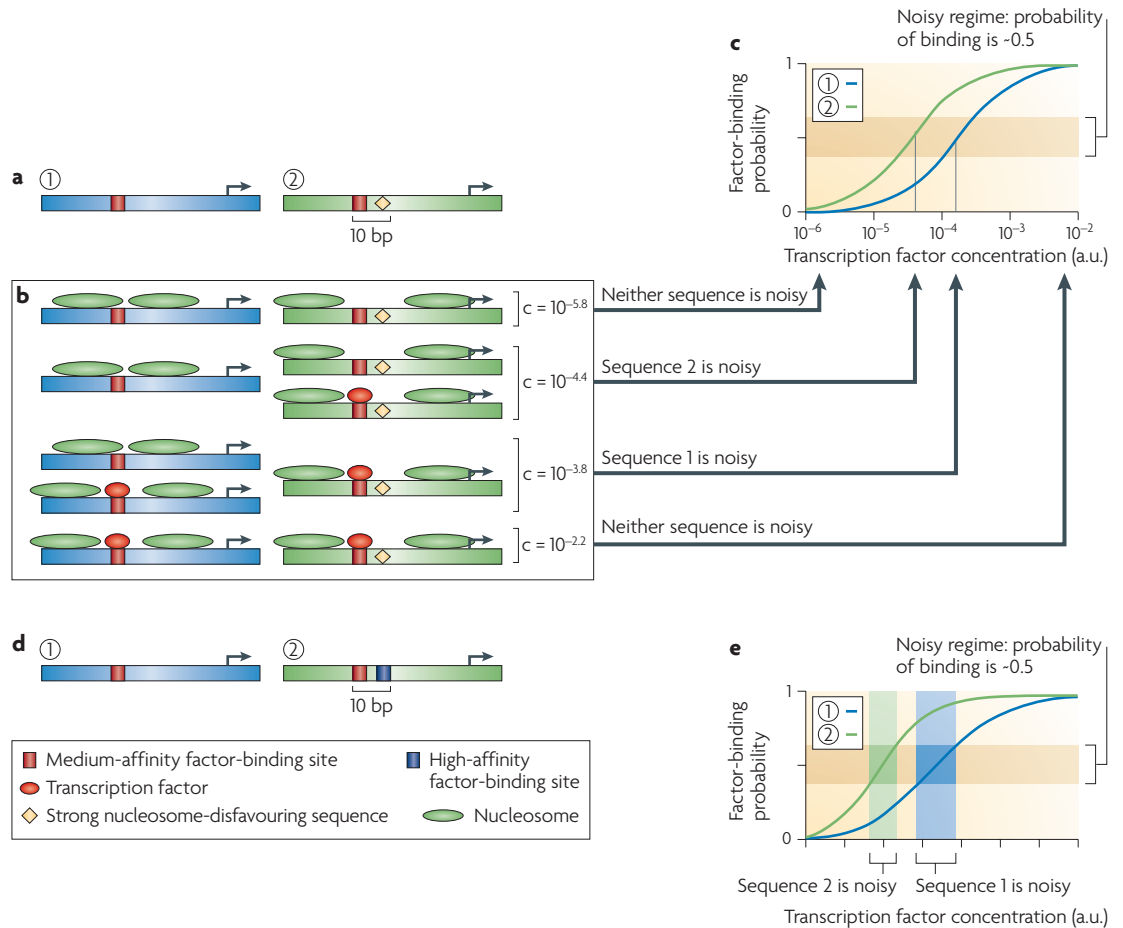


Figure 5 | Explaining transcriptional noise from DNA sequence. a | Nucleosome-disfavouring sequences determine the range of factor concentrations at which high transcriptional noise occurs. Two example sequences are shown, one with a uniform nucleosome landscape (sequence 1) and one with a nucleosome landscape containing a sequence that strongly disfavours nucleosome formation, which is located 10 bp from the single transcription factor site (sequence 2). **b** | For the two sequences from part **a**, the probability of transcription factor binding at different factor concentrations is computed by applying the framework presented here to the binding landscapes of those two sequences⁶¹. Under this equilibrium framework, the regime of high transcriptional noise is where the probability of transcription factor binding is ~0.5 as, at this regime, the variance of factor binding is maximal. Note, however, that although the variance of factor binding is one of the determinants of noise levels, other determinants also exist. **c** | The most likely molecule-binding configurations at each of the two sequences from part **a** for four different factor concentrations (c) are shown. Note that at each of the two intermediate concentrations, one of the two sequences is noisy; that is, the configurations in which the factor is bound and the configurations in which the factor is not bound have almost equal probabilities. **d** | Cooperative binding reduces the range of factor concentrations at which there is high transcriptional noise. Two example sequences with a uniform nucleosome landscape are shown, in which one sequence has a single factor site (sequence 1) and the other sequence has two factor sites separated by 10 bp (sequence 2). **e** | The probability of transcription factor binding to the left (red) site at each of the two sequences from part **d**. The regime of high levels of noise is highlighted. The range of factor concentrations at which each sequence shows high levels of noise is depicted. The range of factor concentrations in which sequence 2 (the sequence with cooperative binding) is noisy is smaller than the corresponding range for sequence 1. a.u., arbitrary units.

of the sequence-dependent energetic costs of DNA looping can be used to assign statistical weights to the expanded set of configurations that include short DNA loops. Molecular mechanics studies of longer chromatin regions *in vitro* or *in vivo* can supply the data that are needed to model the free energy costs of longer loops. Detailed studies of the high-resolution architectures of regulatory sequences will also be required.

Current models assume that the system is in thermodynamic equilibrium and make other simplifying assumptions regarding steric hindrance and the integration of effects of multiple factors, all of which need to be validated experimentally or modified appropriately. Current models also assume that different histone variants, different histone post-translational modifications and the absence or presence of histone H1 do

not substantially influence the nucleosome sequence preference, and these assumptions also need to be validated or modified. The activity of key components such as chromatin remodellers has not been incorporated into current models. Experiments in purified *in vitro* systems should be useful for directly measuring the quantitative effect of these components at a genome-wide scale, thereby allowing their integration into future models. As most of the experimentation and modelling has been carried out only in bacteria and yeast, it will be important to test and develop the models in higher eukaryotes. Although the basic rules of the molecular interactions modelled by current approaches should be universal, the genomes of multicellular organisms might

encode landscapes that combine these building blocks in different ways from unicellular organisms, leading to new design principles in transcriptional regulation. Finally, although we have focused here on approaches that quantitatively model the translation of DNA sequences into transcriptional behaviours, a major remaining challenge is to understand the functional consequences that these expression patterns have and, ultimately, to assess which deviations from the wild-type patterns might be deleterious and lead to disease. Future experiments, combined with the development of improved quantitative models, should allow us to address these and other challenges, and bring us closer to a mechanistic and predictive understanding of transcriptional regulation in all organisms.

