Published in final edited form as:

Nat Rev Genet. 2009 July ; 10(7): 443–456. doi:10.1038/nrg2591.

From DNA Sequence to Transcriptional Behavior: A Quantitative Approach

Eran Segal^{1,2,*} and Jonathan Widom^{3,*}

¹ Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, 76100, Israel

² Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, 76100, Israel

³ Department of Biochemistry, Molecular Biology, and Cell Biology, Northwestern University, 2205 Tech Drive, Evanston, IL 60208-3500 USA

Abstract

Complex transcriptional behaviors are encoded in the DNA sequences of gene regulatory regions. Advances in our understanding of these behaviors have been gained recently by quantitative models that describe how molecules such as transcription factors and nucleosomes interact with the genomic sequence. 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.

Many cellular and organismal processes depend on the establishment of complex patterns of gene expression at precise times and spatial locations, with inaccuracies in carrying out such transcriptional programs often being deleterious and leading to disease. The information for directing such expression patterns is encoded in regulatory DNA sequences — for example, reporter genes attached directly to such regulatory sequences adopt the expression pattern of the endogenous gene^{1,2,3}, and DNA binding and gene expression patterns of an entire human chromosome are essentially unchanged in mice that carry this human chromosome⁴.

Given the centrality of transcriptional programs to many biological processes, a predictive and quantitative understanding of the transcriptional behaviors encoded by DNA sequences is highly desirable. Such an understanding would allow us to go beyond merely identifying the transcription factors and regulatory DNA elements that are involved, and replace the existing qualitative and phenomenological descriptions by a mechanistic view of the process that integrates the involved components into physically realistic mechanistic models. Indeed, our ability to quantitatively predict the behavior of a regulatory system is a useful objective measure of the extent to which we truly understand how the system works. At a more practical level, the ability to accurately predict transcriptional behaviors 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; and it would allow for improved rational design of transgenes for biotechnology and gene therapy.

* Correspondence should be addressed to E.S. (E-mail: eran.segal@weizmann.ac.il) or J.W. (E-mail: j-widom@northwestern.edu).

Recent work has significantly advanced our understanding of how genomic sequences are translated into transcriptional outputs. Progress has been made possible by the availability of vast amounts of data on gene regulation, through the development of quantitative models that explain how molecules such as transcription factors^{5,6,7} and nucleosomes^{8,9} bind DNA sequences, and how these binding events give rise to expression patterns^{10,11}. In this review, we unify these studies into a single 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 the DNA-binding molecules, the range of affinities that the molecules have for any sequence (the binding affinity ‘landscape’) dictates their particular cooperative and competitive binding interactions. This leads to a distinct distribution of molecule BINDING CONFIGURATIONS on that sequence, and consequently to a distinct transcriptional behavior (Fig. 1).

Since transcriptional regulation across organisms utilizes 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 demonstrated in organisms from bacteria^{11,12,13,14} and yeast^{7,8,15,16,17} to flies¹⁰ and mammals^{18,19}.

We start by reviewing the significant progress made to understand the intrinsic affinities of various molecules to DNA. This has been achieved through experiments that directly measure the binding affinity landscapes of different types of molecule, and through 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. Substantially less is known about the mapping into transcriptional outputs, and we thus highlight where critical 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. 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 behavior from DNA sequence.

Binding affinity landscapes

Nucleosomes

Measurements of the sequence-dependent affinity of NUCLEOSOMES²⁰ carried out over the past two decades show that the differing affinities of nucleosomes for different DNA sequences can vary at least 5,000-fold^{21,22}. These differing affinities are thought to reflect the energetic cost of sharply bending the differing DNA sequences around the histone octamer to conform to the nucleosome structure²⁰.

Since direct nucleosome affinity measurements of all possible sequences of length 147 is not possible, 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 consisted of ~10bp periodicities of specific dinucleotides along the nucleosome length, first observed in alignments of ~200 in vivo nucleosome sequences from chicken²³, and then again in similarly-

sized collections from yeast^{8,24}, worm²⁵, fly²⁶, and human²⁶. 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 allowed researchers to identify longer sequence motifs that are generally favored or disfavored by nucleosomes, regardless of their position within the nucleosome length. Incorporating such signals into models that describe nucleosome sequence preferences has resulted in significantly improved predictions^{26,27,28,29}.

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 REMODELERS³¹, transcription³², DNA replication^{33,34}, and the sequence preferences of the nucleosomes themselves^{8,9,15,23,27,28}. A general question has therefore been the extent to which these models truly represent nucleosome sequence preferences, as opposed to also capturing the sequence preferences of 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 nucleosomes are governed only by the intrinsic sequence preferences of nucleosomes, 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-basepair 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 is highly predictive of *in vivo* nucleosome organization in both yeast and worm, demonstrating that nucleosome sequence preferences are a dominant determinant of *in vivo* nucleosome organizations¹⁵.

Transcription factors

Compared to nucleosomes, transcription factors recognize and bind much shorter stretches of DNA (typically 5–15 basepairs in length). It should therefore be theoretically feasible to measure directly the binding affinity of a given factor to most, if not all, possible recognition sequences. Earlier methods based on FOOTPRINTING, GEL-SHIFT ANALYSIS, SOUTHWESTERN BLOTTING, SELEX, and reporter constructs, could only measure the affinity of factors to a relatively small number of sequences. Significant progress was recently achieved with the use of high-throughput technologies such as CHIP-CHIP⁴ and CHIP-SEQ³⁵, which essentially measure all of the *in vivo* bound targets of a given factor. However, the genomic regions identified by these methods are typically hundreds of basepairs, and so identifying the much shorter sequence motifs that are common to the bound regions still requires post-processing computation³⁶.

