Chapter 17
The Role of Transcription Factor Binding Sites in Promoters and Their *In Silico* Detection

Thomas Werner

Abstract As detailed in this chapter TFBSs are among the most important elements of transcription control in promoters, enhancers, locus control regions, and Scaffold/Matrix Attachment regions, to name only the best known. So far, the best way to identify TFBSs in genomic sequences is by sequence similarity searches with whatever method is suitable for the task. As detailed, nucleotide weight matrices are the most popular and developed tools for this purpose. However, all of these methods locate TFBSs one by one and independent of each other, yielding what is called physical binding sites. The only answer such results can provide is the physical binding probabilities of a whole group of different TFs. Neither specific binding of a particular TF can be deduced from such data nor can any functional properties of the stretch of DNA where the TFBS was found be determined. TFBSs do not act as isolated individual binding sites but always as part of a larger context.

Keywords Transcription factors · Frameworks · ModelInspector · Bindingsite · Promoter

17.1 Introduction

Transcriptional regulation is the key link between the static genomic sequence and all the variable events that can be summarized as life. Therefore, regulatory regions in the genome influencing the level of transcription at various levels represent a natural focus for research, linking genomics and transcriptomics or as defined in a previous article "regulomics" [1]. Regulatory regions share several features despite their obvious divergence in sequence and function such as promoters, enhancers, locus control regions, or scaffold/matrix attachment regions. Most of these features are not fixed nucleotide sequences but are variable in sequences restrained by functional requirements. Therefore, understanding of the major components and events during the formation of regulatory DNA-protein complexes is crucial for the design and evaluation of algorithms for the analysis of regulatory regions. One of the most important classes of proteins acting in this context is represented by transcription factors (TFs), many of which interact directly with the genomic DNA.

Algorithms for the analysis and recognition of transcription factor binding sites (TFBSs), necessarily rely on the underlying biological principles in order to generate suitable computational models. Therefore, a brief overview over the biological properties and mechanisms is required to understand why an individual algorithm has been developed in a particular way. The choice of the parameters and implementation of the algorithms largely control the sensitivity and speed of a program. The specificity of software recognizing TBFBs in DNA is determined, to a large extent, by how closely the algorithm follows what will be called the biological model from

T. Werner
Genomatix Software GmbH, Bayerstr. 85A, D-80335, München, Germany
e-mail: werner@genomatix.de

S. Krawetz (ed.), *Bioinformatics for Systems Biology*, DOI 10.1007/978-1-59745-440-7_17,

here on. However, every biological model of individual TFBSs is necessarily incomplete as it became quite evident that a focus on individual TFBSs is too narrow to capture biological function. Extension of the models to consider the relevant context in particular, combinations of several TFBSs, are required to match biology. Nevertheless, even such sophisticated models ultimately have to rely on recognition of the individual TFBSs as a first step.

17.2 Transcription Factors

Transcription factors (TFs) represent a special class of proteins that comes in two flavors: Those TFs that bind directly to DNA and exhibit at least some sequence specificity or better selectivity in their binding and those who form protein-protein networks important in transcription control but do not bind to DNA directly. This latter group are also known as mediators and they will be largely ignored in this chapter due to the fact that they remain invisible to any TFBSs detection. The focus is on the TFs that do bind to DNA directly and more specifically on the corresponding TFBSs of these factors.

17.2.1 Expression, Modifications, and Localization

The extraordinary importance of TFs in the general pool of genes and proteins is clearly visible from the greater variability of expression exhibited by transcription factor genes as compared to other genes. Transcription factor genes produce more alternative transcripts resulting in alternative proteins (62%), than other genes (29%), and also produce more tissue-specific isoforms than the average gene [2]. Often, alternative transcripts of TF genes are also associated with alternative promoters adding another level of complexity to the regulation of TF gene expression. However, the presence and functionality of TFs is controlled on many levels far beyond the initial transcription and translation of all their isoforms.

TFs quite naturally have to enter the nucleus in order to exert any effect on the genomic DNA. Since they are translated, like most other proteins, in the cytoplasm they must cross the nuclear membrane to reach the chromatin, the site of their action. TFs may be regulated in several ways by prevention or facilitation of this cytoplasmic-nuclear transfer, as is the case for NfkB, for example. The classical regulator is IkB, which in itself comes in three isoforms, a, b, c [3], hindering the nuclear transfer of NFkB. It is destructed proteolytically, releasing NFkB triggered by signaling pathways, which is known as the canonical pathway of NFkB activation. However, a variant form of a NFkB precursor protein, P100 can act as a fourth IkB protein and prevent the NFkB complex from nuclear transfer as well. This mechanism of cytoplasmic NFkB retention and its release by other signaling pathways is known as the non-canonical pathway of NFkB activation [4]. Adding the sequestration of NFkB components by other factors, e.g., glucocorticoid receptors illustrates the enormous complexity of even such a simple mechanisms as sequestration. This is by no means restricted to NFkB, as other TFs are also subject to sequestration such as the glucocorticoid receptor, which is guided by heat shock proteins throughout the activation process [5].

17.3 Transcription Factor Binding Sites (TFBSs)

The counterparts of the TFs on the genomic DNA are their binding sites (TFBSs), which attract TFs to the appropriate sites of genomic DNA. Despite the enormous variability of such binding sites and the different selectivity of such TFBSs with respect to the TFs that will bind to them, they do exhibit a few common features, which will be summarized below as physical properties of TFBSs.