- Casadaban, M. J. Transposition and fusion of the *lac* genes to selected promoters in *Escherichia coli* using bacteriophage lambda and Mu. *J. Mol. Biol.* **104**, 541–555 (1976).
- Guarente, L. & Ptashne, M. Fusion of *Escherichia coli lacZ* to the cytochrome *c* gene of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **78**, 2199–2203 (1981).
- Bellen, H. J. *et al.* P-element-mediated enhancer detection: a versatile method to study development in *Drosophila*. *Genes Dev.* **3**, 1288–1300 (1989).
- Wilson, M. D. *et al.* Species-specific transcription in mice carrying human chromosome 21. *Science* **322**, 434–438 (2008).
- Rajewsky, N., Vergassola, M., Gaul, U. & Siggia, E. D. Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics* **3**, 30 (2002).
- Sinha, S., van Nimwegen, E. & Siggia, E. D. A probabilistic method to detect regulatory modules. *Bioinformatics* **19** (Suppl. 1), i292–i301 (2003).
- Granek, J. A. & Clarke, N. D. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol.* **6**, R87 (2005).
- Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006). **This paper shows that genomes encode the location of many of their nucleosomes and that this intrinsically encoded nucleosome organization might facilitate specific chromosome functions.**
- Ioshikhes, I. P., Albert, I., Zanton, S. J. & Pugh, B. F. Nucleosome positions predicted through comparative genomics. *Nature Genet.* **38**, 1210–1215 (2006). **This paper shows that different nucleosome architectures might be encoded at different classes of regulatory sequences.**
- Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451**, 535–540 (2008). **This study develops an equilibrium-based thermodynamic model that relates factor binding to gene expression in *Drosophila melanogaster* early embryonic patterning.**
- Bintu, L. *et al.* Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.* **15**, 116–124 (2005).
- Buchler, N. E., Gerland, U. & Hwa, T. On schemes of combinatorial transcription logic. *Proc. Natl Acad. Sci. USA* **100**, 5136–5141 (2003).
- Dodd, I. B., Shearwin, K. E. & Sneppen, K. Modelling transcriptional interference and DNA looping in gene regulation. *J. Mol. Biol.* **369**, 1200–1213 (2007).
- Kuhlman, T., Zhang, Z., Saier, M. H. Jr & Hwa, T. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **104**, 6043–6048 (2007).
- Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009). **This study directly measured the intrinsically encoded nucleosome organization of the genome and showed that it is similar to the *in vivo* organization. A model of nucleosome occupancy trained on *in vitro* reconstituted yeast nucleosomes significantly predicts nucleosome organization in *Caenorhabditis elegans*.**
- Gertz, J., Siggia, E. D. & Cohen, B. A. Analysis of combinatorial *cis*-regulation in synthetic and genomic promoters. *Nature* **457**, 215–218 (2009). **This paper develops and experimentally verifies an equilibrium-based thermodynamic analysis that relates transcription factor binding to gene expression, and shows that these contributions explain a significant fraction of the measured variability in gene expression.**
- Foat, B. C., Morozov, A. V. & Bussemaker, H. J. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**, e141–e149 (2006).
- Gupta, S. *et al.* Predicting human nucleosome occupancy from primary sequence. *PLoS Comput. Biol.* **4**, e1000134 (2008).
- Sinha, S., Adler, A. S., Field, Y., Chang, H. Y. & Segal, E. Systematic functional characterization of *cis*-regulatory motifs in human core promoters. *Genome Res.* **18**, 477–488 (2008).
- Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
- Gencheva, M. *et al.* *In vitro* and *in vivo* nucleosome positioning on the ovine β -lactoglobulin gene are related. *J. Mol. Biol.* **361**, 216–230 (2006).
- Thastrom, A., Bingham, L. M. & Widom, J. Nucleosomal locations of dominant DNA sequence motifs for histone–DNA interactions and nucleosome positioning. *J. Mol. Biol.* **338**, 695–709 (2004).
- Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675 (1986).
- Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446**, 572–576 (2007).
- Johnson, S. M., Tan, F. J., McCullough, H. L., Riordan, D. P. & Fire, A. Z. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res.* **16**, 1505–1516 (2006).
- Field, Y. *et al.* Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.* **4**, e1000216 (2008). **This study shows that the encoded nucleosome organization is predictive of, and might be the cause of, differences in the transcriptional noise and activation dynamics of regulatory sequences and the utilization efficiency of DNA replication origins.**
- Yuan, G. C. & Liu, J. S. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput. Biol.* **4**, e13 (2008).
- Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genet.* **39**, 1235–1244 (2007).
- Peckham, H. E. *et al.* Nucleosome positioning signals in genomic DNA. *Genome Res.* **17**, 1170–1177 (2007).
- Korber, P., Luckenbach, T., Blaschke, D. & Horz, W. Evidence for histone eviction in *trans* upon induction of the yeast *PHO5* promoter. *Mol. Cell. Biol.* **24**, 10965–10974 (2004).
- Vignali, M., Hassan, A. H., Neely, K. E. & Workman, J. L. ATP-dependent chromatin-remodeling complexes. *Mol. Cell. Biol.* **20**, 1899–1910 (2000).
- Pavlovic, J., Banz, E. & Parish, R. W. The effects of transcription on the nucleosome structure of four *Dictyostelium* genes. *Nucleic Acids Res.* **17**, 2315–2332 (1989).
- Henikoff, S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nature Rev. Genet.* **9**, 15–26 (2008).
- Gasser, R., Koller, T. & Sogo, J. M. The stability of nucleosomes at the replication fork. *J. Mol. Biol.* **258**, 224–239 (1996).
- Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
- Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
- Bulyk, M. L., Gentalen, E., Lockhart, D. J. & Church, G. M. Quantifying DNA–protein interactions by double-stranded DNA arrays. *Nature Biotechnol.* **17**, 573–577 (1999).
- Mukherjee, S. *et al.* Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genet.* **36**, 1331–1339 (2004).
- Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
- Berger, M. F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **135**, 1266–1276 (2008).
- Zhu, C. *et al.* High-resolution DNA binding specificity analysis of yeast transcription factors. *Genome Res.* **19**, 556–566 (2009).
- Badis, G. *et al.* A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell* **32**, 878–887 (2008). **This study provides direct quantitative measurements of the sequence preferences of over 100 transcription factors from yeast.**
- Schroeder, M. D. *et al.* Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol.* **2**, e271 (2004).
- Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* **16**, 962–972 (2006).
- Miller, J. A. & Widom, J. Collaborative competition mechanism for gene activation *in vivo*. *Mol. Cell. Biol.* **23**, 1623–1632 (2003).
- Li, X. Y. *et al.* Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol.* **6**, e27 (2008).
- Shea, M. A. & Ackers, G. K. The OR control system of bacteriophage lambda. A physical–chemical model for gene regulation. *J. Mol. Biol.* **181**, 211–230 (1985).
- Hayes, J. J. & Wolffe, A. P. The interaction of transcription factors with nucleosomal DNA. *Bioessays* **14**, 597–603 (1992).
- Thomas, G. H. & Elgin, S. C. Protein/DNA architecture of the DNase I hypersensitive region of the *Drosophila* hsp26 promoter. *EMBO J.* **7**, 2191–2201 (1988).
- Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nature Genet.* **39**, 1512–1516 (2007).
- Hendrix, D. A., Hong, J. W., Zeitlinger, J., Rokhsar, D. S. & Levine, M. S. Promoter elements associated with RNA pol II stalling in the *Drosophila* embryo. *Proc. Natl Acad. Sci. USA* **105**, 7762–7767 (2008).

