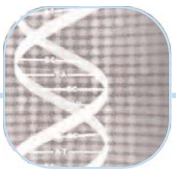


CHAPTER SEVEN

Modeling regulatory motifs

Sridhar Hannenhalli

Biological processes are mediated by specific interactions between cellular molecules (DNA, RNA, proteins, etc.). The molecular identification mark, or signature, required for precise and specific interactions between various biomolecules is not always clear, a comprehensive knowledge of which is critical not only for a mechanistic understanding of these interactions but also for therapeutic interventions of these processes. The biological problem we will address here, stated in general terms, is: how do biomolecules accurately identify their binding partners in an extremely crowded cellular environment? An important class of cellular interactions concerns the recognition of specific DNA sites by various DNA binding proteins, e.g. transcription factors (*TF*). Precisely how the TFs recognize their DNA binding sites with high fidelity is an active area of research. While a detailed treatment of this question covers several areas of investigation, we will focus on aspects of the TF–DNA recognition signal that is encoded in the DNA binding site itself. In this chapter we will summarize a number of approaches to model DNA sequence signatures recognized by transcription factor proteins.

1 Introduction

Most biological processes critically depend on specific interactions between biomolecules. A key question in biology is how, in the overly crowded cellular environment, these various interactions are accomplished with high fidelity. Evidence suggests highly developed mechanisms for trafficking, addressing, and recognizing biomolecules within a cell. For instance, brewer's yeast (*Saccharomyces cerevisiae*) feeds on galactose, among other sugars. The yeast needs a mechanisms to sense the presence of galactose in its environment and in response, turn on specific biological

Bioinformatics for Biologists, ed. P. Pevzner and R. Shamir. Published by Cambridge University Press.

processes to harness galactose. In the presence of galactose, transcriptional regulator protein GAL4 binds to a specific DNA sequence upstream of several genes, most notably GAL2, involved in galactose metabolism [1]. This entire process, from the sensing of galactose to transmitting information down the signal cascade that culminates in the binding of GAL4 to the GAL2 gene's regulatory sequence and metabolizing galactose, requires many specific interactions between different types of molecules including DNA, RNA, and proteins.

As another example, consider the well-studied JAK-STAT signal transduction pathway which plays a critical role in cell fate decision and immune response in humans. Much like galactose metabolism in yeast, the JAK-STAT system involves sensing specific chemicals outside the cell, transmitting this information across the cell membrane down to the regulatory regions of specific genes, to activate the response system [2]. One can think of such signaling pathways as a relay involving specific interactions starting with the interaction between extracellular chemicals and cell-membrane receptors, culminating in the interaction between transcription factors and DNA in gene regulatory sequences. Questions concerning the specificity of interaction between biomolecules are open in most contexts and are areas of active research.

The problem of interaction specificity could be resolved from first principles if we had two pieces of information, namely the location of an interaction partner and certain identifying features of the partner. For instance, if you were to plan a meeting with a stranger in a large city, you would need to know the approximate meeting location (e.g. corner of 6th and Broad), as well as certain identifying features of this person (e.g. red polka dot suit). A parallel in the cellular environment could be a trans-membrane (location) protein with amino acid sequence HHRHK near the amino terminus (identifying feature). In this example, the identifying feature could also be expressed as a stretch of five positively charged and largely hydrophobic residues. Alternatively, one of the interacting proteins may have a structural feature (the key) which fits a complementary structure on another protein (the lock). These examples provide three different ways of representing the *identifying feature* of the interacting partner, or in other words, these examples are different "models" of the interaction specificity. Based on the different models one can surmise that the task of modeling substrate specificity can be extremely difficult, especially in the realm of proteins. Indeed, the task is complex even for the much simpler case in which the substrate is a nucleic acid molecule (DNA or RNA). While the general principles are common to both proteins and nucleic acids, for the sake of simplicity, we will restrict the exposition to nucleic acids hereafter. In particular, we will discuss the issue of modeling the DNA sites recognized and bound by transcription factors (TF), i.e. transcription factor binding sites (TFBS). To orient the reader, we next provide a brief introduction to transcriptional regulation.

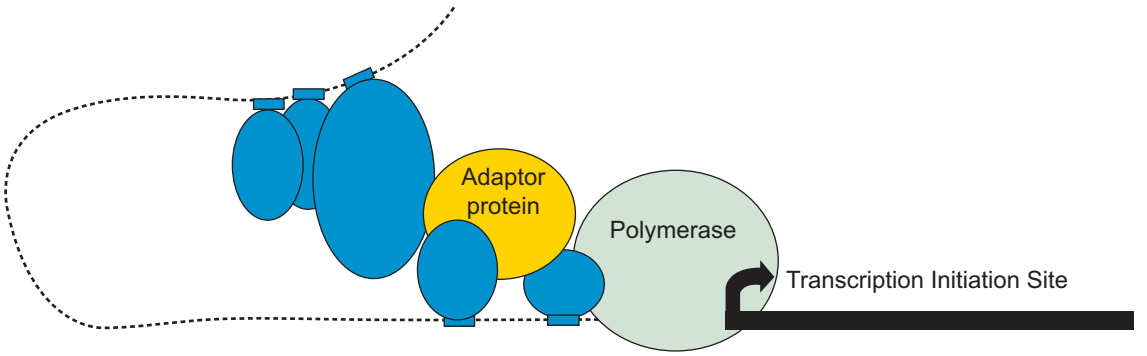


Figure 7.1 Transcription factor proteins (filled ellipses) interact with binding sites (filled rectangles) in the relative vicinity of a gene transcript (black rectangle). The transcription factor binding sites can either be proximal to the transcript (within a few thousand nucleotides) or far (several hundred thousand nucleotides). The interactions between transcription factors is aided by other adaptor proteins. The DNA-bound transcription factors interact with polymerase to regulate transcription.

How much, at what time, and where within an organism any gene product is produced is precisely regulated, and is critical to maintaining all life processes. While the overall regulation of a gene product is executed at various levels – including splicing, mRNA stability, export from nucleus to cytoplasm and translation – much of this regulation is accomplished at the level of transcription. Transcriptional regulation is a fundamental cellular process, and aberrations in this process underlie many diseases [3]. For example, mutations in the *Factor IX* protein is known to cause hemophilia B. Additionally, mutations in the regulatory region immediately upstream of *Factor IX* gene can disrupt the binding of specific TF, which in turn dysregulates the transcription of the gene, thus leading to hemophilia [3]. In eukaryotes, transcriptional regulation is orchestrated by numerous TF proteins. For the most part, TFs regulate gene transcription by binding to specific short DNA sequences in the relative vicinity of the transcription start site of the target gene, and through interactions with each other as well as with the polymerase enzyme. See Figure 7.1 for a schematic.

