This is an open access article which appeared in a journal published by Elsevier. This article is free for everyone to access, download and read.

Any restrictions on use, including any restrictions on further reproduction and distribution, selling or licensing copies, or posting to personal, institutional or third party websites are defined by the user license specified on the article.

For more information regarding Elsevier’s open access licenses please visit:

http://www.elsevier.com/openaccesslicenses
Review

Motif grammar: The basis of the language of gene expression

Gergely Nagy, Laszlo Nagy

Abstract

Collaboration of transcription factors (TFs) and their recognition motifs in DNA is the result of coevolution and forms the basis of gene regulation. However, the way how these short genomic sequences contribute to setting the level of gene products is not understood in sufficient detail. The biological problem to be solved by the cell is complex, because each gene requires a unique regulatory network in each cellular condition using the same genome. Thus far, only some components of these networks have been uncovered. In this review, we compiled the features and principles of the motif grammar, which dictates the characteristics and thus the likelihood of the interactions of the binding TFs and their coregulators. We present how sequence features provide specificity using, as examples, two major TF superfamilies, the bZIP proteins and nuclear receptors. We also discuss the phenomenon of "weak" (low affinity) binding sites, which appear to be components of several important genomic regulatory regions, but paradoxically are barely detectable by the currently used approaches. Assembling the complete set of regulatory regions composed of both weak and strong binding sites will allow one to get more comprehensive lists of factors playing roles in gene regulation, thus making possible the deeper understanding of regulatory networks.
1. Introduction

DNA binding by transcription factors (TFs) takes place at recognition sites that meet specific sequence requirements (motifs) throughout the genome. However, there is an incredible redundancy of these sites for a number of reasons. DNA sequence motifs generally contain both strictly invariant and degenerate nucleotides. The latter means that in certain positions two, three, or even four different nucleotides can be tolerated by the binding TF(s), without necessarily changing the binding affinity. As a result, a motif with four invariant nucleotides and two interchangeable ones (\([1/4]^4 \times [2/4]^2\)) or three invariant and four interchangeable ones (\([1/4]^3 \times [2/4]^4\)) can be found in each kilobase of a genome by chance. In a typical mammalian genome, this can result in millions of putative binding sites per TF. However, most of these are hidden in the chromatin structure, so their accessibility is key to establish specific DNA-protein interactions. Therefore, out of the millions of putative binding sites, only a fraction (hundreds to tens of thousands) is indeed available and occupied in the individual cells at a certain moment and condition, and even fewer have direct functional consequences. The motifs characterizing these binding sites evolved along with the TFs, especially with their DNA binding domains (DBDs) to provide specific contacts during the course of phylogeny and ontogeny. Accordingly, members of most major TF (super)families, such as the basic leucine zipper (bZIP), homeodomain, and high mobility group (HMG) proteins, nuclear receptors (NRs), as well as the TATA-binding protein (TBP) and CCAAT-binding complex (CBC), can be found in all metazoan, and except for NRs, also in lower eukaryotes [1–6].

2. Interaction between DNA and transcription factors

Beyond the sequence-specific DNA binding of the major groove, multiple factors contribute to DNA-TF interactions, including the binding of the sugar-phosphate backbone or the base pairs from the side of the minor groove, which latter allows an additional, basically binary sequence readout [7]. Interferon regulatory factors (IRFs) and certain NRs show dual sequence recognition, as these bind in both the major and minor grooves at the same time [8,9]. TBP, CBC, and HMG protein family members, in contrast, bind primarily in the minor groove; however, their recognition sequences allow the specific bending of DNA, which is critical for the interaction and downstream functions [5,6,10]. Like in these cases, the composition of motifs determines the possible local conformations of DNA, which can fit the interaction surfaces of DNA-binding proteins. These so-called shape motifs, which can be achieved even by diverse sequences, imply an additional layer of specificity for sequence motifs [7,11–14]. Similarly, DNA methylation also affects DNA-protein contacts. The binding of methylated GC-rich sequences by TFs is generally greatly limited [15]. For instance, interaction between CCCTC-binding factor (CTCF) and the insulator elements is hindered by DNA methylation [16,17]. In contrast, the repressor Kaiso shows DNA methylation-dependent binding [18,19].