52. Mavrich, T. N. *et al.* Nucleosome organization in the *Drosophila* genome. *Nature* **453**, 358–362 (2008).
53. Mavrich, T. N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18**, 1073–1083 (2008).
54. Shivaswamy, S. *et al.* Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* **6**, e65 (2008).
55. Yuan, G. C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309**, 626–630 (2005).
The first high-resolution analysis of nucleosome positions in vivo at a genomic scale. The results show that there are stereotyped nucleosome organizations at promoters and add to the evidence that poly(dA:dT) tracts are unfavourable to nucleosome occupancy genome-wide.
56. Liu, X., Lee, C. K., Granek, J. A., Clarke, N. D. & Lieb, J. D. Whole-genome comparison of Leu3 binding *in vitro* and *in vivo* reveals the importance of nucleosome occupancy in target site selection. *Genome Res.* **16**, 1517–1528 (2006).
57. Sekinger, E. A., Moqtaderi, Z. & Struhl, K. Intrinsic histone–DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell* **18**, 735–748 (2005).
An early demonstration that particular regulatory sequences encode low nucleosome occupancy.
58. Valouev, A. *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).
59. Kornberg, R. D. & Stryer, L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.* **16**, 6677–6690 (1988).
This paper develops the idea of statistical nucleosome positioning. Simple steric exclusion together with the high density of nucleosomes along the genome leads to a periodic ordering of nucleosomes away from boundary constraints.
60. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).
61. Raveh-Sadka, T., Levo, M. & Segal, E. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res.* 19 May 2009 (doi:10.1101/gr.088260.108).
62. Khorasanizadeh, S. The nucleosome: from genomic organization to genomic regulation. *Cell* **116**, 259–272 (2004).
63. Narlikar, G. J., Fan, H. Y. & Kingston, R. E. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**, 475–487 (2002).
64. Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285–294 (1999).
65. Lam, F. H., Steger, D. J. & O’Shea, E. K. Chromatin decouples promoter threshold from dynamic range. *Nature* **453**, 246–250 (2008).
This paper shows experimentally that the relationship between nucleosome organization and the factor-binding affinity landscape determines the dynamics of activation of Pho genes in yeast.
66. Kim, H. D. & O’Shea, E. K. A quantitative model of transcription factor-activated gene expression. *Nature Struct. Mol. Biol.* **15**, 1192–1198 (2008).
67. Polach, K. J. & Widom, J. A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *J. Mol. Biol.* **258**, 800–812 (1996).
68. Giniger, E. & Ptashne, M. Cooperative DNA binding of the yeast transcriptional activator GAL4. *Proc. Natl Acad. Sci. USA* **85**, 382–386 (1988).
69. Tanaka, M. Modulation of promoter occupancy by cooperative DNA binding and activation-domain function is a major determinant of transcriptional regulation by activators *in vivo*. *Proc. Natl Acad. Sci. USA* **93**, 4311–4315 (1996).
70. Ma, X., Yuan, D., Diepold, K., Scarborough, T. & Ma, J. The *Drosophila* morphogenetic protein Bicoid binds DNA cooperatively. *Development* **122**, 1195–1206 (1996).
71. Rubin-Bejerano, I., Mandel, S., Robzyk, K. & Kassir, Y. Induction of meiosis in *Saccharomyces cerevisiae* depends on conversion of the transcriptional repressor Ume6 to a positive regulator by its regulated association with the transcriptional activator Ime1. *Mol. Cell Biol.* **16**, 2518–2526 (1996).