Precise and specific interaction between the TF and its cognate DNA binding site is a critical aspect of transcriptional regulation. What is the identifying characteristic of the DNA sites recognized by a TF protein? This question remains an open and important one in modern biology. The specific TF–DNA interaction is determined not only by the DNA sequence but also by a number of additional cellular factors. A full description of these determinants is beyond the scope of this chapter. Here we focus on the aspect of TF–DNA interaction that is encoded in the sequence of the DNA binding site itself.

In particular, we will focus on models of TF binding sites. Given several instances of experimentally determined binding sites for a TF, a *model* is a succinct quantitative description of the known binding sites, which not only may provide mechanistic insights into TF–DNA interaction, but also helps identify novel binding sites. Although we have focused our discussion only on TF binding sites, the discussion applies to any DNA signal such as splice sites, polyA sites, and indeed more generally to signals in amino acid sequences. Finally, the signal encoded in the DNA binding site provides only part of the information required for specific interactions with the DNA binding protein. We will conclude with a discussion of additional hallmarks of functional binding sites that can be exploited specifically to identify functional TF binding sites in the genome.

2 Experimental determination of binding sites

In this section we will briefly summarize the experimental techniques used to determine the DNA binding sites for a specific TF. The sequences obtained from these experiments are then used to construct a model of TF binding. For a detailed review on this topic we refer the reader to [4]. The experimental approaches to binding site determination can be classified as *in vitro* and *in vivo*.

The common *in vitro* techniques include Systematic Evolution of Ligands by EXponential enrichment (SELEX) [5] and protein-binding DNA microarrays [6]. SELEX works as follows. One begins by synthesizing a large library consisting of randomly generated oligonucleotides of fixed length. The solution containing the oligonucleotides is exposed to the TF of interest. Some of the oligonucleotides bind to the TF. The oligomers that are bound by the TF can be separated from the rest (although not perfectly) and a new solution is prepared that is enriched for the bound oligomers. This process of binding to the TF and separating out the bound oligomers is repeated multiple times and in every new round the experimental conditions are varied so that the increasingly stronger binding between the TF and oligomers is favored. Multiple rounds of selection with increasing stringency for the binding results in a solution enriched for oligonucleotides that bind to the TF with high affinity. These oligonucleotides are then cloned and sequenced. In a related experimental techniuqe of protein-binding DNA microarray, the DNA oligomers are immobilized on a glass surface to which a flourescent-labeled TF is exposed. The specific oligomers that bind to the TF of interest are detected through optical signal processing [6]. This approach obviates the need for multiple rounds of enrichment as in SELEX, as well as the need for cloning and sequencing. By their nature, the *in vitro* capture the protein–DNA binding in purified form and in isolation, independent of the other cellular determinants of the binding.

In vivo identification of binding sites is accomplished by two common techniques – *ChIP-chip* and *ChIP-seq*. Both approaches require obtaining the nuclear DNA bound by the TF of interest, followed by DNA digestion, which leaves the TF attached to small stretches of DNA, and then using specific antibody to fish out the TF along with the stretch of DNA bound to it. In the ChIP-chip (Chromatin immunoprecipitation followed by microarray hybridization), the bound DNA is hybridized against a glass array that contains a large set of sequences corresponding to various genomic locations. Thus, the array elements that hybridize to the TF-bound DNA automatically provide the information on the genomic location where the TF binds. In the second technique – ChIP-seq (ChIP followed by high-throughput sequencing) – the microarray hybridization step is replaced by direct sequencing of the TF-bound DNA. The sequences are then mapped to the genome based on sequence similarity. In each of these approaches the TF-bound region is detected with varying resolution, and additional techniques are applied to more precisely map the boundaries of the TF binding sites.

Experimentally determined binding sites are compiled in various databases, most notably TRANSFAC [7] and JASPAR [8]. TRANSFAC is a licensed database which currently includes binding sites for over 1,000 TFs gleaned from the experimental literature. Each individual binding site is assigned a quality score corresponding to the strength of experimental evidence. JASPAR is a freely accessible resource which includes information on ∼150 TFs, also curated from experimental literature, and is based on a more stringent set of criteria as compared to TRANSFAC.

3 Consensus

For the rest of the chapter, we will assume that for a given TF we are provided a set of binding sites of a fixed length, and we will focus on the task of modeling these known sites. Therefore, for a transcription factor F, assume that we are given N examples of K bases long DNA sequences bound by F. Denote the N sequences as $X_1, X_2, \ldots, X_N$. Denote the nucleotide base at position j of sequence X_i by $X_{i,j}$, where $X_{i,j} \in \{A, C, G, T\}$. The DNA sequence characteristics that are critical for the protein–DNA interaction have both biological and computational implications. These characteristics should determine the representation of binding specificity. Consider Example 7.1a in which we are provided with 10 experimentally determined binding sites for the yeast TF *Leu3* [9], and each site is 10 nucleotides long.

Example 7.1.

(a)

	1	2	3	4	5	6	7	8	9	10
X_1	C	C	G	G	T	A	C	C	G	G
X_2	C	C	T	G	T	A	C	C	G	G
X_3	C	C	G	C	T	A	C	C	G	G
X_4	C	C	G	G	A	A	C	C	G	G
X_5	G	C	G	G	T	A	C	C	G	G
X_6	C	C	G	T	T	A	C	C	G	G
X_7	C	C	G	C	A	A	C	C	G	G
X_8	C	C	T	G	A	A	C	C	G	G
X_9	G	C	G	G	T	A	A	C	G	G
X_{10}	C	C	G	C	T	A	C	A	G	G

(b)

	1	2	3	4	5	6	7	8	9	10
A	0.0	0.0	0.0	0.0	0.3	1.0	0.1	0.1	0.0	0.0
C	0.8	1.0	0.0	0.3	0.0	0.0	0.9	0.9	0.0	0.0
G	0.2	0.0	0.8	0.6	0.0	0.0	0.0	0.0	1.0	1.0
T	0.0	0.0	0.2	0.1	0.7	0.0	0.0	0.0	0.0	0.0

(c)

A simple and common approach to summarize these known binding sites is called the *consensus* representation in which we create a consensus string of length K and place in position j the *consensus* nucleotide which occurs with the highest frequency at position j in N binding sites. In Example 7.1a, for instance, at position 3 there are 8 Gs and 2 Ts. Thus the consensus at position 3 is G. The consensus sequence of these 10 known examples of binding sites is thus $CCGGTACCGG$. Note that the consensus sequence happens to be the same sequence as X_1.