3. The hierarchy of binding sites: From monomer binding sites to enhancers

TFs work in collaboration with other TFs and non-DNA-binding coregulators, forming multi-protein complexes and establishing the connection between promoters and enhancers/silencers to regulate gene expression. Promoters, by definition, contain motifs that facilitate the recruitment of general TFs and RNA polymerase, resulting in basal gene expression levels. Enhancers, in turn, contain cell type-specific motifs, which recruit TFs responsible for the induction of phenotype-determining genes from a distance via looping. Silencers have the opposite effect to enhancers, although this is due to the binding TFs, which can take part in gene regulation as activators, repressors, or collaborators. As a result, the same sequence can behave either as an enhancer or silencer depending on the cellular context [20]. For simplicity, we will refer to all these, typically promoter-distal regulatory regions as enhancers.

Genes are regulated by various numbers of TF binding sites. Conserved gene regulatory regions, such as promoters and certain enhancers, can be hundreds of nucleotides long, including dozens of binding sites, while most enhancers have evolved recently and consist of a few or even one single binding site [21]. Several TFs are functional as monomers and interact with single monomer binding sites with high affinity, although these should be long enough – usually at least six invariable nucleotides – to provide specific surface for the required amount of molecular contacts. Most TFs bind DNA as dimers, and these can be further assembled to larger complexes capable of DNA binding at multiple surfaces. In line with these, monomer and dimer binding sites can be arranged into more complex elements, and ultimately, add up to collaborating promoters and enhancers. Within these functional sequences, monomer binding sites can be located in several ways relative to each other; however, their distribution is not fully random – there are regularities within the line of elements to make possible interactions with the TFs proper for cell type-specific gene regulation. Moreover, the joint recruitment of multiple TFs can make suboptimal binding sites accessible, thus enhancers (and promoters) can contain weak binding sites [22]. For instance, several developmental genes are regulated by both optimal and suboptimal binding sites with an optimized relative distribution, while experimental improvement of weak binding sites cause ectopic gene expression [23–26].

4. Transcription factors and composite elements

Most TFs form dimers, in which both proteins are capable of DNA binding at dimer binding sites, thus creating the possibility or further increasing the specificity of DNA-protein interactions...
(Fig. 1). Dimerization can take place with the involvement of two identical TFs (homodimers), TFs from the same (super)family, or even from families of different origin (heterodimers) [27–29]. For instance, bZIP proteins have an integrated DNA-binding and dimerization domain, which enables a flexible choice of dimerizing partners within the superfamily and specific motif recognition depending on the partners [1,30–32]. Steroid hormone receptors and dimeric orphan receptors (NR superfamily) form homodimers or heterodimers with NRs from the same family, while NRs from most NR1 families (NR1A, B, C, H, I) form heterodimers with the retinoid X receptors (RXRs, NR2B) [23,33]. In this superfamily, two zinc fingers provide sequence-specific recognition and also contribute to dimerization.

Although most known dimers are assembled from closely related TFs, currently, there is an emerging number of newly identified heterodimers built up from TFs from different protein (super)families [28–29]. While the former group of dimers binds two associated, substantially identical monomer binding sites – so-called half-sites –, the latter binds composite elements of different monomer binding sites (Fig. 1). The protein product of myeloid ecotropic viral integration site 1 (MEIS1), for example, binds several composite elements in collaboration with members of other homeodomain protein families [27,34]. It also has a homodimer binding site; however, the DBDs within the supposed homodimer are on the opposite side of DNA, so there is no contact between them. In contrast, in the case of MEIS1/distal-less homeobox 3 (DLX3) heterodimers, the DBD of DLX3 distorts DNA, narrows and binds the minor groove, and then it is capable of interacting more closely with the DBD of MEIS1 [27]. Both examples show contacts that are primarily independent of TF-TF interactions and more closely with the DBD of MEIS1 [27].