17.3.1 Physical Properties

TFBSs generally consist of about 10–30 nucleotides, only a few of which are crucial for specific protein binding. Therefore, individual TFBSs can vary in sequence considerably, even if they bind to the same protein. Nucleotides contacted by the protein in a sequence-specific manner are usually the best-conserved parts of a binding site. Other nucleotides involved in the DNA backbone contacts, i.e., contacting the sugar-phosphate framework of the DNA helix (not sequence specific as they do not involve the bases A, G, C, or T) are much less conserved. The least conserved regions are the internal "spacers" that are not contacted by the protein at all. In general, protein-binding sites exhibit enough sequence conservation to permit the detection of candidates by a variety of sequence similarity-based approaches. However, potential binding sites can be found almost all over the genome and are by no means restricted to (known) regulatory regions. Quite a number of binding sites outside the regulatory regions are also known to bind their respective binding proteins [6], indicating that the abundance of predicted TFBSs is not just a shortcoming of the detection algorithms but at least in some cases reflects biological reality.

Another important feature of TFBSs is their actual lack of specificity. They do bind selectively to TFs but by no means are restricted to a single TF. In many cases, several distinct TFs can bind to the same DNA sequence complicating the identification of the actually binding protein (e.g., STAT1 and STAT3 competing in the IL10 promoter for the same binding site [7]).

It is important to keep in mind that binding of the TF to its cognate binding site is only half of the story. Timely dissociation is at least as important, as transcriptional complexes must disassemble in time to allow for the release of the DNA at the end of the action. As a consequence, nature has not only optimized TFBSs for binding of the TFs but also for releasing them, which is probably one of the reasons why we always have to deal with suboptimal binding sequences in all analyses of real DNA sequences. This has some very important consequences for the development and performance of TFBSs detection algorithms.

17.3.2 Functional Properties–Functional Context

Often it is not possible to identify individual binding proteins as they might bind as part of multi-protein complexes [8]. This illustrates another important point already raised in the introduction: TF-binding *in vivo* is usually context-dependent. The isolated TF will bind to a cognate site quite differently if brought together in a reaction tube as a naked protein and DNA probe (e.g., in a gel shift assay) than *in vivo* where the adaptive DNA structure, chromatin, other TFBSs, and a host of other proteins are around. As it became evident from several chromatin immunoprecipitation (ChIP) studies even *in vivo* binding of a TF does not automatically imply a function in transcription control as was found in a genome-wide study, which identified many more CREB binding sites than CREB regulated genes [9].

17.3.2.1 Epigenetic Context

There is a very simple method to prevent a TF from exerting any effect on a corresponding binding site: Hiding that DNA stretch from the protein by any means will do the trick. There are several mechanisms that will effectively sequester TFBSs efficiently summarized as epigenetic events. The simplest way to prevent many TFs from binding to their cognate sites is by DNA methylation which changes the structure sufficiently to inhibit TF binding [10, 11]. Another mechanism is to inhibit protein access to the DNA by packing the DNA more tightly as is achieved by deacetylation of histones resulting in a denser chromatin structure [12]. All of these modifications cannot be read from the DNA sequence directly and are thus necessarily ignored in any sequence-based analysis

method such as computational TFBSs detection, which is another reason why potential and real binding sites do not necessarily correspond.

17.3.2.2 Promoters

The context-dependency of TFBSs can be best illustrated by the example of eukaryotic polymerase II (pol II) promoters. The TFBSs within the promoters (and most likely in other regulatory sequences as well) do not show any general patterns with respect to location and orientation within the promoter sequences although particular functionality may be associated with a specific location or association within the promoter [13]. However, even functionally important binding sites for a specific transcription factor may occur almost anywhere within a promoter if a large number of promoters are analyzed statistically. However, different locations of TFBSs in individual promoters very often are correlated with specific and distinct functions of such TFBSs. For example, functional AP-1 (Activating protein 1, a complex of two TFs, usually one from the fos and one from the Jun family) binding sites can be located far upstream, as in the rat bone sialoprotein gene where an AP-1 site located about 900 nucleotides upstream of the transcription start site (TSS) inhibits expression [14]. An AP-1 site located close to the TSS is important for the expression of Moloney Murine Leukemia Virus [15]. Moreover, functional AP-1 sites have also been found inside exon 1 (downstream of the TSS) of the proopiomelanocortin gene [16] as well as within the first intron of the fra-1 gene [17], both located outside the promoter. AP1 is only one example, the principles outlined here also apply to other TFBSs, illustrating why the overall statistical correlation of TFBSs within promoters is not meaningful with respect to the biological function of the TFBSs. The context of a TFBS is one of the major determinants of its role in transcription control and the context in this case has to be defined functionally. Physical vicinity of two TFBSs may or may not be functionally relevant; the particular interactions of the binding TFs determines which TFBSs are just nearby or essential for transcriptional function. I will refer to a set of functionally interacting TFBSs in any regulatory region (not restricted to promoters) as *transcriptional modules*. The same TFBSs can be part of distinct transcriptional modules depending on the condition or cell type as has been shown experimentally numerous times (e.g., [18]). A subset of overlapping transcriptional modules from the RANTES promoter is shown in Fig. 17.1. The original definition of transcriptional modules by Arnone and Davidson was more general than the one used here [19].