Another limitation of using *in vivo* data is that, as in the nucleosome case, the derived binding specificities may also reflect the specificities of other factors. Here too, high-throughput *in vitro* methods such as PROTEIN BINDING MICROARRAYS^{37,38} and MICROFLUIDIC PLATFORMS³⁹ are being used, in which binding is measured across all possible ~8–10 basepair sequences and is governed only by the intrinsic sequence preferences of a factor. Protein binding microarrays were recently applied to derive such 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 trivial, the binding affinity of many factors for any location on any DNA sequence can now be characterized with high accuracy.

Comparing affinity landscapes: nucleosomes versus transcription factors

The intrinsic nucleosome affinity landscape explains many aspects of the binding patterns of nucleosomes *in vivo*¹⁵. In contrast, the sequence specificity of many transcription factors is

low compared to the size of the genome on which they operate, such that canonical binding sites for such factors will occur by mere happenstance countless times across the genome. For example, a factor whose total binding specificity is five basepairs will probably have over one million canonical binding sites in the human genome, yet the number of factor molecules present in the cell may typically be only one tenth to one thousandth that number. Thus, although a minimal level of binding affinity is a prerequisite for factor binding, most binding sites that meet such criteria occur by happenstance and are not bound *in vivo*; and 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 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 highly predictive models for transcriptional regulation.

From binding affinity landscapes to binding 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 behavior of a regulatory sequence is to understand the configurations of molecules that are *actually* bound to the sequence. Several quantitative frameworks^{5,10,43} addressed exactly this problem. These models consider all possible configurations of molecules on the input sequence. They then associate a statistical weight to each such 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 its statistical weight with the *partition function*, 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 affinity range 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 binding competition between molecules, resulting from steric hindrance constraints, is modeled explicitly. Third, conventional cooperative binding interactions can be modeled explicitly, by assigning higher statistical weights to configurations in which two molecules are bound in close proximity¹⁰. Finally, a novel 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 comprising both factors and nucleosomes, the locations at which nucleosomes intrinsically “want” to bind influences the locations at which factors will be bound; and conversely, the constellation of factors trying to bind influences where the nucleosomes will be bound, and with what occupancies. The DNA sequence defines the outcome of this competition, which will change in response to changing constellations of active transcription factors that are induced, for example, by cellular signaling during development or in response to a change in the environment.

Applications and limitations

The models above were used to identify regulatory sequences in fly⁵ and transcription-factor target interactions in human¹⁹, and for predicting, from regulatory sequences, expression patterns in the fly embryo¹⁰ and in yeast¹⁶. For example, in the segmentation gene network of the fly embryo, the model predicted that cooperative interactions and contributions from both strong and weak sites are important for generating the expression patterns¹⁰. These predictions were later supported by large-scale measurements of transcription factor binding in the fly embryo, which revealed prevalent binding to weak sites⁴⁶. Despite these successes, several

aspects of the modeling framework are not well understood, such as the effect of higher-order chromatin structure, the exact way in which factors compete with each other and with nucleosomes, and the actual 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, equal to the statistical weight of the configuration divided by the partition function. The success of quantitative modeling and prediction of gene regulation in prokaryotes^{13,14} has rested largely on this equilibrium hypothesis^{11,47}. However, the question of how and even whether regulatory systems truly equilibrate remains unclear. The equilibrium question is an especially daunting problem in eukaryotes, due to added complexities such as nucleosomes; yet, models based on an assumption of equilibrium competition also have high predictive value in eukaryotes^{10,16}, where ATP-dependent nucleosome remodeling mechanisms may contribute to facilitating or subverting such an equilibrium (Box 2).

From binding configurations to transcriptional output

The final step in modeling transcriptional behavior from sequence is to understand the transcriptional output that results from each configuration of bound molecules. Naively, 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 intuition: they model transcriptional output as being proportional to the binding probability of the transcription initiation machinery^{11,16}, or proportional to the weighted sum of the bound molecules, where activators have positive weights and repressors have negative weights¹⁰. These oversimplified approaches, however, do not model the effect 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 ignores the additional layers of regulation that are enabled by trans-acting factors, such as regulation that the transcriptional initiation complex undergoes after it is bound^{50,51}, regulation of transcriptional elongation, and the effects that nucleosomes positioned within the transcript may have on the elongation process. The quantitative details of these additional effects are poorly understood and thus, 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 will ‘feel’ 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, with each configuration weighted by its probability. Having presented the general modeling framework, we now review the experimental observations that it explains, starting with observations regarding nucleosome organizations.

Deciphering the determinants of nucleosome organizations in vivo

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

Direct consequences of the nucleosome landscape: distinct nucleosome positioning

This comparison revealed that the large nucleosome depletions around gene ends^{26,52,53,54} 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, two highly abundant transcription factors known to influence chromatin structure, the nucleosome affinity landscape encodes only minor nucleosome depletion; the large depletion seen around these factors' binding sites in vivo is therefore probably due to these factors' own ability to out-compete nucleosomes¹⁵. Nucleosome depletion around gene starts was also found to be encoded by the intrinsic nucleosome affinity landscape, but in this case, the action of chromatin remodelers and the binding of transcription factors and of the transcription initiation machinery also contribute measurably to the depletion¹⁵ (Fig. 2a).

Another notable feature of the in vivo nucleosome organization 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 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 a peaked nucleosome affinity landscape in the case of well-positioned regions, and a relatively flat landscape in the case of fuzzy regions. Indeed, both types of regions exist in the map of nucleosomes assembled on purified yeast DNA, and a model of nucleosome sequence preferences constructed from this yeast data is significantly correlated with the in vivo nucleosome organization of worm¹⁵ (Fig. 2b,c).