In line with the symmetric basic structure of bZIP dimers, the direction of bZIP half-sites is always convergent, and the spacer appears to be an integral part of the half-sites (Fig. 2). The most

## 5. Motif grammar

Monomer binding sites within longer gene regulatory units can follow each other in several ways to provide specificity and selectivity. In the simplest case, half-sites can form an asymmetric head-to-tail (tandem or direct repeat, DR) or symmetric (inverted or everted) repeat (IR or ER, respectively); in addition, the distance between them can also vary. For instance, within the signal transducer and activator of transcription (STAT) family all dimers are capable of binding the quasi-palindromic, interferon γ-activated sequence (GAS) with three spacer nucleotides (5′-TTCC(N/A/G)GAA-3′), while STAT6 homodimers prefer four nucleotide-long spacers [42–44]. In the case of composite elements and higher-order regulatory sequences, besides the orientation and spacing, the type, number, and order of monomer binding sites also have significance, not to mention the strength (affinity) of the individual binding sites and the shape information encoded in DNA (Table 1). These sequence features that determine the quality, order, orientation, and putative interactions of the binding TFs and their coregulators are termed motif grammar, or promoter/enhancer grammar, if we consider entire regulatory regions [23,45]. To illustrate how sequence features determine specific DNA-TF interactions, we discuss in the next sections the basic sequence requirements of two major TF groups, the bZIP and NR superfamilies, involved in important physiological and pathological processes [1,2]. Furthermore, we listed in Table 2 all elements of this review, mentioned in connection with motif grammar.

## 6. Motif grammar of bZIP proteins

In line with the symmetric basic structure of bZIP dimers, the direction of bZIP half-sites is always convergent, and the spacer appears to be integral part of the half-sites (Fig. 2). The most

### Table 1

| Motif grammar | Enhancer grammar | Features |
|---------------|------------------|---------|
| Monomer binding site | Dimer binding site | Composite element | Cluster of elements | Gene regulation |
| 7 | 7 | 7 | + | 6– hundreds |
| 3–15 | 6–20 | 10–25 | + | 1– tens |
| 1 | 2 | 2–3 | + | + |
| – | –/+/+ | + | + | Number of binding sites |
| – | –/+/+ | + | + | Type |
| – | + | + | + | Order |
| + | – | + | + | Orientation |
| + | + | + | + | Spacing |
| + | + | + | + | Strength |
| + | + | + | + | Shape |

Sequence features that contribute (+) or do not contribute (−) to motif/enhancer complexity (specificity) are indicated. Homodimer binding sites, for instance, contain basically identical monomer binding sites (half-sites), so their orientation, spacing, strength, and shape can vary (+), but their type and order are self-evident (−), while in other dimer binding sites the type and order can also be determinant features (+). Unlike in the case of enhancers (promoters), the effect of individual elements on gene expression is uncertain (?).
Table 2
Summary of representative transcription factors and their binding sites.

| Name of transcription factors (element) | Citation(s) |
|----------------------------------------|-------------|
| MEIS1/MEIS1                            | Jolma et al. 2015 [27] |
| MEIS1/DLX3                             | Jolma et al. 2015 [27] |
| OCT4/SOX2                              | Rodda et al. 2005 [35] |
| IRF/IRF (ERE)                          | Fuji et al. 1999 [9] |
| PJ3/IRF4/8 (ECE/IRE)                   | Morera et al. 2002 [38] |
| IRF4/8/PU1.1 (IECS)                    | Tamura et al. 2005 [39] |
| GCM1/ELK1                              | Jolma et al. 2015 [27] |
| STAT/STAT (GAS)                        | Pease et al. 1993 [42], Seidel et al. 1995 [43] |
| STAT6/STAT6                            | Li et al. 2016 [44] |
| JUN/FOS (TRE)                          | Deppe mann et al. 2006 [46], Amoutzias et al. 2007 [1], Cohen et al. 2018 [31] |
| CREB, ATF, JUN dimers (CRE)            | Deppe mann et al. 2006 [46], Amoutzias et al. 2007 [1], Cohen et al. 2018 [31] |
| sMAF/CNC (MARE)                        | Inamdar et al. 1996 [47], Newman et al. 2003 [30] |
| IMAF/MAF (MARE)                        | Kataoka et al. 1994 [49], Newman et al. 2003 [30], Kurokawa et al. 2009 [48] |
| C/EBP/C/EBP (ERE)                      | Cohen et al. 2018 [31] |
| C/EBP/ATF4 (ARE)                       | Cohen et al. 2018 [31] |
| JUN/BATF/IRF4/8 (AICE)                 | Glasmacher et al. 2012 [51] |
| NE3C dimers (IR2)                      | Mangelsdorf et al. 1995 [2] |
| ER/ER (ERE)                            | Mangelsdorf et al. 1995 [2] |
| RAR/RXR (RARE, DRE-2, 5.8, IRD)         | Rastinejad et al. 1995 [59], Moutier et al. 2012 [53], Simandi et al. 2018 [52] |
| PPAR/RXR (PPRE)                        | Ijpenberg et al. 1997 [8], Chandra et al. 2008 [55], Nagy et al. 2020 [67] |
| REV-ERB/REV-ERB (Rev-DR2)              | Ijpenberg et al. 1997 [8], Sierk et al. 2001 [54] |
| RXR/LXR (LXRE)                         | Feldmann et al. 2013 [38], Lou et al. 2014 [57] |
| RXR/VDR (VRE)                          | Rastinejad et al. 1995 [59], Orlov et al. 2012 [60] |
| RXR/THR (THRE)                         | Rastinejad et al. 1995 [59], Granette et al. 2015 [61] |
| ROR (RORE)                             | Ijpenberg et al. 1997 [8] |
| NR3B                                   | Johnston et al. 1997 [63] |
| NR4A                                   | Wilson et al. 1992 [64] |
| NR5A                                   | Lala et al. 1992 [63] |