17.3.2.3 FrameWork Concept

The context of a TF-site is one of the major determinants of its role in transcription control [18]. As a consequence of context requirements, TF sites are usually grouped together in functional groups (frameworks), which have been described in many cases and conveying a specific promoter function will require more than one site (e.g., [20]). When the mutual dependency of TFBSs

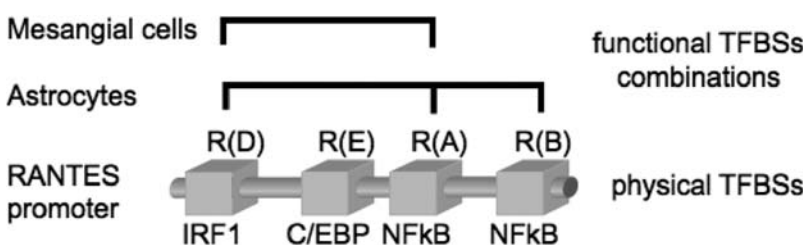

Fig. 17.1 Proximal transcriptional modules in the human RANTES promoter R(A) – R(D) identify experimentally verified promoter regions in the RANTES promoter [18]. Below the boxes, the identifiers of the corresponding TFBSs are given (Copies of figures including color copies, where applicable, are available in the accompanying CD)

in a framework has been experimentally verified, such frameworks are also called transcriptional modules (Fig. 17.1). The organization of the binding sites (and probably also of other elements) of a transcriptional module appears to be much more restricted than what the apparent variety of TFBSs and their distribution in the whole promoter suggests. Within a transcriptional module, both sequential order and distance can be crucial for function indicating that these modules may be the critical determinants of a regulatory region rather than the individual binding sites. Promoter modules are always constituted of more than one binding site. Since regulatory regions, such as promoters, can contain several modules that may use overlapping sets of binding sites, the conserved context of a particular binding site cannot be determined from the primary sequence (Fig. 17.1). This is also the reason why analysis of a DNA sequence solely for individual TFBSs will miss the functional context and thus can only serve as the first step in elucidating transcriptional functions in a DNA sequence. The corresponding modules must be either known *a priori*, determined by comparative sequence analysis or experimentally in suitable expression assays.

17.4 How Transcription Factors Bind to DNA

Transcription factors bind to DNA via a multitude of atomic interactions that are either van der Waals hydrophobic contacts or supported by juxtaposition of oppositely charged amino acids and DNA components. Generally, two basic modes of molecular interactions can be distinguished:

1. The first involves nonspecific contacts between the protein side chains and the so-called backbone of the DNA, which consists of the sugar-phosphate structure linking the bases together (Fig. 17.2). Such contacts can form anywhere on a DNA (double)strand and is responsible for the general tendency of TFs to associate with the DNA. No sequence specific effects are involved in this interaction.
2. The second mode is the sequence-specific recognition achieved by direct contact of amino acid side chains with particular bases of the DNA. Therefore, these contacts can only be formed where there is a suitable succession of bases, i.e., a specific nucleotide sequence (Fig. 17.2). However, as the protein mainly recognizes the DNA structure this allows for some sequence variation as long as the binding structure is maintained.

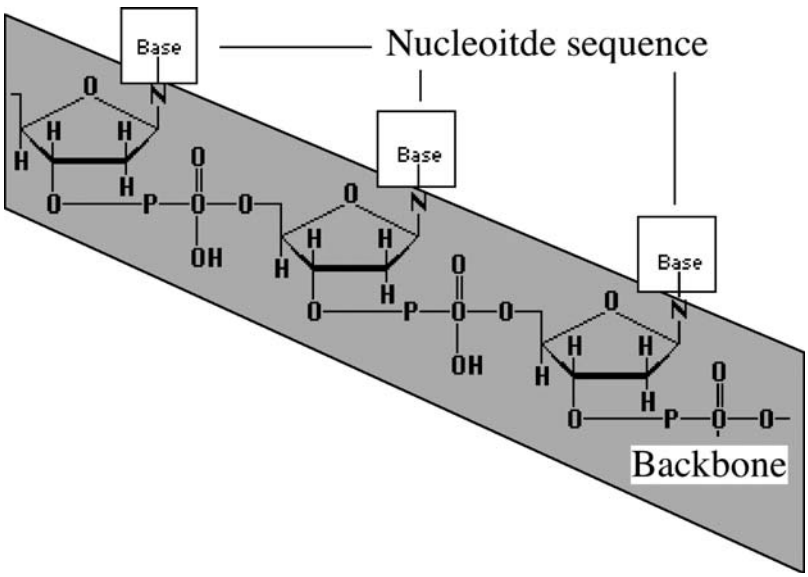

Fig. 17.2 Basic structure of a single DNA strand. The gray area represents the sugar-phosphate backbone involved in the nonspecific protein contacts. The white boxes represent the bases, which are responsible for the sequence specific contacts with proteins (Copies of figures including color copies, where applicable, are available in the accompanying CD)

17.5 Transcription Factor Binding Site Detection *In Silico*

TFBSs need to be differentiated from all the other possible sequences by any detection algorithm. Algorithms used to analyze and detect TFBSs are necessarily based on some kind of a usually simplified model of what a particular TFBS should look like. All models used are inevitably compromising between accuracy with respect to the biological model (the standard of truth) and the computational feasibility of the model. For example, a computational model based on *a priori* three-dimensional structure prediction derived from molecular dynamics using sophisticated force fields may be the most accurate model for a TFBS but cannot be used for the analysis of real data due to excessive demand on computational resources as well as the limited knowledge about structure-sequence relations. Nevertheless, it would represent the ideal biological model, as proteins can only interact with structures and not with letters, we call sequences.

17.5.1 Models of TFBSs

Over time, various different models have been introduced to describe TFBSs. There are several approaches to take the inherent variability of TFBSs into account. The basic models will be discussed in the order of rising complexity.