More formally, given N binding sites, each of length K, let $N_{x,j}$ be the number of binding sites sites having nucleotide x at position j, where $x \in \{A, C, G, T\}$ and

$1 \leq j \leq K$. The normalized frequency of nucleotide x at position j is denoted by $f_{x,j} = (N_{x,j})/N$. Clearly,

$$\sum_{x \in \{A,C,G,T\}} f_{x,j} = 1. \tag{7.1}$$

The consensus sequence of these N binding sites is defined as the K-long nucleotide sequence $C_1 C_2 \cdots C_K$, in which C_j is the nucleotide x that maximizes $f_{x,j}$. The consensus at each position in Example 7.1a is unambiguously defined. However, consider a case where at some position there are 4 Cs, 5 Gs, 1 A and 0 T. In this case, assigning a G as the consensus ignores the fact that nucleotide C is *almost* as likely as G. To address this ambiguity one may use letter S at this position of the consensus string where S represents *strong* bases C and G. Similarly, nucleotides A and G (*purines*) together are represented by letter R. There is an *International Union of Pure and Applied Chemistry* (IUPAC) letter code to denote each combination of nucleotides and which is used to represent consensus in general [10].

Although quite useful for many practical situations, the consensus representation is restrictive as it systematically ignores the rare bases at each position, which might represent biologically important instances of binding sites. Next we discuss the *Position Weight Matrix* representation of binding sites that addresses this specific shortcoming of the consensus model.

4 Position Weight Matrices

The *Position Weight Matrix* (PWM) is currently the most common representation of TF binding sites. Unlike the consensus approach, a PWM captures all observed bases at each position. In its simplest form, a PWM is a probability matrix with 4 rows corresponding to the 4 nucleotide bases and K columns corresponding to each position in the binding site. We will refer to rows 1 through 4 interchangeably as rows A, C, G, T, respectively. The entry corresponding to the jth column (position) and xth row (base) is $f_{x,j}$, defined above as the frequency of x at position j among the binding sites. The PWM corresponding to the binding sites in Example 7.1a is shown in 7.1b.

Note that if there is an insufficient number of known binding sites, i.e. if N is relatively small, then a particular nucleotide base may not be observed at a position. This would result in $f_{x,j} = 0$, which can be interpreted to imply that x is prohibited at position j, even though we know that this is simply due to insufficient sampling of sites and not because of a functional impossibility. A typical solution to deal with this situation is to correct for potentially unobserved data by adding a *prior* (also known as

pseudo count) to the observed nucleotide counts before computing the frequencies. A simple approach is to add a count of 1 to each observed count, also called the *Laplace prior*. If a Laplace prior is used in Example 7.1a, then the counts in the first column become (*1, 9, 3, 1*) for (*A, C, G, T*), and the first column of the PWM in Example 7.1b becomes (*0.071, 0.644, 0.214, 0.071*). Formally, under the Laplace prior, the frequencies are $f_{x,j} = (N_{x,j} + 1)/(N + 4)$.

There is a quantitative property of a PWM that corresponds to its usefulness in modeling the TF–DNA binding preference. For instance, if the known binding sites for a TF are highly dissimilar to each other, then there is very little knowledge to be gained about the general binding preference. More specifcially, consider a particular column j of a PWM. If each of the 4 nucleotides is equally likely to be observed at that position, i.e. if $f_{x,j} = 0.25$, for each nucleotide base x, then this column conveys no information regarding the binding preference of the TF under consideration. This intuitive notion of information contained in position j of a PWM can be quantified formally using the *Information Content*, which is measured in bits and is defined as

$$I_j = 2 + \sum_{x \in \{A,C,G,T\}} f_{x,j} log_2(f_{x,j}). \tag{7.2}$$

Note that in the most informative case, when exactly one of the nucleotides, say A, is observed at a position with $f_{A,j} = 1$, $f_{C,j} = 0$, $f_{G,j} = 0$, $f_{T,j} = 0$, then I_j achieves its maximum value of 2 bits.[1] In the other extreme, when all nucleotides are equally likely and $f_{x,j} = 0.25\ \forall x \in \{A, C, G, T\}$, then I_j achieves its minimum value of 0 bits [11]. One can verify that any other value of probabilities yields a positive information. Example 7.1c shows the *Logo* representation of the motif in Example 7.1b depicting the information content at each position. The x-axis enumerates the binding site positions and the y-axis indicates the information content. The height of each base corresponds to its relative frequency. The figure was generated using the Weblogo tool at *weblogo.berkeley.edu*. For a more detailed discussion on information content and another relative measure called *Relative entropy*, the reader is referred to [12].

While the PWM is a simple, intuitive, and the most commonly used model of TF–DNA interaction, its main drawback is that it assumes independence among different positions in the binding site. Specifically, the preference for a nucleotide at one position has no bearing on the nucleotide preferences at another position. Consider the hypothetical Example 7.2 below which has six binding sites, each four nucleotides long.

[1] Here, the value of $0 log_2 0$ is approximated to be 0.

Example 7.2.

X_1	$C\ G\ G\ G$
X_2	$C\ G\ T\ G$
X_3	$C\ G\ G\ C$
X_4	$A\ T\ G\ G$
X_5	$A\ T\ G\ G$
X_6	$A\ T\ G\ T$

In the first column, nucleotides, C and A are equally likely, while in the second column nucleotides G and T are equally likely. Based on this information and assuming independence between these two columns, one would infer that the two binding sites $CGGG$ and $CTGG$ are equally preferred. However, it is more likely that when there is a C at the first position a G is preferred in the second position, and when there is an A at the first position a T is preferred in the second position. In other words, the first and second positions are not independent. A direct experimental measurement of such dependence is laborious. Two specific experimental studies that infer dependence between positions in binding sites can be found in [13] for bacterial Mnt repressor binding sites and in [14] for Egr1 transcription factor binding sites.