Indirect consequences of the nucleosome landscape: Long-range ordering of nucleosomes

The examples above represent cases where the nucleosome landscape directly accounts for the experimental observations. Other observations can also 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 gene starts, which decays with the distance from the gene start. There are strong nucleosome-disfavoring sequences and nucleosome-positioning sequences upstream and over gene starts, respectively²⁶, which act to create, with high probability, a nucleosome-depleted region upstream of the gene start, together with a well-positioned nucleosome over the gene start. Introducing boundary constraints, such as these nucleosome-disfavoring and positioning sequences, into the framework presented here, automatically results in a long-range periodic ordering of nearby nucleosomes, simply as 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 long-range ordering of nucleosomes nearby gene starts can also be explained by the intrinsic nucleosome landscape²⁶, but in this case, it can be due in part simply to indirect long-range effects that the sequences surrounding gene

starts exert on nucleosome configurations, at the high nucleosome concentrations that exist in vivo.

As another intriguing example, nucleosomes may 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) changes in different conditions or cell types, resulting in a different distribution of binding configurations (Fig. 2e). Thus, a single binding affinity landscape can in principle encode for many different distributions of binding configurations, depending on the different concentration of molecules.

Relating nucleosome landscapes and transcription factors

Explaining the repressive function of nucleosomes

Aside from explaining DNA-binding patterns of molecules such as nucleosomes, we need to understand the dynamic transcriptional behavior that genes exhibit in response to changes in the concentration of the regulating factors. The quantitative framework presented here can be used to directly read the DNA sequence and predict these responses: this is achieved by computing, from the encoded affinity landscape, the probability of factor binding at increasing factor concentrations. Consider first a hypothetical DNA sequence with a landscape for only one transcription factor, which in turn recognizes a single binding site. 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 (no intrinsic favored locations), they compete with the factor for binding, resulting in a lower probability of factor binding; a given level of 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^{62,63,64}.

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, activation dynamics depend on the nucleosome occupancy around the factor site, with activation occurring at 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 where Pho4 concentration is low, is predictive of an earlier onset (in both factor concentration and in time itself) for 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-disfavoring sequences generate low nucleosome occupancy in their immediate vicinity and indeed, for nearly all yeast transcription factors, the subset of their sites that are near nucleosome-disfavoring sequences have 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 may regulate its different targets with different activation dynamics, by positioning some of its sites near nucleosome-disfavoring sequences^{26,61} (Fig. 3c–e).

Many regulatory sequences contain multiple sites for multiple transcription factors. In sequences where 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 lower factor concentration. Indeed, such

cooperativity was demonstrated in yeast, between an endogenous yeast transcription factor and two foreign transcription factors from *E. 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 may be part of the mechanism of cooperative binding that was demonstrated in regulatory sequences from yeast^{16,68,69} and fly^{10,70} that contain clusters of factor binding sites. Intriguingly, since this obligate cooperativity occurs between any two factors, it may also explain why some transcription factors exhibit both activatory and inhibitory roles^{71,72}: for instance, a transcriptional repressor can seemingly act as an activator if its competition with nucleosomes promotes the binding of a nearby activator (Fig. 3f–h).

Encoding distinct modes of transcriptional regulation

Chromatin remodelers are important in transcriptional regulation⁷³ and target specific sets of genes⁷⁴. Although the mechanism by which chromatin remodelers are recruited to specific loci is not well understood, a recent study²⁶ suggested that the differential requirement for remodelers at different loci may be partly explained from the DNA sequence. This study defined two categories of yeast genes based only on sequence information. The differences in the affinity landscapes of these two gene categories suggest that they undergo different modes of regulation, according to whether transcription factors do or do not compete with nucleosomes for access to the DNA. Indeed, genes in the category where competition is predicted to take place 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 remodeling complexes⁷⁴ (Fig. 4).

Thus, by partitioning genes based on the affinity landscapes, two modes of transcriptional regulation can be identified that provide a partial explanation for the differential requirement for chromatin remodelers at different genomic loci. A corollary of these facts is that genomes can encode different dynamic responses – even to the same transcription factor – at two different regulatory sequences, by embedding the factor's cognate binding sites in two different nucleosome affinity landscapes.

Evolution of binding affinity landscapes

A genetic mechanism for achieving phenotypic diversity

The examples above demonstrate that many diverse aspects of transcriptional regulation can be understood directly from the affinity landscapes encoded in DNA sequences. Since 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 the encoded 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 indeed explain much of the expression divergence across yeast species⁷⁸.

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

nucleosome-depleted region, whereas in anaerobic yeast species, where cellular respiration genes are inactive under typical growth, the orthologous regulatory sequences encode a relatively nucleosome-occupied landscape. This suggests that DNA sequence changes that directly alter the encoded nucleosome affinity landscape of regulatory sequences may 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 so it is interesting to address whether these noise levels can be predicted directly from the regulatory sequence of a gene. TATA SEQUENCES are predictive of high noise, presumably by amplifying fluctuations in gene activation through facilitation of transcription re-initiation^{81,82,83,84}, whereas nucleosome-disfavoring sequences are predictive of low noise²⁶. This latter observation can be understood using the framework presented here, by examining the effect on activation dynamics that it predicts for sequences that contain nucleosome-disfavoring elements. As discussed above, regulatory target sites that are located close to nucleosome-disfavoring elements will be occupied at lower factor concentrations than are target sites that are far from such elements. Thus, the noisy regime, at which the probability of factor binding is ~ 0.5 , is reached at lower factor concentrations in target sequences where the site is near nucleosome-disfavoring elements⁶¹ (Fig. 5a–c). If we assume that the physiological concentration of the regulating factor is such that it binds both types of target sequences with sufficient probability, then at such a concentration, targets with nucleosome-disfavoring sequences have already ‘escaped’ the noisy regime, providing a plausible explanation for the relatively 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 exhibit less noise⁸¹. Since nucleosome-disfavoring elements create nucleosome-depleted regions, sequences that contain such elements have a lower requirement for and are less targeted by chromatin remodelers²⁶, which may result in more rapid transitions between the active and inactive states, providing another plausible explanation for the lower noise of these sequences.