Table 3

| Name of transcription factors (element) | Citation(s) |
|----------------------------------------|-------------|
| OCT4/SOX2                              | Rodda et al. 2005 [35] |
| IRF/IRF (ERE)                          | Fuji et al. 1999 [9] |
| PJ3/IRF4/8 (ECE/IRE)                   | Morera et al. 2002 [38] |
| IRF4/8/PU1.1 (IECS)                    | Tamura et al. 2005 [39] |
| GCM1/ELK1                              | Jolma et al. 2015 [27] |
| STAT/STAT (GAS)                        | Pease et al. 1993 [42], Seidel et al. 1995 [43] |
| STAT6/STAT6                            | Li et al. 2016 [44] |
| JUN/FOS (TRE)                          | Deppe mann et al. 2006 [46], Amoutzias et al. 2007 [1], Cohen et al. 2018 [31] |
| CREB, ATF, JUN dimers (CRE)            | Deppe mann et al. 2006 [46], Amoutzias et al. 2007 [1], Cohen et al. 2018 [31] |
| sMAF/CNC (MARE)                        | Inamdar et al. 1996 [47], Newman et al. 2003 [30] |
| IMAF/MAF (MARE)                        | Kataoka et al. 1994 [49], Newman et al. 2003 [30], Kurokawa et al. 2009 [48] |
| C/EBP/C/EBP (ERE)                      | Cohen et al. 2018 [31] |
| C/EBP/ATF4 (ARE)                       | Cohen et al. 2018 [31] |
| JUN/BATF/IRF4/8 (AICE)                 | Glasmacher et al. 2012 [51] |
| NE3C dimers (IR2)                      | Mangelsdorf et al. 1995 [2] |
| ER/ER (ERE)                            | Mangelsdorf et al. 1995 [2] |
| RAR/RXR (RARE, DRE-2, 5.8, IRD)         | Rastinejad et al. 1995 [59], Moutier et al. 2012 [53], Simandi et al. 2018 [52] |
| PPAR/RXR (PPRE)                        | Ijpenberg et al. 1997 [8], Chandra et al. 2008 [55], Nagy et al. 2020 [67] |
| REV-ERB/REV-ERB (Rev-DR2)              | Ijpenberg et al. 1997 [8], Sierk et al. 2001 [54] |
| RXR/LXR (LXRE)                         | Feldmann et al. 2013 [38], Lou et al. 2014 [57] |
| RXR/VDR (VRE)                          | Rastinejad et al. 1995 [59], Orlov et al. 2012 [60] |
| RXR/THR (THRE)                         | Rastinejad et al. 1995 [59], Granette et al. 2015 [61] |
| ROR (RORE)                             | Ijpenberg et al. 1997 [8] |
| NR3B                                   | Johnston et al. 1997 [63] |
| NR4A                                   | Wilson et al. 1992 [64] |
| NR5A                                   | Lala et al. 1992 [63] |