5 Higher-order PWM

In Example 7.2, there is likely to be dependence between the first two positions. In this case the preferred binding sites can be better modeled, and thus better predicted, if we consider the first two nucleotides together. For instance, CG and AT are the most likely dinucleotides at the first two positions. In general, if we want to incorporate possible dependencies between nucleotides at every pair of adjacent positions, we can extend the single nucleotide PWM with 4 rows and K columns to a dinucleotide PWM with 16 rows corresponding to all 16 nucleotide combinations and $K-1$ columns corresponding to all dinucleotide positions. Therefore, in the first column of Example 7.2, the CG and AT dinucleotides will have large frequency values, each "close" to 0.5 each,[2] and all other 14 dinucleotides will have low values, "close" to zero. This dinucleotide-based PWM has also been referred to as the *Position Weight Array* [15, 16]. One can extend the *Position Weight Array* to capture even higher-order dependencies, say among L consecutive nucleotides. This corresponds to enumerating at every position of the binding site the L nucleotides-long sequences starting at the

[2] The probabilities will be "close" to 0.5, as opposed to being exactly 0.5, if we add small pseudocounts for the unobserved dinucleotides.

position among all binding sites, i.e. from positions 1 through L, positions 2 through $L+1$, and so on till positions $K-L+1$ through K. This results in a PWM with 4^L rows (corresponding to all possible K-long sequences) and $K-L+1$ columns for any $L \geq 1$, where L represents the number of adjacent nucleotides considered together. This model is equivalent to a *Markov Model* of order $L-1$, which provides the probability of observing a nucleotide at any position based on the previous $L-1$ nucleotides. See Figure 7.3b for an example of a first-order Markov Model. The Markov Model is a general statistical tool and is often used to model a variety of molecular sequences.

The main limitation of these higher-order PWMs is a lack of sufficient data, i.e. small values of N. For instance, we cannot reliably infer the preference for a dinucleotide among the 16 possible choices based on only 6 sequences, as in Example 7.2. Moreover, high-order PWMs are still limited in that they do not directly capture the dependence between non-adjacent nucleotide positions, for instance between positions 1 and 3, independent of position 2. In theory, this can be remedied by explictly enumerating nucleotide combinations for various combinations of positions, although such models suffer from insufficient data to a much greater extent than higher-order PWM models. In the next section we will discuss richer models of TF–DNA binding preferences that attempt to maximize the information captured from the data.

6 Maximum dependence decomposition

The *Maximum Dependence Decomposition* (MDD) approach, proposed in Genscan [16], explicitly estimates the extent to which the nucleotide at position j depends on the nucleotide at position i. Specifically, MDD estimates the extent to which the nucleotide at position j depends on whether the nucleotide at position i is the consensus (most frequent) nucleotide for that position or a non-consensus nucleotide. For each i all binding site sequences are divided into two groups, C_i and $\overline{C_i}$, depending on whether the nucleotide at position i is the consensus or a non-consensus base, respectively. Within each group the nucleotide frequencies are computed at every position j. For a given position j, the two sets of frequencies are compared using the χ^2 statistic [17]. If position j is independent of position i, then we expect the two sets of nucleotide frequencies to be fairly similar; however, if the two sets of frequencies differ significantly from each other, it would suggest that nucleotide preference at position j depends on the nucleotide at position i. Let f_A, f_C, f_G, and f_T be the normalized frequencies (number of each base divided by the total number of sequences) of the four bases at position j among the sequences in $\overline{C_i}$. Let N be the total number of sequences in C_i. If

the four bases were distributed identically in the two sets of sequences C_i and $\overline{C_i}$, then we would expect the number of the four bases at position j among the sequences in C_i to be $N * f_A$, $N * f_C$, $N * f_G$, and $N * f_T$. Let N_A, N_C, N_G, and N_T be the observed number of the four bases at position j among the sequences in C_i. In this context, the χ^2 statistic is defined as:

$$\frac{(N * f_A - N_A)^2}{N * f_A} + \frac{(N * f_C - N_C)^2}{N * f_C} + \frac{(N * f_G - N_G)^2}{N * f_G} + \frac{(N * f_T - N_T)^2}{N * f_T} \quad (7.3)$$

The greater the difference in the two sets of nucleotide frequencies, the higher the value of χ^2 statistic. If the statistic indicates a significant difference[3] between the two frequency distributions then the position j is said to depend on position i. For example, for a set of 20 sequences, if position 1 includes 12 As and 8 Gs, then the consensus C_1 is A. Now for the 12 sequences in which the nucleotide at position 1 is an A, assume that at position 2, 8 have a C and 4 have a T. On the other hand, for the 8 sequences in which the nucleotide at position 1 is a G, at position 2, 7 have a T and 1 has a C. For the sequences with $C_1 = A$, the counts for (*A, C, G, T*) at position 2 are (*0, 8, 0, 4*), and for the other 8 sequences the nucleotide counts at position 2 are (*0, 1, 0, 7*). Intuitively, the two sets of counts look very different from each other, and the χ^2 statistic formally quantifies this intuition.

Denote the χ^2 statistic quantifying the dependence of position j on position i as $\chi^2(j \mid i)$. The MDD approach proceeds iteratively as follows.

1. Compute $S_i = \sum_{j \neq i} \chi^2(j, i)$ to capture the total dependence on position i.
2. Among all K positions, select position i with the maximum value of S_i, and partition all sequences into two parts based on whether they have C_i or $\overline{C_i}$ at position i.
3. Repeat steps 1 and 2 separately for each of the two sets of sequences obtained in step 2.
4. Stop if there is no significant dependence, or if there is an insufficient number[4] of sequences in the current subset. In either case, construct a standard PWM for the remaining subset of sequences.

Figure 7.2a illustrates the MDD modeling procedure. The above procedure decomposes the entire binding site data set into a tree-like structure. To test whether a given sequence X fits the model, as illustrated in Figure 7.2b, one proceeds down the tree,

[3] If there is no real difference between the two frequency distributions then the χ^2 statistic is expected to follow the so-called χ^2 distribution. By comparing the computed χ^2 value to the expected distrbution, one can compute the probability that the two distributions are identical. This probability is called the P-value. If the P-value is small, say below 5%, then we can say that the two distributions are significantly different.

[4] We leave this purposefully vague, as there is no formal rule to define this. Essentially, if the number of remaining sequences is small, say below 5, then it does not pay to further partition them.