Noise in replication initiation

The ideas above are based on modeling the binding of molecules to DNA and as such, they may apply beyond the context of transcriptional regulation. For example, DNA replication origins also exhibit cell-to-cell variability, with some origins initiating replication in most cell divisions and others initiating only occasionally. Thus, analogous to transcriptional noise, the framework predicts that origins that are close to nucleosome-disfavoring elements would have lower nucleosome occupancy, and thus be more accessible to the replication initiation complex and initiate replication with higher efficiency. Indeed, in fission yeast, where replication efficiency was measured⁸⁵, origins that are close to nucleosome-disfavoring sequences initiate with higher efficiency²⁶ (probability), and a systematic sequence deletion study around one replication origin found that deletion of a strong nucleosome-disfavoring 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. As one such 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 thus the noisy regime 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 behaviors. The key idea behind this translation process is that DNA-binding molecules have intrinsic affinities to DNA sequences that are specific to each molecule, and thus, every sequence defines a unique affinity landscape with respect to each molecule. Molecules that interact and bind DNA result in a unique distribution of molecule binding configurations at each sequence and a resulting transcriptional output.

Recent studies have determined the intrinsic affinities of nucleosomes and of many transcription factors, allowing us to translate DNA sequences into affinity landscapes with high accuracy. These landscapes, either directly, or through the application of the framework, partly explain many experimental observations regarding binding patterns of nucleosomes, dynamics of transcriptional activation, and transcriptional noise, demonstrating that diverse sets of transcriptional behaviors 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 modeling 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 within the same regulatory module to interact with each other⁸⁷. Regulatory modules may prove to 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. Experiments targeted at measuring 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 data needed to model the free energy costs of longer loops. Detailed studies of architectures of regulatory sequences carried out at high spatial resolution will also be required.

Current models assume that the system is in thermodynamic equilibrium and make other simplifying assumptions regarding steric hindrance and 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 posttranslational modifications, and the absence or presence of histone H1, do not significantly influence the nucleosome sequence preference; these assumptions, too, need to be validated or modified. The activity of key components such as chromatin remodelers has thus far 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. Since most of the experimentation and modeling has been done only in bacteria and yeast, it is important to test and develop the models in higher eukaryotes. Although the basic rules of molecular interactions modeled by current approaches should be universal, genomes of multi-cellular organisms may encode landscapes that combine these building blocks in ways that differ from unicellular organisms, leading to new design principles in transcriptional regulation. Finally, while we have focused here on approaches that

quantitatively model the translation of DNA sequences into transcriptional behaviors, a major remaining challenge is to understand the functional consequences that these expression patterns have, and ultimately, to be able to tell which deviations from the wild-type patterns may 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.

Box 1

Computing gene expression from DNA sequence

Several models were proposed^{10,11,16} for translating DNA sequences into transcriptional behavior. 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 (activators, repressors, etc.) to compute the probability of RNA polymerase occupancy, and then assume that the expression level is proportional to polymerase occupancy.

The computation is divided into two steps: one that computes the occupancy distribution of molecules on the target DNA sequence, and another that translates this occupancy distribution into a level of expression. Denoting each possible configuration of molecules on the DNA by c , and the probability of 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:

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

where $W(c)$ 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, a molecule's contribution 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:

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

where $\tau(m_i)$ is the concentration of the molecule bound at position p_i , F_0 is the statistical weight of the empty configuration, and

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

is 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 . The normalizing term $\sum_{c' \in C} W(c')$, also known as the *partition function*, sums over all possible legal configurations of molecules. The Box Figure illustrates the computation of $W(c)$ for one toy configuration.

Step 2: From occupancy distribution to expression level

The second model component, $P(E/c)$, which translates configurations into expression levels, is much less 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 acts to recruit the 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:

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

where 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 a particular binding molecules concentration. 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 such point is then combined to produce the entire expression pattern along the axis.

Box 2

The equilibrium assumption and the role of nucleosome remodelers

The models we review assume that molecules bind at thermodynamic equilibrium. Although models based on this assumption are highly predictive of gene regulation in both prokaryotes^{13,14} and eukaryotes^{10,16}, its validity has not been demonstrated. This assumption is particularly challenging to justify (or imagine) in eukaryotes, due to the presence of nucleosomes and many ATP-dependent nucleosome remodeling factors. In vitro, these nucleosome remodeling factors can control the spacing between nucleosomes on long DNA⁸⁸ and drive nucleosomes to disfavored locations on shorter DNA fragments, e.g., from favored positioning sequences to DNA ends⁸⁹. In vivo, inactivation of the ATP-dependent chromatin remodeling complex Isw2 leads to an average shift of ~15bp 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 remodeling activities could invalidate the equilibrium assumption on which the models discussed here are based, and require instead detailed and unique kinetic models for every sequence.