72. Ma, J. Crossing the line between activation and repression. *Trends Genet.* **21**, 54–59 (2005).
73. Felsenfeld, G. & Groudine, M. Controlling the double helix. *Nature* **421**, 448–453 (2003).
74. Robert, F. *et al.* Global position and recruitment of HATs and HDACs in the yeast genome. *Mol. Cell* **16**, 199–209 (2004).
75. Dion, M. F. *et al.* Dynamics of replication-independent histone turnover in budding yeast. *Science* **315**, 1405–1408 (2007).
76. Newman, J. R. *et al.* Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**, 840–846 (2006).
77. Zhang, Z., Gu, J. & Gu, X. How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? *Trends Genet.* **20**, 403–407 (2004).
78. Tirosh, I., Weinberger, A., Bezael, D., Kaganovich, M. & Barkai, N. On the relation between promoter divergence and gene expression evolution. *Mol. Syst. Biol.* **4**, 159 (2008).
79. Field, Y. *et al.* Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nature Genet.* **41**, 438–445 (2009).
This paper shows that DNA sequence changes that directly alter the intrinsically encoded nucleosome organization of the genome are associated with, and might be important drivers of, evolutionary divergence in gene expression patterns.
80. Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nature Genet.* **38**, 636–643 (2006).
81. Raser, J. M. & O’Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science* **304**, 1811–1814 (2004).
82. Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W. J. & Hartl, D. L. Genetic properties influencing the evolvability of gene expression. *Science* **317**, 118–121 (2007).
83. Tirosh, I., Weinberger, A., Carmi, M. & Barkai, N. A genetic signature of interspecies variations in gene expression. *Nature Genet.* **38**, 830–834 (2006).
84. Blake, W. J., M. K. A., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422**, 635–637 (2003).
85. Heichinger, C., Penkett, C. J., Bahler, J. & Nurse, P. Genome-wide characterization of fission yeast DNA replication origins. *EMBO J.* **25**, 5171–5179 (2006).
86. Kim, S. M., Zhang, D. Y. & Huberman, J. A. Multiple redundant sequence elements within the fission yeast *ura4* replication origin enhancer. *BMC Mol. Biol.* **2**, 1 (2001).
87. Rippe, K., von Hippel, P. H. & Langowski, J. Action at a distance: DNA-looping and initiation of transcription. *Trends Biochem. Sci.* **20**, 500–506 (1995).
88. Fyodorov, D. V. & Kadonaga, J. T. Dynamics of ATP-dependent chromatin assembly by ACF. *Nature* **418**, 897–900 (2002).
89. Yang, J. G., Madrid, T. S., Sevastopoulos, E. & Narlikar, G. J. The chromatin-remodeling enzyme ACF is an ATP-dependent DNA length sensor that regulates nucleosome spacing. *Nature Struct. Mol. Biol.* **13**, 1078–1083 (2006).
90. Whitehouse, I., Rando, O. J., Delrow, J. & Tsukiyama, T. Chromatin remodelling at promoters suppresses antisense transcription. *Nature* **450**, 1031–1035 (2007).
91. Rippe, K. *et al.* DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes. *Proc. Natl Acad. Sci. USA* **104**, 15635–15640 (2007).
92. von Hippel, P. H. & Delagoutte, E. A general model for nucleic acid helices and their ‘coupling’ within macromolecular machines. *Cell* **104**, 177–190 (2001).
93. Dechering, K. J., Cuelenaere, K., Konings, R. N. & Leunissen, J. A. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res.* **26**, 4056–4062 (1998).