Fig. 2. Schematic representation of major bZIP motifs. TRE/CRE (5′-TCA(G)-3′) and MARE (5′-TCTGCA(G/C)-3′) half-sites are marked by blue arrows, C/EBP half sites are marked by green arrows, and spacer nucleotides are marked by black dots. Schematic motif logos and motif/protein names are indicated (TRE, TPA response element; CRE, cAMP response element; CREB, CRE binding protein; ATF, activating transcription factor; MAF, musculoaponeurotic fibrosarcoma protein; CNC, cap’ri collar-type bZIP protein; C/EBP, CCAAT/enhancer-binding protein; sMARE, small MAF response element; CARE, C/EBP-ATF response element). For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
ing, which are self-evident features for most dimer binding sites. Basically, there are two types of NR half-sites. Most NRs bind variants of the 5'-AGA/G/GTCA-3' consensus half-site, while members of the NR3C steroid hormone receptor family have a 5'-AGAACA-3' consensus sequence (Fig. 3) [2]. All steroid hormone receptor dimers bind IR elements separated by 3 nucleotides (IR3), and the additional NR dimers recognize a complete series of DR elements with zero to at least five nucleotide long spacers (DR0-5). Retinoic acid receptor (RAR, NR1B)/RXR heterodimers, for instance, bind a wide range of elements, namely DR0, DR1, DR2, DR5, DR8 (DR2:DR0), and also IR0 elements [52–53]. Since most DR elements can be bound by more than one dimer, in their case, specificity should be made possible in other ways. Beyond their cell typespecific gene expression and dimer conformation that allow the recognition of different spacer lengths, NRs also adapted to additional sequence features, primarily the quality of the half-sites and their flanking sequences (Fig. 3).

Certain NRs recognize a half-site longer than the consensus hexamer. All these sequences are extended upstream and some of them can be part of DR elements and thus serve specific motif recognition by NR dimers. Peroxisome proliferator (PP)-activated receptors (PPARs, NR1C), REV-ERB proteins (NR1D), and RAR-related orphan receptors (RORs, NR1F) belong to different NR classes based on their dimerization and DNA binding features, but all of these prefer 5'-(A/G)(A/C/G/T)(A/G)GTCA-3' sequences over shorter half-sites (Fig. 3). These NRs have a carboxy-terminal extension (CTE) of their DBD, and this interacts with the minor groove upstream to the DBD-bound half-site [8,54]. Thereby PPAR/RXR heterodimers bind extended DR1 elements (PPREs) with higher affinity than other DR1s, and the other DR1-binding NRs, such as hepatocyte nuclear factor 4 (HNF4, NR2A) and testicular orphan receptor (TR, NR2C) dimers, might prefer DR1 elements with other features [2,55]. Similarly, REV-ERB dimers prefer DR2 elements with 5'-A(C/G)T-3' extensions both upstream to the DR2 and within the spacer (Rev-DR2s) [54]. As a result, the monomeric RORs are capable of PPRE and Rev-DR2 binding, which both include the ROR response element (RORE), and REV-ERBs are capable of repressing both on single ROREs and PPREs, which has functional consequences in the regulation of the circadian rhythm of the cells [56].

Besides these related NR families, there are additional ones with sequence preference within the minor groove. These all form heterodimers with RXR, bind the downstream half-site, and thus their specific 5' extensions are in the spacer of DR elements (Fig. 3). Interestingly, liver X receptors (LXRs, NR1H2-3) have an extension similar to that of ROREs despite the significant structural differences. The DR4 of RXR/LXR (LXRE) contains a spacer with a preferred 5'-CTTN-3' sequence, which is bound by the aminoterminal extension (NE) of LXR [57,58]. In contrast, in the case of the vitamin D receptor (VDR, NR1I1) and thyroid hormone receptors (THRs, NR1A), the CTE of DBD has an alpha-helical structure that crosses the minor groove in the spacer of DR3 or DR4, respectively [59,60]. The CTE helix of THR interacts with more phosphate groups of both strands and the first two nucleotides of the 3' half-site in the minor groove, yet it has a preferred nucleotide within the spacer, of which general sequence is 5'--NN(T/C)N-3' [61]. VDR has an alternative, 7-nucleotide long 3' half-site with a 5'--G(A/G)(G:T/G)CTCA-3' consensus sequence, of which beginning shows similar interactions with the CTE helix as observed in the case of THR [60,62].