17.5.1.1 Direct Sequences

The easiest way of course is to put all real sequences of TFBSs into a database and then locate the exact matches only. Figure 17.3 shows the binding sites for glucocorticoid receptor – binding sites as collected in MatBase (Genomatix Software GmbH, Munich) as an example. The biggest disadvantage of such an approach is that as the level of abstraction of this "model" is zero, sequences are taken as they are, and there is no way of inferring onto anything similar. While

Name	Alignment	Matrix similarity	References
remmo1	TGAGCTCTTAGT**GTTC**TAT	(0.906)	1995608
ocotglo5	GCGTTCCAAGCT**GTTC**TCC	(0.905)	3453115
HUMMET2A	CCGGTACACTGT**GTCC**TCC	(0.932)	2881624
MMTV2	TTGGTATCAAAT**GTTC**TGA	(0.932)	2881624
HIVBRUCG_2	GGCTAACTATAT**GTCC**TAA	(0.867)	7684876
RATANFA	CTGCCTGTTTGT**GTTC**TGA	(0.843)	1835978
ratatrc	CAGGACTTGTTT**GTTC**TAG	(0.893)	2881624
RATTATGRE3	GGGGTACAGGTT**GTTC**TGA	(0.974)	2881624
RATTOG5A	TATGCACAGCGA**GTTC**TAG	(0.864)	2881624
ratTAT	GCTGTACAGGAT**GTTC**TAG	(0.962)	2881624
RATTOG5B	CCCTTTCATGAT**GTCC**TGG	(0.881)	2881624
HUMGH1	TGGGCACAATGT**GTCC**TGA	(0.943)	2881624
rnafpg	GAAGTGGTCTTT**GTCC**TTG	(0.861)	2454390
MSV2	GCTGTTCCATCT**GTTC**TTG	(0.952)	2881624
MSV1	TGGGGACCATCT**GTTC**TTG	(0.953)	2881624
gglyshre	CCAGTTTGTACA**GTTC**TGG	(0.839)	3416833
MMTV1	TGGTTACAAACT**GTTC**TTA	(0.943)	2881624
HUMBGPG	AGGGTATAAACA**GTGC**TGG	(0.841)	2038339

Fig. 17.3 Binding sites for glucocorticoid receptors collected from various genes and mammalian organisms. Name: gene symbol, Alignment: actual sequence, Matrix Similarity: MatInspector score, Reference: PubMedID (Copies of figures including color copies, where applicable, are available in the accompanying CD)

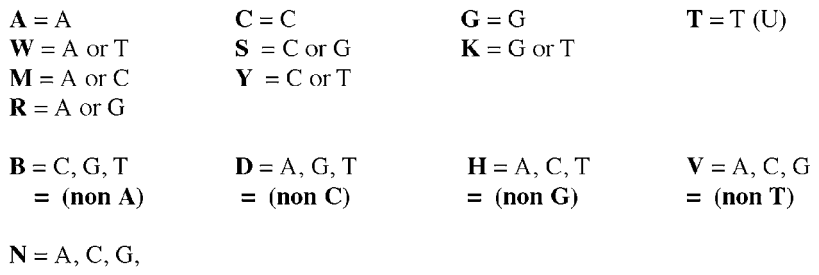

Fig. 17.4 The 15 letter IUPAC ambiguity code (Copies of figures including color copies, where applicable, are available in the accompanying CD)

this will yield only proven binding sites, this approach will miss all variant binding sites and becomes unfeasible if huge numbers of TFBS sequences are to be considered. Naturally, direct sequence matching is not used in any computational approach.

However, those collections of real binding sites are a prerequisite for all modeling approaches and thus databases containing collected TFBSs are valuable resources providing the training sequences for all algorithms used so far.

17.5.1.2 IUPAC Consensi

The most rudimentary improvement to matching direct sequences is to allow for mismatches but this allows indiscriminative matching of all sequences differing by the number of allowed mismatches, taking none of the TF-specific restrictions into account. The simplest model introducing such restrictions is based on simple sequence similarities detected by a IUPAC consensus (using additional letters to describe ambiguities such as R = A or G). Such IUPAC-sequences can be easily used on a computer. Figure 17.4 shows the complete IUPAC ambiguity code.

The IUPAC-sequences can deal with variant sequences but a major drawback of the method is that the definition of the consensus sequence is highly arbitrary depending on which rules are being used to determine the "prevalent" nucleotides as well as the number of sequences considered (e.g., [21]). IUPAC consensi are also very promiscuous in allowing patterns that do not occur anywhere in reality by odd combinations of substitutions. They treat all positions as equally important, which is in stark contrast to the biological reality (base-specific, backbone-, non-binding), i.e., a mismatch at a specific contact site is scored the same as a mismatch in a spacer region. Below is the IUPAC representation of the GRE sequence alignment from Fig. 17.3:

GRE-IUPAC: N N G G T W C W N N N T G T T C T N R

It is easy to see that a lot of sequences not matching any of the founding TFBS sequences will be accepted as well and no particular scoring is possible.

The next model in terms of complexity would be the positional weight matrices. They have been proven to be the most widely used models as of today and will be discussed after the more complex Hidden Markov Models.

17.5.1.3 Hidden Markov Models (HMM)

This is a sophisticated statistical model that has been successfully applied to describe protein-sequence alignments [22]. In brief, HMMs are fully probabilistic models, allowing manipulation and optimizing parameters using Bayesian probability theory. To explain this in simpler terms, consider flipping coins where there are just two possibilities for the outcome. Either the digit is visible or the picture (an eagle if you throw US quarters) is visible. You may attempt to find out what is the probability of either side to show up next just given the outcome of the last event. This would constitute a simple Markov chain. Now consider, there is somebody behind a curtain (so you cannot observe the actual action) throwing coins and telling you the outcome. Again, you attempt

to predict the outcome from the previous result. However, this time things are complicated by the fact that the person behind the curtain has several coins to choose from and would not tell you which one was thrown. So, what you are told is the final result while you do not know the start condition (i.e., which coin was thrown). The selection of the coins is supposed to be a stochastic background process, which cannot be observed directly (hidden). Calculating probabilities for outcome scenarios from the previous outcome involving such a hidden background process constitutes a Hidden Markov Model (HMM).