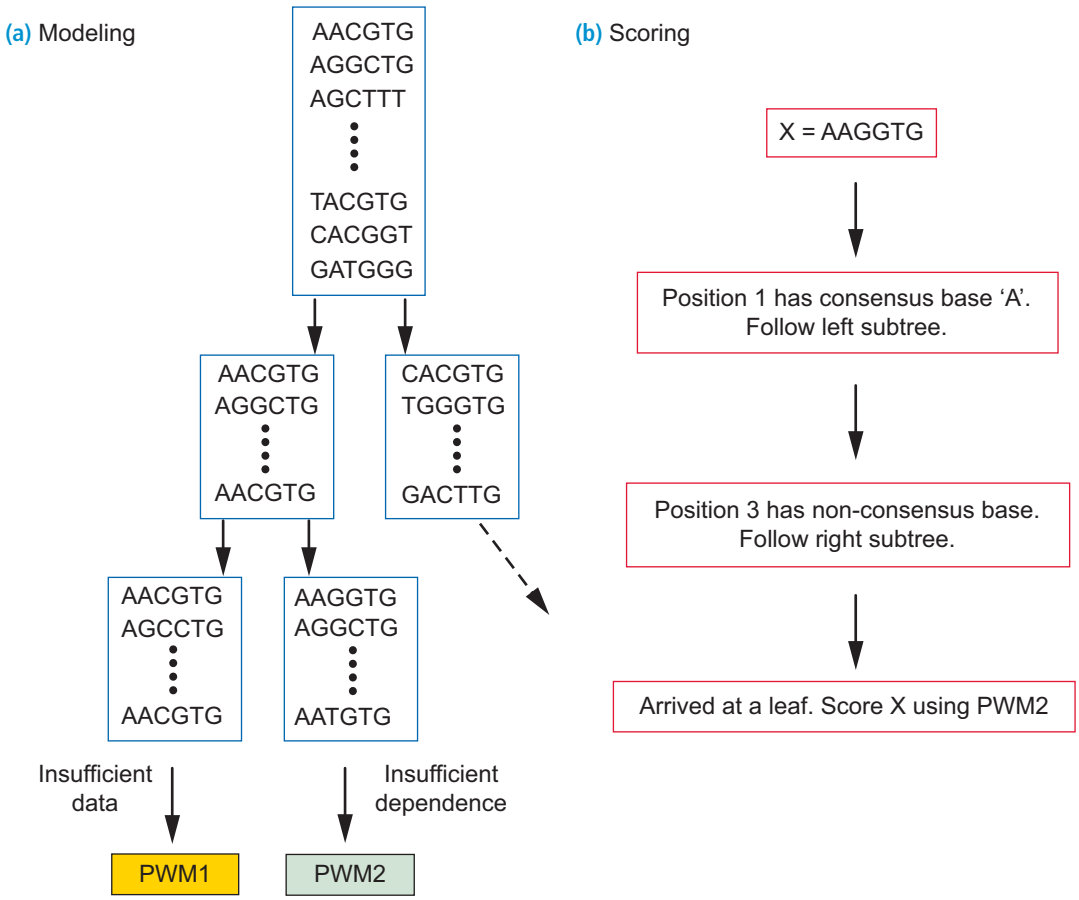


Figure 7.2 The figure, adapted from [16], illustrates the maximal dependency decomposition (MDD) procedure. (a) Modeling. Starting with all binding sites, maximum dependency is detected for position 1 with consensus "A." The sites are then partitioned based on whether or not the nucleotide at position 1 is an "A." Among the sites with "A" in the first position, maximum dependency is detected for position 3 with consensus "C." The sites are further partitioned based on whether or not the nucleotide at position 3 is a "C." The two partitions are not partitioned any further, however, because of either insufficient data or insufficient dependency. The entire MDD model is built following this procedure. (b) Scoring. Given a sequence X, one proceeds down the left subtree because the first base of X is an "A," followed by the right subtree because the third base is not a "C." At this stage, because a leaf is encountered, X is scored using PWM2, corresponding to the current leaf.

where a decision is made at each internal branching point based on whether a specific position of X is a consensus base or not, guiding the search down the appropriate descendent branches of the tree. The search eventually stops at a leaf which corresponds to a PWM, the one that "best" represents the sequence X.

Unlike the *Position Weight Array* mentioned above, which assumes dependence between every pair of adjacent positions, MDD is not restricted to adjacent positions and explicitly evaluates whether there is a statistical dependence between any two positions. However, it is easy to see that MDD requires a large number of sequences.

7 Modeling and detecting arbitrary dependencies

In this section we will discuss a general *Bayesian* approach developed in [18] to model dependencies between arbitrary pairs of binding site positions. In this approach, each of the K binding site positions may depend on any arbitrary set of other positions. This scenario can be best illustrated using a graph structure. Consider a network with K nodes $(s_1, s_2, \cdots, s_K)$ corresponding to the positions i through K, where x_i is a random variable representing the nucleotide at position i. We draw an arrow (a directed edge) from node s_i to s_j if the nucleotide at position j depends on the nucleotide at position i; dependence can be determined using the χ^2 statistic. Figure 7.3 shows a few dependency structures for $K = 4$. Consider the simplest case, with 4 nodes and no edges depicted in Figure 7.3a, such that each of the nucleotides is independent, which is precisely the PWM model. In probabilistic terms, the probability of observing a specific binding site $x_1x_2x_3x_4$ is the product of the four independent probabilities, i.e. $P(x_1x_2x_3x_4) = P(x_1)P(x_2)P(x_3)P(x_4)$, where $P(x_i)$ is the entry in the PWM at column i, for nucleotide x_i.

Now consider the dependency shown in Figure 7.3b with three edges. The first position is independent of any other position, while every other position depends on the previous position. In probabilistic terms, $P(x_1x_2x_3x_4) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3)$, where the notation $P(u|v)$ represents the probability of u conditional on the value of v. This is precisely the first-order Markov Model and is similar to the Weighted Array Matrix model mentioned above. The probability of each nucleotide at the first position is calculated in a fashion identical to that of a PWM. The conditional probabilities can then be derived from the given set of sites in a similar fashion. For instance, if among 10 sequences that have an A at the first position, three have a C at the second position, then $P(x_2 = C \mid x_1 = A) = 0.3$.

Figure 7.3c depicts a more complex dependency structure among the binding site positions. In this case position 2 depends on position 1. Position 3 depends on both positions 1 and 4, while positions 1 and 4 are independent of any other positions. We can write out the probability of observing a DNA sequence $x_1x_2x_3x_4$ as $P(x_1x_2x_3x_4) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1, x_4)P(x_4)$. Similar to the previous case, we can compute the conditional probability $P(x_3|x_1, x_4)$ by computing the fraction of different nucleotides at position 3 for various combinations of dinucleotides at positions 1 and 4. Finally,

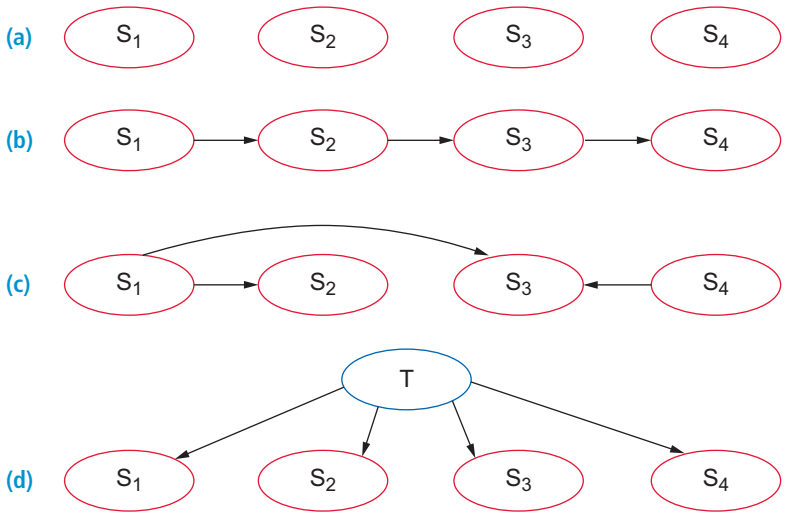


Figure 7.3 The figure illustrates a few possible dependency structures between the binding site positions (adapted from [18]).