The last class of NRs is the monomeric orphan receptors, which theoretically have a single half-site to bind, but since the variants of this consensus hexamer are very frequent and thus cannot provide specificity, these NRs – including RORs – also have 5' extended monomer binding sites. As a result, NR4A proteins require a 5'-AA-3' extension, and NR3B and NR5A proteins require a 5'-CA-3' extension beyond the NR "half-site" (Fig. 3) [63–65]. This means that except for NR0B proteins, which have no complete DBD allow-

![Fig. 3. Schematic representation of nuclear receptor motifs. The general (5'-A(G/C)GTCA-3') and the NR3C steroid hormone receptor-specific (5'-AGAACA-3') half-sites are marked by blue or green arrows, respectively. 5' extensions are marked by cyan arrows, and spacer nucleotides are marked by black dots. Schematic motif logos and motif protein names are indicated (IR, inverted repeat; DR, direct repeat; ROR, retinoic acid receptor (RAR)-related orphan receptor; PPAR, peroxisome proliferator (PP)-activated receptor; RXR, retinoid X receptor; VDR, vitamin D receptor; THR, thyroid hormone receptor; LXR, liver X receptor; ERE, estrogen response element; RORE, ROR response element; PPRE, PP response element; VDRE, vitamin D response element; THRE, thyroid hormone response element; LXRE, LXR response element). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)](image-url)
ing DNA binding, only the NR2E family members have no extended half-site described, although this is one of the least examined NR family.

Not only the monomeric, but also the dimeric NRs show motif degeneracy, although typically only one half-site is affected, and one strong half-site is required for stable DNA binding, like in the case of the C/EBP dimer binding sites. In the so-called half-site binding mode any half-site can provide strong DNA-protein interaction, and this contributes to the recruitment of the dimerizing partner with the weak half-site [66]. In the case of PPRs, both the PPAR and RXR half-site can be extended, and these provide more frequent binding than the weak and non-extended ones [67]. Importantly, half-site binding mode is dominant over the full site mode in the PPARγ cistrome of macrophages and adipocytes, which can be part of the optimization of enhancers to fulfill their gene regulatory roles but involves major technical limitations in the determination of sequence-specific direct binding events.

8. Summary and outlook

Naturally, our understanding of DNA-TF interactions is limited by the methodologies used. Initially, a few model or canonical sites were used for molecular biology and biochemical studies. These put restriction on and provided bias to building the motif grammar. In the last two decades, several ex vivo high-throughput methods, such as protein-binding microarrays, systematic evolution of ligands by exponential enrichment (SELEX)-based methods, and DNA affinity purification sequencing (DAP-seq) were developed to characterize the sequence features determining DNA binding by TFs. These showed not only the sequences optimal for DNA-TF interactions, but also their DNA methylation dependence, a wide range of possible TF-TF interactions through composite elements, as well as the significance of half-site binding mode [7,15,27,66]. However, due to the fact that these cannot take live cellular conditions – such as the neighboring or farther interacting sequences and the concentration of collaborating or competing TFs and other chromatin components – into consideration, several features, for instance those allowing the binding of suboptimal sites, could not be broadly tested. A successful way to investigate this phenomenon is to compare organisms with different genotypes or generate mutants and show the alterations in DNA binding – although these are not high-throughput approaches [22–25]. In contrast, the availability of cistronic (chromatin immunoprecipitation sequencing, ChIP-seq) data sets allows an unbiased assessment, although it is always a great challenge to discriminate direct and indirect DNA binding events, since bioinformatic approaches typically cannot identify weak binding sites, although these can be as functional as canonical elements in certain conditions. Identifying all functional units of regulatory regions requires more sophisticated, future approaches.