There are strong as well as the weak points with respect to TFBSs descriptions. In order to train a probabilistic model sufficiently well, a huge training set of instances (here sequences) is required. In terms of binding sites this is about 1,000 or more sequences per model, i.e., binding sites, while realistic numbers for verified sequences is typically in the range of 5–50. Although HMMs represent by far the best mathematical model to describe the full range of parameters influencing the TFBSs (Fig. 17.5), they require far more sequences than those available at present. For this sole reason, HMMs are still not the method of choice for TFBS descriptions. Another problem is that even sophisticated HMMs still do not account for mutual dependencies of nucleotide positions within the TFBSs, which are known to occur in a number of real TFBSs.

New high-throughput methods such as ChIP-on-chip provide us with more and more experimentally verified TFBS-sequences, so one could expect that HMMs may be well- posed to replace the weight matrix based methods one day. However, simultaneously we are learning that subtle differences in such large collections of TFBSs may actually have important functional (i.e., context-dependent) consequences, arguing against throwing them together for HMM training. It is too early to come up with a definitive answer as to whether the final nod goes to HMMs or weight matrices, or something else.

17.5.1.4 Positional Weight Matrices (PMW)

This is a simple and robust model that has been around for more than a decade and is still the prevalent method in every day application [23–26]. In brief, the method starts with an alignment of pre selected sequences (TFBSs for a given factor) and derives a nucleotide distribution matrix from that alignment (counting the number of occurrences for each nucleotide at each position of the alignment). Then weighting is applied, which depends on the particular algorithm used. The resulting positional weight matrix (PWM) is then used to score sample sequences for how well they fit into the matrix. Figure 17.6 shows the nucleotide distribution matrix for the GRE alignment shown in Fig. 17.3.

At the bottom of Fig. 17.6 the Consensus Index (Ci) is shown, which is used to weight the individual positions in the scoring of any sequence. This weighting is based on the concept of

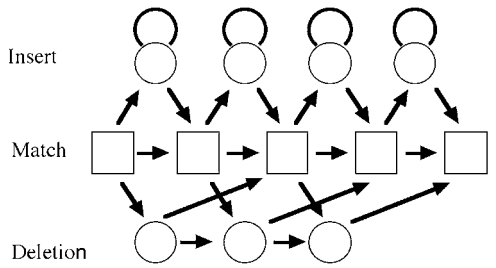

Fig. 17.5 Schematic architecture of a HMM describing the alignment of sequences with a length of five nucleotides. The square boxes indicate matching positions. The circles above the boxes indicate insertions and the circles below indicate deletions. Note that each sequence can feed through a different path in this HMM. The bold arrows represent the transitions and carry the parameters determining the relative probabilities for these transitions (Copies of figures including color copies, where applicable, are available in the accompanying CD)

Pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	1	3	3	0	2	10	0	9	8	5	4	3	0	0	0	0	0	5	6
C	5	6	2	1	4	2	12	3	2	1	6	0	0	0	5	18	0	2	2
G	6	7	10	13	1	1	2	1	4	4	5	0	18	0	1	0	0	7	9
T	6	2	3	4	11	5	4	5	4	8	3	15	0	18	12	0	18	4	1
IUPAC	N	N	G	G	T	W	C	W	N	N	N	T	G	T	T	C	T	N	R
Ci	22.4	20.7	27.4	54.7	35.4	32.5	47.3	27.8	20.9	24.8	15.8	72.0	100.0	100.0	51.1	100.0	100.0	19.1	30.6

Fig. 17.6 Nucleotide distribution at the 19 positions of the GRE alignment from Fig. 17.3. IUPAC: IUPAC consensus sequence, Ci: Consensus Index as calculated by the MatInspector algorithm (Copies of figures including color copies, where applicable, are available in the accompanying CD)

mutual information content (Shannon entropy), which reflects the varying importance of the different positions of a TFBS due to different contact restrictions as discussed. The algorithm was originally described by [26] and further improvements and extensions are described by [23].

The graphical representation of the Ci in comparison with biological evidence from Fig. 17.6 is shown in Fig. 17.7, illustrating the enhanced sensitivity of the weighted matrix towards biological features (sequence specific binding, backbone contacts and spacer region).

So, how can all of that be dealt with in a practical application? There are many different approaches to detect TFBSs and the reader is directed to a recent review for an overview [27]. This chapter will use the MatInspector approach for illustration as this is the only program directly connected to further analyses that take functional context into account and thus at least allows one to go beyond the mere physical results.