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

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

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

Glossary

BINDING CONFIGURATION

A particular arrangement of molecules along a DNA sequence, including 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 condition

NUCLEOSOME

The basic unit of chromatin, containing 147bp of DNA wrapped around a histone protein octamer

CHROMATIN REMODELERS

A protein or protein complex that has the capacity to alter the structure of chromatin. Some remodelers 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 polyacrylamide gel electrophoresis and transfer to a membrane, by their ability to bind to specific oligonucleotide probes

SYSTEMATIC EVOLUTION OF LIGANDS BY EXPONENTIAL ENRICHMENT (SELEX)

A combinatorial technique for producing DNAs that bind specifically and with high affinity to a DNA-binding protein of interest

CHIP-CHIP

ChIP-on-chip (also known as ChIP-chip) is a technique 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

Same as ChIP-chip, but where the resulting interactions are read out by high-throughput parallel sequencing and not by microarrays as in ChIP-chip

PROTEIN BINDING MICROARRAYS

A method that allows 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 spotted with a large number of different potential DNA-binding sites

MICROFLUIDIC PLATFORMS FOR MEASURING DNA-BINDING INTERACTIONS

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

TATA SEQUENCES

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

LEGAL CONFIGURATION

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

References

1. Casadaban MJ. Transposition and fusion of the lac genes to selected promoters in Escherichia coli using bacteriophage lambda and Mu. J Mol Biol 1976;104:541–555. [PubMed: 781293]doi:0022-2836(76)90119-4 [pii]

2. Guarente L, Ptashne M. Fusion of *Escherichia coli* lacZ to the cytochrome c gene of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 1981;78:2199–2203. [PubMed: 6264467]
3. Bellen HJ, et al. P-element-mediated enhancer detection: a versatile method to study development in *Drosophila*. *Genes Dev* 1989;3:1288–1300. [PubMed: 2558050]
4. Wilson MD, et al. Species-specific transcription in mice carrying human chromosome 21. *Science* 2008;322:434–438. [PubMed: 18787134]
5. Rajewsky N, Vergassola M, Gaul U, Siggia ED. Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics* 2002;3:30. [PubMed: 12398796]
6. Sinha S, van Nimwegen E, Siggia ED. A probabilistic method to detect regulatory modules. *Bioinformatics* 2003;19 (Suppl 1):i292–301. [PubMed: 12855472]
7. Granek JA, Clarke ND. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol* 2005;6:R87. [PubMed: 16207358]
8. Segal E, et al. A genomic code for nucleosome positioning. *Nature* 2006;442:772–778. [PubMed: 16862119]
9. Ioshikhes IP, Albert I, Zanton SJ, Pugh BF. Nucleosome positions predicted through comparative genomics. *Nat Genet* 2006;38:1210–1215. [PubMed: 16964265]
10. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* 2008;451:535–540. [PubMed: 18172436]
11. Bintu L, et al. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* 2005;15:116–124. [PubMed: 15797194]
12. Buchler NE, Gerland U, Hwa T. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A* 2003;100:5136–5141. 10.1073/pnas.0930314100 0930314100 [pii] [PubMed: 12702751]
13. Dodd IB, Shearwin KE, Sneppen K. Modelling transcriptional interference and DNA looping in gene regulation. *J Mol Biol* 2007;369:1200–1213. [PubMed: 17498740]doi:S0022-2836(07)00518-9 [pii] 10.1016/j.jmb.2007.04.041
14. Kuhlman T, Zhang Z, Saier MH Jr, Hwa T. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc Natl Acad Sci U S A* 2007;104:6043–6048. [PubMed: 17376875]doi: 0606717104 [pii] 10.1073/pnas.0606717104
15. Kaplan N, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*. 2008doi:nature07667 [pii] 10.1038/nature07667
16. Gertz J, Siggia ED, Cohen BA. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* 2009;457:215–218. [PubMed: 19029883]doi:nature07521 [pii] 10.1038/nature07521
17. Foat BC, Morozov AV, Bussemaker HJ. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* 2006;22:e141–149. [PubMed: 16873464]doi:22/14/e141 [pii] 10.1093/bioinformatics/btl223
18. Gupta S, et al. Predicting human nucleosome occupancy from primary sequence. *PLoS Comput Biol* 2008;4:e1000134. 10.1371/journal.pcbi.1000134 [PubMed: 18725940]
19. Sinha S, Adler AS, Field Y, Chang HY, Segal E. Systematic functional characterization of cis-regulatory motifs in human core promoters. *Genome Res* 2008;18:477–488. [PubMed: 18256240]
20. Richmond TJ, Davey CA. The structure of DNA in the nucleosome core. *Nature* 2003;423:145–150. [PubMed: 12736678]
21. Gencheva M, et al. In Vitro and in Vivo nucleosome positioning on the ovine beta-lactoglobulin gene are related. *J Mol Biol* 2006;361:216–230. [PubMed: 16859709]
22. Thastrom A, Bingham LM, Widom J. Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. *J Mol Biol* 2004;338:695–709. [PubMed: 15099738]
23. Satchwell SC, Drew HR, Travers AA. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* 1986;191:659–675. [PubMed: 3806678]
24. Albert I, et al. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* 2007;446:572–576. [PubMed: 17392789]

25. Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ. Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res* 2006;16:1505–1516. [PubMed: 17038564]
26. Field Y, et al. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol* 2008;4:e1000216. [PubMed: 18989395]
27. Yuan GC, Liu JS. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput Biol* 2008;4:e13. [PubMed: 18225943]
28. Lee W, et al. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* 2007;39:1235–1244. [PubMed: 17873876]
29. Peckham HE, et al. Nucleosome positioning signals in genomic DNA. *Genome Res* 2007;17:1170–1177. [PubMed: 17620451]
30. Korber P, Luckenbach T, Blaschke D, Horz W. Evidence for histone eviction in trans upon induction of the yeast PHO5 promoter. *Mol Cell Biol* 2004;24:10965–10974. [PubMed: 15572697]
31. Vignali M, Hassan AH, Neely KE, Workman JL. ATP-dependent chromatin-remodeling complexes. *Mol Cell Biol* 2000;20:1899–1910. [PubMed: 10688638]
32. Pavlovic J, Banz E, Parish RW. The effects of transcription on the nucleosome structure of four *Dictyostelium* genes. *Nucleic Acids Res* 1989;17:2315–2332. [PubMed: 2704621]
33. Henikoff S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nat Rev Genet* 2008;9:15–26. [PubMed: 18059368]doi:nrg2206 [pii] 10.1038/nrg2206
34. Gasser R, Koller T, Sogo JM. The stability of nucleosomes at the replication fork. *J Mol Biol* 1996;258:224–239. [PubMed: 8627621]doi:S0022-2836(96)90245-4 [pii] 10.1006/jmbi.1996.0245
35. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007;316:1497–1502. [PubMed: 17540862]
36. Harbison CT, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004;431:99–104. [PubMed: 15343339]
37. Bulyk ML, Gentale E, Lockhart DJ, Church GM. Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat Biotechnol* 1999;17:573–577. [PubMed: 10385322]
38. Mukherjee S, et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* 2004;36:1331–1339. [PubMed: 15543148]
39. Maerkl SJ, Quake SR. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 2007;315:233–237. [PubMed: 17218526]doi:315/5809/233 [pii] 10.1126/science.1131007
40. Berger MF, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* 2008;133:1266–1276. [PubMed: 18585359]doi:S0092-8674(08)00683-1 [pii] 10.1016/j.cell.2008.05.024
41. Zhu C, et al. High-resolution DNA binding specificity analysis of yeast transcription factors. *Genome Res*. 2009doi:gr.090233.108 [pii] 10.1101/gr.090233.108
42. 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* 2008;32:878–887. [PubMed: 19111667] doi:S1097-2765(08)00842-3 [pii] 10.1016/j.molcel.2008.11.020
43. Schroeder MD, et al. Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol* 2004;2:E271. [PubMed: 15340490]
44. Tanay A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*. 2006
45. Miller JA, Widom J. Collaborative competition mechanism for gene activation in vivo. *Mol Cell Biol* 2003;23:1623–1632. [PubMed: 12588982]
46. Li XY, et al. Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol* 2008;6:e27. [PubMed: 18271625]
47. Shea MA, Ackers GK. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J Mol Biol* 1985;181:211–230. [PubMed: 3157005]doi:0022-2836(85)90086-5 [pii]
48. Hayes JJ, Wolffe AP. The interaction of transcription factors with nucleosomal DNA. *Bioessays* 1992;14:597–603.10.1002/bies.950140905 [PubMed: 1365915]

49. Thomas GH, Elgin SC. Protein/DNA architecture of the DNase I hypersensitive region of the *Drosophila* hsp26 promoter. *Embo J* 1988;7:2191–2201. [PubMed: 2901349]
50. Zeitlinger J, et al. RNA polymerase stalling at developmental control genes in the *Drosophila* melanogaster embryo. *Nat Genet* 2007;39:1512–1516. [PubMed: 17994019]
51. Hendrix DA, Hong JW, Zeitlinger J, Rokhsar DS, Levine MS. Promoter elements associated with RNA Pol II stalling in the *Drosophila* embryo. *Proc Natl Acad Sci U S A* 2008;105:7762–7767. [PubMed: 18505835]
52. Mavrich TN, et al. Nucleosome organization in the *Drosophila* genome. *Nature* 2008;453:358–362. [PubMed: 18408708]
53. Mavrich TN, et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* 2008
54. Shivaswamy S, et al. Dynamic Remodeling of Individual Nucleosomes Across a Eukaryotic Genome in Response to Transcriptional Perturbation. *PLoS Biol* 2008;6:e65. [PubMed: 18351804]
55. Yuan GC, et al. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* 2005;309:626–630. [PubMed: 15961632]
56. Liu X, Lee CK, Granek JA, Clarke ND, Lieb JD. Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Res* 2006;16:1517–1528. [PubMed: 17053089]
57. Sekinger EA, Moqtaderi Z, Struhl K. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol Cell* 2005;18:735–748. [PubMed: 15949447]
58. Valouev A, et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* 2008
59. Kornberg RD, Stryer L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res* 1988;16:6677–6690. [PubMed: 3399412]
60. Schones DE, et al. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 2008;132:887–898. [PubMed: 18329373]
61. Raveh-Sadka, T.; Levo, M.; Segal, E. Proceedings of the 13th International Conference on Research in Computational Molecular Biology (RECOMB);
62. Khorasanizadeh S. The nucleosome: from genomic organization to genomic regulation. *Cell* 2004;116:259–272. [PubMed: 14744436]doi:S0092867404000443 [pii]
63. Narlikar GJ, Fan HY, Kingston RE. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* 2002;108:475–487. [PubMed: 11909519]doi:S0092867402006542 [pii]
64. Kornberg RD, Lorch Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 1999;98:285–294. [PubMed: 10458604]
65. Lam FH, Steger DJ, O'Shea EK. Chromatin decouples promoter threshold from dynamic range. *Nature* 2008;453:246–250. [PubMed: 18418379]
66. Kim HD, O'Shea EK. A quantitative model of transcription factor-activated gene expression. *Nat Struct Mol Biol* 2008;15:1192–1198. [PubMed: 18849996]doi:nsmb.1500 [pii] 10.1038/nsmb.1500
67. Polach KJ, Widom J. A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *J Mol Biol* 1996;258:800–812. [PubMed: 8637011]doi:S0022-2836(96)90288-0 [pii] 10.1006/jmbi.1996.0288
68. Giniger E, Ptashne M. Cooperative DNA binding of the yeast transcriptional activator GAL4. *Proc Natl Acad Sci U S A* 1988;85:382–386. [PubMed: 3124106]
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 U S A* 1996;93:4311–4315. [PubMed: 8633061]
70. Ma X, Yuan D, Diepold K, Scarborough T, Ma J. The *Drosophila* morphogenetic protein Bicoid binds DNA cooperatively. *Development* 1996;122:1195–1206. [PubMed: 8620846]
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* 1996;16:2518–2526. [PubMed: 8628320]