Figure 7.3d illustrates a scenario where the nucleotides at the four bases are independent of each other but depend on an extrinsic variable T. For instance, certain TF are known to recognize distinct classes of motifs and the variable T may represent the motif class which in turn determines the nucleotide preferences at the four positions. It is not difficult to see that any arbitrary dependency structure defines a unique model, and given a model, one can precisely estimate the probability of observing a DNA sequence. However, there are a large number of possible dependency structures, and determining all possible dependency structures is not at all trivial. Incidentally, this problem is also encountered in other areas of computational biology, notably when inferring regulatory networks from gene expression data. The issue of searching for the *optimal model* is discussed in more detail in chapter 16 on biological network inference.

8 Searching for novel binding sites

The eventual goal of any model of TF–DNA binding is to efficiently and accurately assess whether an arbitrary sequence is likely to bind to the TF, and more generally, to identify potential binding site locations along a long stretch of DNA, possibly an entire genome. For consensus models, the search entails a simple scan of the DNA sequences for a perfect match, or a match with a limited number of mismatches to the consensus sequence. However, in the case of PWMs, detecting the binding sites is less straightforward.

8.1 A PWM-based search for binding sites

Essentially each sequence is assigned a "match" score which represents quantitatively its similarity to the PWM. For a PWM, a scoring function can simply be the product of nucleotide frequencies at each position. For instance, the match score for *CCGGTACCGG* (sequence X_1 in Example 7.1a) and using the PWM in Example 7.1b can be computed as $0.8 \times 1.0 \times 0.8 \times 0.6 \times 0.7 \times 1.0 \times 0.9 \times 0.9 \times 1.0 \times 1.0 = 0.22$. This quantity represents the probability that the sequence confers to, or is generated by, the PWM. Such a raw score is interpreted (*is this score sufficiently large to indicate a match of the PWM to the binding site?*) in the context of a specific background. For instance, a PWM in which, at every position, the bases "C" or "G" have the highest probability, is expected to achieve a high raw score while searching a region of the genome that is composed mostly of "C" and "G". In this case, an even higher raw score should be required.

Various software tools employ different strategies to select a threshold for the raw score. The MATCH software adapted from [19] employs the following strategy. Let r denote the raw match score for a PWM for a binding site. The raw score r is first converted into a percentile score p. If the minimum and maximum achievable scores by the PWM are r_{min} and r_{max}, then $p = (r - r_{min})/(r_{max} - r_{min})$. MATCH then searches an input sequence for matches whose percentile score surpasses a user-defined threshold. The default thresholds are based on a carefully chosen background to optimize either the false-negative rate, the false-positive rate, or the sum of both types of errors. Another strategy is to convert the raw score into a P-value, which estimates the random expectation of observing the raw score (or higher). For instance, Levy and Hannenhalli use a direct empirical approach. For a PWM, raw scores for every position on the entire genome (of the species of interest) on either strand are computed. This empirically estimated background distribution of raw scores provides a direct way to compute the frequency with which a score of at least r is expected by chance. If a score of at least r is achieved Q times, then the P-value of this score is estimated as Q/L, where L is the total length of the genome including both strands [20]. The other models that incorporate higher-order dependency between positions can be used to assign a score to novel DNA sequences analogously, and will not be discussed here.

8.2 A graph-based approach to binding site prediction

In Example 7.1a, it is intuitive that the first sequence $X_1 = CCGGTACCGG$ should have a high-affinity interaction with the TF, since it is not only known to bind to the TF, but it is also the consensus sequence. Given a model, we can compute a score for a sequence indicative of the binding probability or binding affinity. We discussed

above how this score is computed for a PWM. While in Example 7.1a, the consensus sequence happens to be among one of the sequences known to bind the TF, this is often not the case. More problematic and perhaps counterintuitive is the fact that with probabilistic models, such as PWM, a sequence that is not among the known examples may score better than a sequence known to bind the TF. Naughton *et al.* provide a simple illustrative example [21]. Consider three known examples of binding sites for a TF – *AAA*, *AAA*, and *AGG*. If we construct a PWM based on these three sequences, the score for sequence *AAG* would be $1.0 \times 0.67 \times 0.33 = 0.22$ while the score for *AGG* will be $1.0 \times 0.33 \times 0.33 = 0.11$. Interestingly, the sequence *AAG*, which is not known to bind to the TF, has a higher score than the sequence *AGG*, which is known to bind the TF. The problem is that in order to score a sequence, the probabilistic models use "average" properties of the known sites and not the known sites themselves. To address this shortcoming of probabilistic models, Naughton *et al*. proposed a graph-based approach for scoring a sequence directly from the known binding sites without building an explicit model. The intuition behind their approach is as follows. Assume that we wish to score a sequence X using N distinct sequences known to bind to the TF. Each of the N sequences additively contributes to the score for X, and the individual score contribution is a product of two components. The first component is proportional to the similarity between the sequences X and Y, where Y is one of the N sequences. The second component is proportional to the number of times Y occurs among the known binding sites. Thus the score contribution is high if there is a sequence very similar to X among the known sequences and there are many known instances of this sequence. The details of the precise function used can be found in [21].

9 Additional hallmarks of functional TF binding sites

TF binding sites are typically short (5–15 bp) and various binding sites for a TF can vary substantially. The DNA binding site sequence alone often does not contain sufficient information to explain the specificity with which a TF binds to its cognate binding sites. Thus, on the one hand, there are numerous locations in a genome that harbor DNA sequences strongly matching the TF–DNA binding model, and yet do not seem to bind to the TF in experiments; on the other hand, there are numerous locations experimentally known to be bound to a TF and yet which do not contain any sequences that could be predicted by the TF–DNA interaction model. Therefore, the match to a TF–DNA model, such as a PWM, is only one of the many determinants of functional TF–DNA interactions. There are several other hallmarks of TF binding sites that can be employed to improve the accuracy of binding site identification. Below we briefly

mention two such features. Additional determinants of functional TF–DNA interaction are discussed below.