Acknowledgements

We thank the members of our laboratories for discussions and critical comments on the manuscript. E.S. acknowledges research support from the National Institutes of Health and the European Research Council and J.W. acknowledges research support from the National Institutes of Health. E.S. is the incumbent of the Soretta and Henry Shapiro career development chair.

DATABASES

Saccharomyces Genome Database:
<http://www.yeastgenome.org>
 Abf1 | Pho4 | Reb1

FURTHER INFORMATION

Segal laboratory homepage: <http://genie.weizmann.ac.il>
 Widom laboratory homepage: <http://www.biochem.northwestern.edu/widomweb/index.html>
 Nature Reviews Genetics Series on Modelling:
<http://www.nature.com/nrg/series/modelling/index.html>

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Online summary:

- This Review presents a unifying quantitative and conceptual framework for translating DNA sequences into transcriptional behaviours.
- Each DNA-binding molecule has specific DNA sequence preferences (affinities) and, thus, every regulatory sequence defines a unique affinity landscape for each molecule.
- At given concentrations of DNA-binding molecules, the unique affinity landscape of a regulatory sequence dictates a distinct distribution of molecule-binding configurations and, consequently, a distinct transcriptional output.
- Accurate models of the DNA sequence preferences of nucleosomes and of many transcription factors are now available.
- The intrinsic DNA sequence preferences of nucleosomes are major determinants of nucleosome organization *in vivo*, and partly account for the depletion of nucleosomes around the starts and ends of genes.
- Nucleosome depletion around transcription factor-binding sites is partly encoded in the nucleosome affinity landscape of the genome and might assist in directing factors to their appropriate genomic sites.
- Differences in the intrinsic nucleosome affinity landscapes in which factor-binding sites are embedded might allow the same factor to regulate its different targets with different activation dynamics.
- Two factors that have adjacent binding sites can show indirect binding cooperativity through competition with nucleosomes, allowing some factors to function as activators on some regulatory sequences but as inhibitors on others.
- The affinity landscape of a regulatory sequence dictates when factors must compete with nucleosomes for access to the DNA, partly explaining the differential requirement for chromatin remodellers at different loci.
- DNA sequence changes that directly alter the nucleosome affinity landscape of a regulatory sequence might help to drive phenotypic diversity across evolution.
- Variability in cell-to-cell expression and in DNA replication can be partly explained in terms of the affinity landscape of a DNA sequence for transcription factors and nucleosomes.

Biographies

Eran Segal received his Ph.D. in computer science and genetics in 2004, from Stanford University, California, USA. In 2005, following postdoctoral research at the Center for Physics and Biology at Rockefeller University, New York, USA, he joined the faculty of the Weizmann Institute of Science in Rehovot, Israel, in the Departments of Computer Science and Applied Mathematics, and Molecular Cell Biology. The main focus of his laboratory is on developing quantitative models for gene regulation, using a combination of experimental and computational approaches.

Jonathan Widom received his Ph.D. in biochemistry in 1982 from Stanford University, California, USA. He was a postdoctoral fellow in structural studies at the MRC Laboratory of Molecular Biology in Cambridge, UK, from 1983–1985, when he joined the faculty of the University of Illinois at Urbana-Champaign, Illinois, USA, in the Departments of Chemistry, Biochemistry and Biophysics, and the Beckman Institute. In 1991, he moved to Northwestern University, Illinois, where he is William Deering Professor in the Departments of Biochemistry, Molecular Biology, Cell Biology and Chemistry. His research focuses on chromosome structure and gene regulation, and the biophysical chemistry of protein–DNA complexes.

Toc Blurb

000

From DNA sequence to transcriptional behaviour: a quantitative approach*Eran Segal and Jonathan Widom*

This Review presents a quantitative framework for translating DNA sequences into transcriptional behaviours. Such a model, based on the binding affinity landscape of molecules to genomic sequences, can help to describe complex phenomena such as transcriptional noise and the evolution of transcriptional regulation.