72. Ma J. Crossing the line between activation and repression. *Trends Genet* 2005;21:54–59. [PubMed: 15680515]doi:S0168-9525(04)00306-3 [pii] 10.1016/j.tig.2004.11.004
73. Felsenfeld G, Groudine M. Controlling the double helix. *Nature* 2003;421:448–453.10.1038/nature01411 nature01411 [pii] [PubMed: 12540921]
74. Robert F, et al. Global position and recruitment of HATs and HDACs in the yeast genome. *Mol Cell* 2004;16:199–209. [PubMed: 15494307]
75. Dion MF, et al. Dynamics of replication-independent histone turnover in budding yeast. *Science* 2007;315:1405–1408. [PubMed: 17347438]
76. Newman JR, et al. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* 2006;441:840–846. [PubMed: 16699522]
77. Zhang Z, Gu J, Gu X. How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? *Trends Genet* 2004;20:403–407.10.1016/j.tig.2004.07.006 S0168-9525(04)00186-6 [pii] [PubMed: 15313547]
78. Tirosh I, Weinberger A, Bezalel D, Kaganovich M, Barkai N. On the relation between promoter divergence and gene expression evolution. *Mol Syst Biol* 2008;4:159. [PubMed: 18197176] doi:msb4100198 [pii] 10.1038/msb4100198
79. Field Y, et al. Gene Expression Divergence in Yeast is Coupled to Evolution of DNA-Encoded Nucleosome Organization. *Nat Genet.* 2009In press
80. Bar-Even A, et al. Noise in protein expression scales with natural protein abundance. *Nat Genet* 2006;38:636–643. [PubMed: 16715097]
81. Raser JM, O'Shea EK. Control of stochasticity in eukaryotic gene expression. *Science* 2004;304:1811–1814. [PubMed: 15166317]
82. Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. Genetic properties influencing the evolvability of gene expression. *Science* 2007;317:118–121. [PubMed: 17525304]doi:1140247 [pii] 10.1126/science.1140247
83. Tirosh I, Weinberger A, Carmi M, Barkai N. A genetic signature of interspecies variations in gene expression. *Nat Genet* 2006;38:830–834. [PubMed: 16783381]
84. Blake WJ, M KA, Cantor CR, Collins JJ. Noise in eukaryotic gene expression. *Nature* 2003;422:633–637. [PubMed: 12687005]
85. Heichinger C, Penkett CJ, Bahler J, Nurse P. Genome-wide characterization of fission yeast DNA replication origins. *Embo J* 2006;25:5171–5179. [PubMed: 17053780]
86. Kim SM, Zhang DY, Huberman JA. Multiple redundant sequence elements within the fission yeast *ura4* replication origin enhancer. *BMC Mol Biol* 2001;2:1. [PubMed: 11178109]
87. Rippe K, von Hippel PH, Langowski J. Action at a distance: DNA-looping and initiation of transcription. *Trends Biochem Sci* 1995;20:500–506. [PubMed: 8571451]doi:S0968000400891173 [pii]
88. Fyodorov DV, Kadonaga JT. Dynamics of ATP-dependent chromatin assembly by ACF. *Nature* 2002;418:897–900.10.1038/nature00929 nature00929 [pii] [PubMed: 12192415]
89. Yang JG, Madrid TS, Sevastopoulos E, Narlikar GJ. The chromatin-remodeling enzyme ACF is an ATP-dependent DNA length sensor that regulates nucleosome spacing. *Nat Struct Mol Biol* 2006;13:1078–1083. [PubMed: 17099699]doi:nsmb1170 [pii] 10.1038/nsmb1170
90. Whitehouse I, Rando OJ, Delrow J, Tsukiyama T. Chromatin remodelling at promoters suppresses antisense transcription. *Nature* 2007;450:1031–1035. [PubMed: 18075583]
91. Rippe K, et al. DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes. *Proc Natl Acad Sci U S A* 2007;104:15635–15640. [PubMed: 17893337] doi:0702430104 [pii] 10.1073/pnas.0702430104
92. von Hippel PH, Delagoutte E. A general model for nucleic acid helicases and their “coupling” within macromolecular machines. *Cell* 2001;104:177–190. [PubMed: 11207360]doi:S0092-8674(01)00203-3 [pii]
93. Dechering KJ, Cuelenaere K, Konings RN, Leunissen JA. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res* 1998;26:4056–4062. [PubMed: 9705519]

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), where k is the number of basepairs 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 (bottom panel, left), the model uses these binding affinity landscapes to compute a probability distribution over configurations of bound molecules (see Box 1 for details); a small subset of these configurations are illustrated (bottom panel, centre). Configurations in which two bound molecules overlap are not allowed due to steric hindrance constraints, thereby modeling binding competition between molecules (see bottom-most configuration with probability 0). 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.

Figure 2. Main determinants of in vivo nucleosome organization

(a) Shown is the nucleosome occupancy *in vivo* in yeast (blue) and the nucleosome affinity landscape as measured *in vitro* by assembling purified histones on purified yeast genomic DNA¹⁵ (green), averaged across all genes. Occupancy around gene transcription start sites is shown on the left, and around gene translation end sites on the right. Also shown below each graph is a schematic illustration of the key components that contribute to the *in vivo* nucleosome occupancy. Nucleosome depletion around gene ends is largely encoded by the nucleosome affinity landscape, while nucleosome depletion around gene starts results both from the encoded nucleosome affinity landscape and from the binding action of transcription factors.