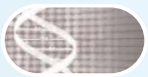
9.1 Evolutionary conservation

Consider a region of the genome that encodes for an important organismal function. Any mutation in this region affecting the specific function may be deleterious to the fitness of the organism and should be purged by evolution. In other words, such a region is likely to be evolving under purifying selection and will thus be conserved across species during evolution. The same principle applies to regulatory regions of the genome that harbor TF binding sites. *Phylogenetic footprints* are non-protein-coding regions of the genome that are highly conserved and are much more likely to be evolving under purifying selection [22]. Due to the recent availability of numerous alignable genome sequences, phylogenetic footprinting has been widely used to identify binding sites [20, 23, 24]. For a detailed review of phylogenetic footprinting we refer the reader to [25]. Although using evolutionary conservation is an effective way to reduce the false-positive rate in binding site prediction, exclusive reliance on conservation is limited for two reasons. First, conserved regions may sometimes be functionally neutral and thus may not harbor an important binding site [26]. Second, several functional binding sites are known not to be conserved, as shown by several studies [27, 28].

9.2 Modular interactions between TFs

Eukaryotic gene regulatory programs achieve complexity through combinatorial interactions among TF. For instance, the expressions of some of the *Drosophila* genes involved in development are regulated through combinatorial interactions among five TF proteins, *Bcd*, *Cad*, *Hb*, *Kr*, and *Kni* [29]. Consistent with the interactions between the TFs, the binding sites for these TF occur in clusters in the regulatory regions of the genes [30]. It seems that binding sites that occur in clusters are more likely to be functional. Thus the prediction of individual binding sites can be improved when subsumed within a search for binding site clusters. Several tools have been developed to detect significant clusters of binding sites in the genome [31, 32]. A cluster of functionally interacting binding sites, typically with multiple instances in the genome (presumably regulating several functionally related genes) is referred to as a *cis-regulatory module* (CRM) [33, 34]. Knowledge of CRMs can aid in accurate identification of individual binding sites [35]. Numerous computational approaches have been proposed to identify CRMs [25, 36–38]. Studies suggest that the binding of a TF to a binding site may depend on the presence or absence of binding sites for other TFs in the relative vicinity [39, 40]. Thus binding sites for a TF can be predicted with greater accuracy if one takes

into account the presence/absence of binding sites of specific interacting TF. Binding models have been proposed to exploit such sequence contexts [41, 42].



DISCUSSION

The general problem of accurately identifying transcription factor binding sites is important for a mechanistic understanding of transcriptional regulation. In this chapter we have focused on the narrower problem of modeling the TF–DNA interaction based only on a set of experimentally determined binding site sequences without any other information about the genomic or cellular context. An ideal model should be such that (1) the true DNA binding sites fit the model very well, i.e. the model is *sensitive*, and (2) the DNA sequences that are known not to bind the TF should not fit the model, i.e. the model is *specific*. Moreover, the model should be biologically interpretable. The PWM model, while being simple, does not capture potential dependencies between binding site positions. A full dependence model, on the other hand, is difficult to estimate reliably based only on a small number of exemplar binding sites. Despite the efforts and advances made over the last several years our ability to predict binding sites on a genome scale remains unsatisfactory.

Ultimately, any sequence-based model of TF–DNA interaction does not capture the inherently dynamic cellular state. For instance, how tightly the DNA at any given location on the chromosome is packaged on the nucleosomes, critically determines the TF–DNA interaction and, more generally, transcriptional regulation [43, 44]. It is possible that even a high-affinity binding site may not bind the TF, if the binding site location is tightly wrapped around a nucleosome, which are the basic unit of DNA packaging. Narlikar *et al.* were able to significantly improve the *de novo* motif discovery accuracy by exploiting nucleosome occupancy [45]. Histone modifications can also help identify the condition-specific chromatin structure and can help improve the genome-wide identification of binding sites. Recent application of high-throughput technologies, most notably ChIP-seq [46], have been used to generate genome-wide maps of histone modifications [47–49]. Lastly, post-translational modification states of TF proteins can critically alter the TF–DNA interaction [50]. However, how these modifications affect TF–DNA interaction is not well understood. Improvements in computational modeling of TF–DNA interaction is likely to come from a better biological understanding of these various determinants of TF–DNA interactions coupled with the development of tools that can integrate the heterogeneous information.

QUESTIONS

(1) Consider the following probability matrix representing the DNA binding specificity of a transcription factor.

	1	2	3	4	5
A	0.01	0.10	0.97	0.95	0.50
C	0.03	0.05	0.01	0.01	0.10
G	0.95	0.05	0.01	0.03	0.10
T	0.01	0.80	0.01	0.01	0.30

Calculate the information content (IC) for position 3 and position 5. Briefly explain what information content means and why there is such a difference in this value between positions 3 and 5. In other words, what characteristic of position 5 makes its IC so low, while the IC of position 3 is so high?

(2) What is the consensus binding site for the transcription factor in problem (1)?

(3) Based on the consensus sequence, can you find the most likely binding sites for the TF in the following DNA sequence: ACCAAGTAGATTACTT? Consider both the forward and reverse strands. Now which of these sites is the most likely if you consider the probability matrix above?

(4) Analogous to transcription factors, which bind to DNA, RNA binding proteins (RBP) bind to specific RNA molecules, such as mRNA. They regulate critical aspects of post-transcriptional processing of the mRNA. Much like TF–DNA interaction, RBP–RNA interaction is believed to be specific. What aspects of the target mRNA are likely to be important for specific RBP–RNA interaction?

REFERENCES

[1] J. M. Huibregtse, P. D. Good, G. T. Marczynski, J. A. Jaehning, and D. R. Engelke. Gal4 protein binding is required but not sufficient for derepression and induction of gal2 expression. *J. Biol. Chem.*, 268: 22219–22222, 1993.

[2] D. Hebenstreit, J. Horejs-Hoeck, and A. Duschl. Jak/stat-dependent gene regulation by cytokines. *Drug News Perspect.*, 18: 243–249, 2005.