CRediT authorship contribution statement

Gergely Nagy: Conceptualization, Writing - original draft, Visualization. Laszlo Nagy: Writing - review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank the members of the Nagy laboratory for discussions and comments on the manuscript. This work was supported by the National Research, Development and Innovation Office to the Nuclear Receptor Research Laboratory (KKPI29909, K124298) and to G.N. (PD124834). In addition, G.N. is a recipient of the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and supported by the ÚNKP-19-4-DE-173 New National Excellence Program of the Ministry of Human Capacities. L.N. is supported by the National Institutes of Health (RO1DK115924).

References

[1] Amoutzias GD et al. One billion years of bZIP transcription factor evolution: conservation and change in dimerization and DNA-binding site specificity. Mol Biol Evol 2007;24:827–35.
[2] Mangelsdorf DJ et al. The nuclear receptor superfamily: the second decade. Cell 1995;83:835–9.
[3] Kramm K, Engel C, Grohmann D. Transcription initiation factor TBP: old friend, new questions. Biochem Soc Trans 2019;47:411–23.
[4] Bobola N, Merabet S. Homeodomain proteins in action: similar DNA binding preferences, highly variable connectivity. Curr Opin Genet Dev 2017;31:1–8.
[5] Huber EM, Scharf OH, Hortschansky P, Groll M, Brakhage AA. DNA minor groove sensing and widening by the CCAAT-binding complex. Structure 2012;20:1757–68.
[6] Daiyle L, Basilio C. Coevolution of HMG domains and homeodomains and the generation of transcriptional regulation by Sox/POU complexes. J Cell Physiol 2001;186:315–28.
[7] Siggers T, Groden R. Protein-DNA binding: complexities and multi-protein codes. Nucleic Acids Res 2014;42:2099–111.
[8] IJenberg A, Joannou E, Wabili W, Desvergne B. Polarity and specific sequence requirements of peroxisome proliferator-activated receptor (PPAR)/retinoid X receptor heterodimer binding to DNA. A functional analysis of the malic enzyme PPAR response element. J Biol Chem 1997;272:20108–17.
[9] Fuji Y et al. Crystal structure of an IRF-DNA complex reveals novel DNA recognition and cooperative binding to a tandem repeat of core sequences. EMBO J 1999;18:5028–41.
[10] Nikolob DB et al. Crystal structure of a human TATA box-binding protein/TATA element complex. Proc Natl Acad Sci USA 1996;93:4862–7.
[11] Samee MAH, Bruneau BG, Pollard KS. A de novo shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. Cell Syst 2019;8:42–46.e6.
[12] Rube HT, Rastogi C, Kriebelbauer JF, Bussemaker HJ. A unified approach for quantifying and interpreting DNA shape readout by transcription factors. Mol Syst Biol 2018;14:e7902.
[13] Chu T-P, Xin B, Mikkelsen T, Wang Y, Rohe R. TFBShape: an expanded motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res 2019;48:D246–55.
[14] Rohe R et al. The role of DNA shape in protein-DNA recognition. Nature 2009;461:1248.
[15] O’Malley RC et al. Cistrome and epistrome features shape the regulatory DNA landscape. Cell 2016;165:1280.
[16] Hashimoto H et al. Structural basis for the versatile and methylation-dependent binding of CTCF to DNA. Mol Cell 2017;66:711–720.e3.
[17] Bell AC, Felsenfeld G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature 2000;405:482–5.
[18] Buck-Koehntop BA et al. Molecular basis for recognition of methylated and specific DNA sequences by the zinc finger protein Kaiso. Proc Natl Acad Sci USA 2012;109:15229–34.
[19] Prokhortchouk A et al. The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. Genes Dev 2001;15:1613–8.
[20] Gisselbrecht SS et al. Transcriptional silencers in drosophila serve a dual role as transcriptional enhancers in alternate cellular contexts. Mol Cell 2020;77:324–337.e8.
[21] Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nat Ecol Evol 2018;2:152–63.
[22] Tsai A et al. Nuclear microenvironments modulate transcription from low-affinity enhancers. Elife 2017;6.
[23] Farley EK, Olson KM, Zhang W, Rokhsar DS, Levine MS. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proc Natl Acad Sci USA 2016;113:6508–13.
[24] Crocker J et al. Low affinity binding site clusters confer HOX specificity and regulatory robustness. Cell 2015;160:191–203.
[25] Ramos AI, Barolo S. Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. Philos Trans R Soc B Biol Sci 2013;368:20130018.