(b) Across one genomic region from worm with well-positioned nucleosomes, shown is the

average nucleosome occupancy for that region in vivo⁵⁸ (blue) and the average nucleosome affinity landscape for that region as predicted by a model constructed from in vitro data in yeast¹⁵ (green). **(c)** Same as (b), across a genomic region from worm with less-well defined nucleosome locations (“fuzzy nucleosomes”). The agreement between predictions of a model based on nucleosome sequence preferences and the experimental measurements, both at regions with well-positioned nucleosomes (b) and at regions with fuzzy nucleosomes (c), suggests that both types of regions may be encoded by the genomic sequence, through peaked nucleosome affinity landscapes (b) or relatively flat landscapes (c). **(d)** Nucleosome-disfavoring sequences can have a long-range effect on the nucleosome organization. This example sequence contains a strong nucleosome disfavoring sequence (yellow diamond), which are highly abundant in eukaryotic genomes⁹³. When such a nucleosome-affinity landscape is combined with a high nucleosome concentration, as is the case in vivo, the bound nucleosomes automatically organize into ordered arrays, whose order decays with the distance from the original disfavoring sequence (bottom graph and schematic bottom sequence). This phenomenon is termed ‘statistical positioning’⁵⁹. **(e)** Illustration of how a single sequence may potentially encode for different nucleosome organizations in different cell types or biological conditions, by encoding different outcomes of nucleosome-factor competition at different factor concentrations. Shown is a sequence having a uniform landscape for nucleosomes and a landscape for one factor that includes a single strong binding site. In condition 1, where the hypothetical factor is expressed at low levels, the most likely configurations have nucleosomes covering the factor binding site, whereas in condition 2, where 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.

Figure 3. Reading gene expression dynamics from DNA sequence

(a) Nucleosomes act as general repressors. Shown are two example sequences with a transcription factor landscape containing a single binding site, and with either a uniform but moderate-affinity landscape for nucleosomes (sequence '1') or a uniform but low-affinity landscape for nucleosomes (sequence '2'). (b) For the two sequences from (a), shown is the probability of transcription factor binding at different factor concentrations, computed by applying the framework presented here to the binding landscapes of those two sequences. (c) Nucleosome disfavoring sequences determine the threshold of activation. Shown are three example sequences with differing nucleosome and factor landscapes: ('1') a uniform nucleosome landscape; ('2') a landscape with a sequence that strongly disfavors nucleosome formation, located 10bp from the single transcription factor site ('2'); ('3') same as '2', but where the disfavoring sequence is located 135bp from the factor site. (d) The probability of transcription factor binding at each of the three sequences from panel c. (e) For each of the

three sequences from (c), shown is the most likely molecule binding configuration at three different factor concentrations (c). **(f)** Proximal factor sites exhibit cooperative or destructive binding. Shown are three example sequences with a uniform nucleosome affinity landscape and differing factor landscapes: ('1') a single factor site; ('2') two factor sites separated by 10bp; ('3') two factor sites separated by 135bp. **(g)** The probability of transcription factor binding to the left (red) site at each of the three sequences from (f). **(h)** Shown are 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 to sequence '1'.

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 by the absence of strong nucleosome disfavoring sequences and the presence of TATA sequences (left), and one by the presence of strong nucleosome disfavoring sequences and the absence of TATA sequences (right). Shown is the nucleosome occupancy in vivo (blue) and the nucleosome affinity landscape as measured in vitro by assembling purified histones on purified yeast genomic DNA¹⁵ (green), averaged across all genes of each gene set. 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. (b) Schematic illustration of the most likely configurations of each gene set. In gene set one (left), the nucleosome landscape exhibits high nucleosome occupancy and the transcription factor landscape has a relatively large number of binding sites spread across the regulatory region, suggesting that nucleosomes and factors are in competition for access to the DNA. Supporting this suggestion is the high transcriptional noise, high rate of histone turnover, and enrichment for chromatin remodeler activity, that were found for this gene set²⁶. In contrast, in gene set two (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 little competition between factors and nucleosomes, and supporting this is the low noise, low histone turnover, and absence of enrichment for chromatin remodeler targets, that were found for this gene set²⁶.

Figure 5. Explaining transcriptional noise from DNA sequence

(a) Nucleosome disfavoring sequences determine the range of factor concentrations at which high transcriptional noise occurs. Shown are two example sequences, one with a uniform nucleosome landscape ('1'), and one with a nucleosome landscape containing a sequence that strongly disfavors nucleosome formation, located 10bp from the single transcription factor site ('2'). (b) For the two sequences from (a), shown is the probability of transcription factor binding at different factor concentrations, 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 (highlighted by the brown rectangle), since at this regime the variance of factor binding is maximal. Note, however, that while the variance of factor binding is one of the determinants of noise levels, other determinants exist as well. (c) For four different factor concentrations (c), shown are the most likely molecule binding configurations at each of the two sequences from (a). Note that at each of the two intermediate concentrations, one of the two sequences is noisy, i.e., the configurations in which the factor is bound and the configurations in which the factor is not bound have near-equal probability. (d) Cooperative binding reduces the range of factor concentrations at which there is high transcriptional noise. Shown are two example

sequences with a uniform nucleosome landscape, where one sequence has a single factor site ('1'), and the other two factor sites separated by 10bp ('2'). (e) The probability of transcription factor binding to the left (red) site at each of the two sequences from⁶¹ (d). The regime of high noise is highlighted (brown rectangle). The range of factor concentrations at which each sequence exhibits high 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'.