[3] J. Villard. Transcription regulation and human diseases. *Swiss Med. Wkly*, 134: 571–579, 2004.
[4] L. Elnitski, V. X. Jin, P. J. Farnham, and S. J. Jones. Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. *Genome Res.*, 16: 1455–1464, 2006.
[5] C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249: 505–510, 1990.
[6] M. L. Bulyk. Protein binding microarrays for the characterization of DNA-protein interactions. *Adv. Biochem. Eng. Biotechnol.*, 104: 65–85, 2007.
[7] V. Matys, O. V. Kel-Margoulis, E. Fricke, *et al.* TRANSFAC and its module TRANSCOMPEL: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34: D108–D10, 2006.
[8] A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32: D91–D94, 2004.
[9] X. Liu and N. D. Clarke. Rationalization of gene regulation by a eukaryotic transcription factor: Calculation of regulatory region occupancy from predicted binding affinities. *J. Mol. Biol.*, 323: 1–8, 2002.
[10] A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984. *Nucl. Acids Res.*, 13: 3021–3030, 1985.
[11] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188: 415–431, 1986.
[12] G. D. Stormo. DNA binding sites: Representation and discovery. *Bioinformatics*, 16: 16–23, 2000.
[13] T. K. Man, J. S. Yang, and G. D. Stormo. Quantitative modeling of DNA-protein interactions: Effects of amino acid substitutions on binding specificity of the MNT repressor. *Nucl. Acids Res.*, 32: 4026–4032, 2004.
[14] M. L. Bulyk, P. L. Johnson, and G. M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucl. Acids Res.*, 30: 1255–1261, 2002.
[15] M. Q. Zhang and T. G. Marr. A weight array method for splicing signal analysis. *Comput. Appl. Biosci.*, 9: 499–509, 1993.
[16] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268: 78–94, 1997.
[17] M. J. Campbell and D. Machin. *Medical Statistics: A Commonsense Approach.* 3rd edn. Wiley, Chichester 2002.
[18] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan. Modeling dependencies in protein-DNA binding sites. In: Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology, Berlin, Germany. ACM Press, New York, 2003, 28–37.

[19] K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner. Matind and matinspector: New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.*, 23: 4878–4884, 1995.

[20] S. Levy and S. Hannenhalli. Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome*, 13: 510–514, 2002.

[21] B. T. Naughton, E. Fratkin, S. Batzoglou, and D. L. Brutlag. A graph-based motif detection algorithm models complex nucleotide dependencies in transcription factor binding sites. *Nucl. Acids Res.*, 34: 5730–5739, 2006.

[22] D. A. Tagle, B. F. Koop, M. Goodman, *et al.* Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.*, 203: 439–455, 1988.

[23] W. W. Wasserman and J. W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, 278: 167–181, 1998.

[24] X. Xie, J. Lu, E. J. Kulbokas, *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRS by comparison of several mammals. *Nature*, 434: 338–345, 2005.

[25] W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 5: 276–287, 2004.

[26] M. A. Nobrega, Y. Zhu, I. Plajzer-Frick, V. Afzal, and E. M. Rubin. Megabase deletions of gene deserts result in viable mice. *Nature*, 431: 988–993, 2004.

[27] E. T. Dermitzakis and A. G. Clark. Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. *Mol. Biol. Evol.*, 19: 1114–1121, 2002.

[28] E. Emberly, N. Rajewsky, and E. D. Siggia. Conservation of regulatory elements between two species of *Drosophila*. *BMC Bioinformatics*, 4: 57, 2003.

[29] D. Niessing, R. Rivera-Pomar, A. La Rosee, *et al.* A cascade of transcriptional control leading to axis determination in *Drosophila*. *J. Cell. Physiol.*, 173: 162–167, 1997.

[30] B. P. Berman, Y. Nibu, B. D. Pfeiffer, *et al.* Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. U S A*, 99:757–762, 2002.

[31] M. Rebeiz, N. L. Reeves, and J. W. Posakony. Score: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc. Natl Acad. Sci. U S A*, 99: 9888–9893, 2002.

[32] S. Sinha, E. Van Nimwegen, and E. D. Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19 Suppl. 1, I292–I301, 2003.

[33] M. Z. Ludwig, N. H. Patel, and M. Kreitman. Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: Rules governing conservation and change. *Development*, 125: 949–958, 1998.

[34] H. Bolouri and E. H. Davidson. Modeling DNA sequence-based cis-regulatory gene networks. *Dev. Biol.*, 246: 2–13, 2002.

[35] O. Hallikas, K. Palin, N. Sinjushina, *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, 124: 47–59, 2006.
[36] J. W. Fickett and W. W. Wasserman. Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.*, 11: 19–24, 2000.
[37] S. Hannenhalli. Eukaryotic transcriptional regulation: Signals, interactions and modules. In N. Stojanovic (ed.) *Computational Genomics*. Horizon Bioscience, Norfolk, 2007, 55–82.
[38] S. Hannenhalli. Eukaryotic transcription factor binding sites – Modeling and integrative search methods. *Bioinformatics*, 24: 1325–1331, 2008.
[39] A. Hochschild and M. Ptashne. Cooperative binding of lambda repressors to sites separated by integral turns of the DNA helix. *Cell*, 44: 681–687, 1986.
[40] S. Lomvardas and D. Thanos. Nucleosome sliding via TBP DNA binding in vivo. *Cell*, 106: 685–696, 2001.
[41] D. Das, N. Banerjee, and M. Q. Zhang. Interacting models of cooperative gene regulation. *Proc. Natl Acad. Sci. U S A*, 101: 16234–16239, 2004.
[42] L. Wang, S. Jensen, and S. Hannenhalli. An interaction-dependent model for transcription factor binding. In: *Lecture Notes in Computer Science*. Volume 4023. Springer, Berlin/Heidelberg, 2005, 225–234.
[43] W. Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447: 425–432, 2007.
[44] M. M. Suzuki and A. Bird. DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.*, 9: 465–476, 2008.
[45] L. Narlikar, R. Gordan, and A. J. Hartemink. A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS. Comput. Biol.*, 3: e215, 2007.
[46] P. J. Park. Chip-seq: Advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, 10: 669–680, 2009.
[47] A. Barski, S. Cuddapah, K. Cui, *et al.* High-resolution profiling of histone methylations in the human genome. *Cell*, 129: 823–837, 2007.
[48] D. E. Schones, K. Cui, S. Cuddapah, *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132: 887–898, 2008.
[49] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, *et al.* Identification and analysis of functional elements in 1 genome by the encode pilot project. *Nature*, 447: 799–816, 2007.
[50] M. Neumann and M. Naumann. Beyond ikappabs: Alternative regulation of nf-kappab activity. *FASEB J.*, 21: 2642–2654, 2007.