17.5.2 TFBS Variability and Multiple TF Binding

All search programs attempting to locate potential TF binding sites in the genomic DNA basically face similar challenges. The most notorious one is incomplete data, preventing the generation of high

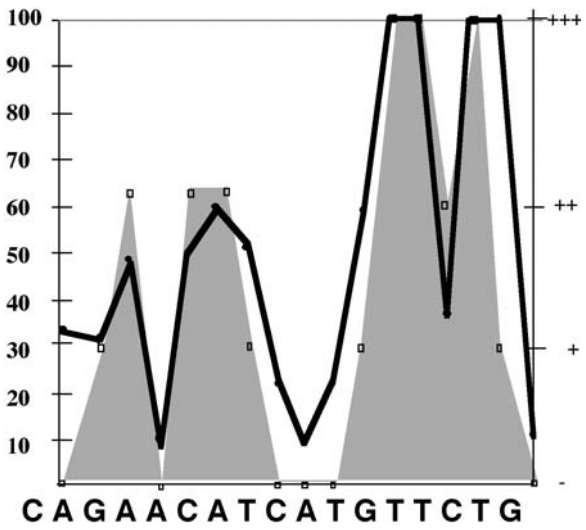

Fig. 17.7 Comparison of calculated importance (consensus index based on Shannon entropy) and experimental evidence for the Glucocorticoid Receptor Binding site. The scale on the left shows the normalized consensus index [28] derived from the nucleotide weight matrix of the GRE [26]. The experimental evidence is given as follows: – = no DNA contact by the protein, + = backbone contact, + + = unspecific base contact, + + + = sequence specific base contact (Copies of figures including color copies, where applicable, are available in the accompanying CD)

quality descriptions by any means as discussed in the context of the HMMs. If only two examples for a specific binding site are available, it is meaningless (but not impossible) to construct a weight matrix and even a simpler IUPAC string will not be very useful. The situation is very close to that of a non-existent description, which is the next most frequent problem. Just because there is no IUPAC or matrix available for a binding site does not mean there is no specific binding site for that protein.

Now, let us consider the case where there is a sufficiently well defined weight matrix and this is used to scan sequences. There are two additional features of TFBSs to be taken into consideration. The first is the inherent suboptimal binding affinity (due to dissociation requirements), which renders strict scoring optimization unsuitable for obtaining biologically meaningful results. The second problem is that very often, similar sequences are known as binding sites for different but related TFs. As already discussed, such TFs (e.g., within the STAT or NFkB families) can actually bind to their mutual binding sites quite well. One solution to take these biological facts into account is the concept of matrix families uniting closely similar TFBS matrices into groups that are then used to score DNA sequences, as has been successfully implemented in the MatInspector algorithm based on grouping of TFBS matrices by self organizing maps [23].

17.5.3 Threshold Issues

Regardless of the model used to describe TFBSs there is one common factor to all of them– the use of one or more thresholds to determine whether a given sequence should be considered a match or not. Choosing the right threshold for matrix detection is always a choice between Scylla and Carybdis. A very sensitive approach will minimize the amount of false negative TFBS predictions and thus is oriented towards a complete annotation. However, this inevitably requires accepting large numbers of additional TFBS hits which might be false positives and easily outnumber the true positive predictions (i.e., verified ones) by an order of magnitude. Just to complicate things even more, all the TFBS methods are designed to detect physical TFBS, i.e., TFBS that will be able to bind to the respective TF, e.g., in a gel shift assay. However, most of the time researchers are not interested in such physical binding sites, but want to know about binding sites that are functional with respect to the transcriptional control of a particular gene. Due to the context-dependency of the TFBS functions, a rather weak match may be the biologically functional one, while a very strong match elsewhere is void of biological function in a particular context.

For these reasons every threshold used in finding TFBS matches in sequences is necessarily a compromise between sensitivity, selectivity, likelihood of TF binding and some kind of 'black box' called biological functionality. What became more than clear over the past decade is that there is definitely no single threshold of the one-size-fits-all type. Matrices are quite different in length and relative sequence conservation, which mandates a more differentiated treatment of thresholds. This is the reason why MatInspector uses individually optimized thresholds for all matrices.

Again, the dilemma between finding TFBS individually (the physical TFBSs) and functional requirements (such as in frameworks and transcriptional modules) has an influence on threshold selection. While a relatively low threshold does not make any sense in scanning large sequences, it is perfectly fit – even required – in a close functional context such as a transcriptional module. The reason is that cooperative binding often substitutes for the requirement of a strong TFBS, as a weaker TFBS might be sufficient in connection with the protein-protein interaction between the two binding TFs. However, such information cannot be retrieved or used by any means if the method focuses on the detection of single TFBSs. Therefore, TFBS-model thresholds remain a thorny issue and whatever choice is made it remains suboptimal in other circumstances.

Just to complicate things a bit further it is also becoming increasingly difficult to determine true positive and true negative matches of TFBS models due to the context dependency of functional

"truth". A TFBS void of function in one tissue may very well be functionally required within the same promoter in another context (Fig. 17.1).

17.6 How to Define Unknown Transcription Factor Binding Sites?

Especially in the light of an ever-growing number of full-scale genomic sequencing projects, it is very important to go after new, unknown binding sites. However, since the generation of a weight matrix requires a set of known binding sites, how is this accomplished for unknown binding sites?

There are several ways to get out of this dilemma, at least partially. So far, there is no way to go for (TFBSs) from a large totally anonymous sequence. Some previous knowledge is required in order to use pattern definition algorithms to produce new patterns that can be turned into IUPAC consensus sequences or nucleotide weight matrices. A very effective and relatively simple approach is the experimental determination of the binding site spectrum for a given protein. This is called SELEX, and is actually an *in vitro* selection of binding sequences from a large collection of random oligonucleotides. Only the sequences with sufficient affinity to the protein will be bound, the rest can be washed away and then sequencing of the bound oligonucleotides reveals individual binding sequences that can be used to derive a nucleotide weight matrix. The catch is that before that can be done an extensive purification of the protein has to be carried out.

Another experimentally oriented approach is the evaluation of expression array data. Here, large amounts of gene probes are arrayed onto a filter or glass chip. This array is then hybridized to the RNA (via complementary DNA) isolated from cells that underwent some treatment. The amount of signal over each spot indicates the approximate level of RNA present for this gene. By comparing such values with other experiments with untreated cells, it is possible to detect which genes changed their RNA levels under treatment. Then it is possible to cluster the genes according to their expression patterns. Analyzing the promoter regions from genes with a very similar expression pattern and at least within a common biologically functional context can be used to identify common patterns from scratch, many of which are transcription factor binding sites. Brazma et al. have demonstrated this approach successfully in the yeast system [29]. Fortunately, promoters in higher eukaryotes, like mammals, are now as readily available as in yeast, due to improved promoter prediction (e.g., [30]) as well as large-scale experimental efforts to locate TSS [31].

Regardless of the way the promoter sequences are acquired, the sequences need to be analyzed for unknown motifs hidden in the set. There are many programs available to go after pattern definition in a set of sequences. The most popular methods are the Gibb's sampler [32], expectation maximization algorithms (e.g., [33], and a variety of other approaches (e.g., [34]).

Discussing all of these algorithms is beyond the scope of this article but all of these methods have specific strong and weak points. There is no single program that will do in all the cases, so a considerable effort in trying out different methods with one set of sequences is still mandatory. The reader is again referred to the excellent review of Tompa et al. [27].

17.7 How to Proceed from Physical TFBSs to Functional Context?

It is quite clear that a single sequence will yield few clues to what the functional context might be, especially as the same sequence may contain organizational structures that are linked to many different functional scenarios. Comparative sequence analysis is one of the most powerful methods to deduce regulatory features and organization, because a selection of sequences to be compared can be based on functional similarities of the sequences (e.g., same pathway or expression behavior). Basically, two types of comparative analysis can be distinguished. The first approach

compares the regulatory regions, e.g., promoters within one species such as promoters co-expressed under particular conditions (e.g., from microarray studies). The second approach compares only orthologous regulatory sequences across several species (again promoters are the most prominent representatives) in order to elucidate which features and elements remained conserved in evolution. Such features should be closely associated with functional conservation of the corresponding regulatory regions. While comparative analysis within species not necessarily allows distinguishing between pure statistical findings and functional conservation, phylogenetic analysis of orthologous regulatory sequences should indicate the predominantly functionally conserved features. However, intra-genomic comparison may differentiate between individual functions especially when based on selection methods such as microarrays [35], while phylogenetic analysis will always yield a summary over all the conserved functions. Thus, very often a combination of both approaches is the best way to go [36, 37].

Most of this approach has been implemented into the program package GEMS Launcher® (Genomatix Software GmbH), which contains MatInspector®, to locate individual TFBSs and the program FrameWorker to carry out automatic comparative sequence analysis for TFBS-frameworks as well as the program ModelInspector® to locate other regulatory sequences in whole genomes containing the same organizational structures as described in a recent publication [23]. This way, the crucial step from physical to functional sequence analysis becomes possible and the principle has been applied successfully in numerous projects and publications (e.g., [36, 38, 20, 35, 7]).

17.8 Summary

Therefore, elucidation of the role of TFBSs in promoters requires taking the functional context (frameworks) into account, which none of the individual detection programs is capable of. Therefore, in the light of all accumulated knowledge about the mechanisms of transcriptional regulation, the title of the chapter should actually read: "The role of transcription factor binding sites in promoters and *in silico* detection of their functional context." Still, finding the individual TFBSs is and remains to be an important task, but stopping there is like collecting all words of a book without looking at the sentences. You will have all the data and almost none of the information contained within.

Frameworking as explained in that chapter is not the only way to elucidate functional context of TFBSs but the only way to do so is based on comparative sequence analysis alone. All other approaches require a lot more experimental evidence such as ChIP data or functional assays (deletion or mutation studies, etc.). Therefore, framework analysis appears to be the natural first step, not excluding or replacing any of the other methods. It is also a very efficient way to take advantage of the enormous wealth of information hidden in the vast amount of genomic sequences available today.

Acknowledgments Part of this work was supported by Biochance-PLUS-3 grant 0313724A (Germany).

References

Werner, T. (2004). Proteomics and regulomics: the yin and yang of functional genomics. *Mass Spectrom Rev* **23**(1), 25–33.

Taneri, B., Snyder, B., Novoradovsky, A., and Gaasterland, T. (2004). Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. *Genome Biol* **5**(10), R75.

Scheinman, R. I., Gualberto, A., Jewell, C. M., Cidlowski, J. A., and Baldwin, A. S., Jr. (1995). Characterization of mechanisms involved in transrepression of NF-kappa B by activated glucocorticoid receptors. *Mol Cell Biol* **15**(2), 943–953.

Basak, S., Kim, H., Kearns, J. D., Tergaonkar, V., O'Dea, E., Werner, S. L., Benedict, C. A., Ware, C. F., Ghosh, G., Verma, I. M., and Hoffmann, A. (2007). A fourth IkappaB protein within the NF-kappaB signaling module. *Cell* **128**(2), 369–381.

Pratt, W. B., Morishima, Y., Murphy, M., and Harrell, M. (2006). Chaperoning of glucocorticoid receptors. *Handb Exp Pharmacol*(172), 111–138.

Kodadek, T. (1998). Mechanistic parallels between DNA replication, recombination and transcription. *Trends Biochem Sci* **23**(2), 79–83.

Ziegler-Heitbrock, L., Lotzerich, M., Schaefer, A., Werner, T., Frankenberger, M., and Benkhart, E. (2003). IFN-alpha induces the human IL-10 gene by recruiting both IFN regulatory factor 1 and Stat3. *J Immunol* **171**(1), 285–290.

Panne, D., Maniatis, T., and Harrison, S. C. (2004). Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-beta enhancer. *Embo J* **23**(22), 4384–4393.

Impey, S., McCorkle, S. R., Cha-Molstad, H., Dwyer, J. M., Yochum, G. S., Boss, J. M., McWeeney, S., Dunn, J. J., Mandel, G., and Goodman, R. H. (2004). Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell* **119**(7), 1041–54.

Douet, V., Heller, M. B., and Le Saux, O. (2007). DNA methylation and Sp1 binding determine the tissue-specific transcriptional activity of the mouse Abcc6 promoter. *Biochem Biophys Res Commun* **354**(1), 66–71.

Kim, J., Kollhoff, A., Bergmann, A., and Stubbs, L. (2003). Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3. *Hum Mol Genet* **12**(3), 233–245.

Ling, G., Wei, Y., and Ding, X. (2006). Transcriptional Regulation of Human CYP2A13 Expression in the Respiratory Tract by C/EBP and Epigenetic Modulation. *Mol Pharmacol*.

Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M., and Pontoglio, M. (1997). Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J Mol Biol* **266**(2), 231–245.

Yamauchi, M., Ogata, Y., Kim, R. H., Li, J. J., Freedman, L. P., and Sodek, J. (1996). AP-1 regulation of the rat bone sialoprotein gene transcription is mediated through a TPA response element within a glucocorticoid response unit in the gene promoter. *Matrix Biol* **15**(2), 119–130.

Sap, J., Munoz, A., Schmitt, J., Stunnenberg, H., and Vennstrom, B. (1989). Repression of transcription mediated at a thyroid hormone response element by the v-erb-A oncogene product. *Nature* **340**(6230), 242–244.

Boutillier, A. L., Monnier, D., Lorang, D., Lundblad, J. R., Roberts, J. L., and Loeffler, J. P. (1995). Corticotropin-releasing hormone stimulates proopiomelanocortin transcription by cFos-dependent and -independent pathways: characterization of an AP1 site in exon 1. *Mol Endocrinol* **9**(6), 745–755.

Bergers, G., Graninger, P., Braselmann, S., Wrighton, C., and Busslinger, M. (1995). Transcriptional activation of the fra-1 gene by AP-1 is mediated by regulatory sequences in the first intron. *Mol Cell Biol* **15**(7), 3748–3758.

Fessele, S., Maier, H., Zischek, C., Nelson, P. J., and Werner, T. (2002). Regulatory context is a crucial part of gene function. *Trends Genet* **18**(2), 60–63.

Arnone, M. I., and Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**(10), 1851–1864.

Naschberger, E., Werner, T., Vicente, A. B., Guenzi, E., Topolt, K., Leubert, R., Lubeseder-Martellato, C., Nelson, P. J., and Sturzl, M. (2004). A NF-kappaB motif and ISRE cooperate in the activation of guanylate binding protein-1 expression by inflammatory cytokines in endothelial cells. *Biochem J* **Pt**.

Cavener, D. R. (1987). Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. *Nucleic Acids Res* **15**(4), 1353–1361.

Eddy, S. R. (2004). What is a hidden Markov model? *Nat Biotechnol* **22**(10), 1315–1316.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*. **21**(13), 2933–2942.

Chen, Q. K., Hertz, G. Z., and Stormo, G. D. (1995). MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput Appl Biosci* **11**(5), 563–566.

Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* **31**(13), 3576–3579.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* **23**(23), 4878–4884.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G.,

van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23**(1), 137–144.

Frech, K., Herrmann, G., and Werner, T, (1993). Computer-assisted prediction, classfication, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res* **21**(7), 1655–1664.

Brazma, A., Jonassen, I., Vilo, J., and Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Res* **8**(11), 1202–1215.

Scherf, M., Klingenhoff, A., and Werner, T. (2000). Highly specific localization of promoter regions in large enomic sequences by promoterInspector: a novel context analysis approach. *J Mol Biol* **297**(3), 599–606.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., and Carninci, P. (2006). CAGE: cap analysis of gene expression. *Nat Methods* **3**(3), 211–222.

Neuwald, A. F., Liu, J. S., and Lawrence, C. E. (1995). Gibbs motif sampling: detection of bacterial outer membrance protein repeats. *Protein Sci* **4**(8), 1618–1632.

Cardon, L. R., and Stormo, G. D. (1992). Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J Mol Biol* **223**(1), 159–170.

Wolfertstetter, F., Frech, K., Herrmann, G., and Werner, T. (1996). Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comput Appl Biosci* **12**(1), 71–80.

Seifert, M., Scherf, M., Epple, A., and Werner, T. (2005). Multievidence microarray mining. *Trends Genet* **21**(10), 553–558.

Cohen, C. D., Klingenhoff, A., Boucherot, A., Nitsche, A., Henger, A., Brunner, B., Schmid, H., Merkle, M., Saleem, M. A., Koller, K. P., Werner, T., Grone, H. J., Nelson, P. J., and Kretzler, M. (2006). Comparative promoter analysis allows de novo identification of specialized cell junction-associated proteins. *Proc Natl Acad Sci U S A* **103**(15), 5682–5687.

Doehr, S., Klingenhoff, A., Maier, H., Hrabe de Angelis, M., Werner, T., and Schneider, R. (2005). Linking disease-associated genes to regulatory networks via promoter organization. *Nucleic Acids Res* **33**(3), 864–872.

Masuda, K., Werner, T., Maheshwari, S., Frisch, M., Oh, S., Petrovics, G., May, K., Srikantan, V., Srivastava, S., and Dobi, A. (2005). Androgen Receptor Binding Sites Identified by a GREF_GATA Model. *J Mol Biol* **353**(4), 763–771.