Computational and Structural Biotechnology Journal 18 (2020) 2026–2032

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Motif grammar: The basis of the language of gene expression

Gergely Nagy [a], Laszlo Nagy [a,b,*]

[a] Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen, Debrecen, HU 4032, Hungary
[b] Johns Hopkins University School of Medicine, Departments of Medicine and Biological Chemistry, Institute for Fundamental Biomedical Research, Johns Hopkins All Children's Hospital, Saint Petersburg, FL 33701, USA

ARTICLE INFO

Article history:
Received 29 April 2020
Received in revised form 6 July 2020
Accepted 8 July 2020
Available online 18 July 2020

Keywords:
Motif grammar
Transcription factor
Nuclear receptor
Basic leucine zipper
Weak motif

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.

Contents

* Corresponding author at: Johns Hopkins University School of Medicine, Departments of Medicine and Biological Chemistry, Institute for Fundamental Biomedical Research, Johns Hopkins All Children's Hospital, Saint Petersburg, FL 33701, USA.
E-mail address: lnagy@jhmi.edu (L. Nagy).

https://doi.org/10.1016/j.csbj.2020.07.007

1. Introduction

DNA binding by transcription factors (TFs) takes place at recognition sites that meet their 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 invariable 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 invariable nucleotides and two interchangeable ones ($[1/4]^4 \times [2/4]^2$) or three invariable 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 phylogenesis and ontogenesis. 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.

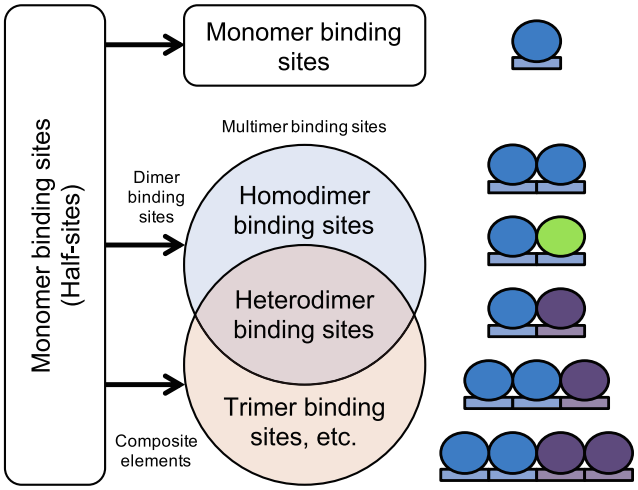

Fig. 1. Classification of transcription factor binding sites. Monomer binding sites, which can be bound by single TFs, often add up to larger units. Two, essentially identical half-sites can be bound by homodimers formed by identical TFs (blue circles, right) or heterodimers formed by related TF partners (blue and green circle). Composite elements are built up form at least two monomer binding sites specific for different TFs (blue and dark purple circles). Boxes colored according to TFs represent monomer binding sites. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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) [7,27–29]. 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) [2,33]. 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 suggest that dimer binding sites can have major roles both in TF binding and dimerization, although other TFs, like bZIP proteins, have interaction surfaces large and compatible enough to provide dimerization even in the absence of specific DNA segments.

There are also long known heterodimers and composite elements with critical developmental and cell lineage-determining roles. Collaboration of octamer-binding transcription factor 4 (OCT4, homeodomain) and sex determining region Y (SRY)-box 2 (SOX2, HMG) on their shared composite elements is a key for the maintenance of embryonic stem cells [6,35]. Purine-rich nucleic acid binding protein 1 (PU.1; erythroblast transformation-specific [ETS] superfamily), in turn, is indispensable for the development and maintenance of most white blood cells and tightly collaborates with IRF4 or IRF8 on several kinds of composite elements [36,37]. Within these elements, the core nucleotide tetramers tandemly follow each other, but their order and spacing are different. The ETS:IRF composite element (EICE) with two spacer nucleotides and ETS:IRF response element (EIRE) with three spacer nucleotides are bound by PU.1 and IRF4/8 in this order, while the IRF:ETS composite sequence (IECS) with two or three spacer nucleotides is bound in the opposite order by the two proteins. These motifs imply different ways of dimerization, although the transcriptional effects of different conformations of the formed ternary complexes are not known [38–41]. There are additional ETS proteins that form heterodimers with TFs of different origins and have composite elements accordingly. Out of these, several elements, such as that of the glial cells missing transcription factor 1 (GCM1)/ETS-like 1 (ELK1) heterodimer, lack flanking (spacer) nucleotides or contain altered ones [27]. Thereby, in certain cases, the knowledge of canonical monomer binding sites is not sufficient to cover all elements in use, but also dimer-specific information should be used to cover all binding sites (the cistrome) of a certain TF.

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′-TTC(T/C)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
Features of motif and enhancer (promoter) grammar.

Motif grammar			Enhancer grammar	Features
Monomer binding site	Dimer binding site	Composite element	Cluster of elements	
?	?	?	+	Gene regulation
3–15	6–20	10–25	6– hundreds	Size (bp)
1	2	2–3	1– tens	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 determinate 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 (ISRE)	Fujii et al. 1999 [9]
PU.1/IRF4/8 (EICE/EIRE)	Meraro et al. 2002 [38]
IRF4/8/PU.1 (IECS)	Tamura et al. 2005 [39]
GCM1/ELK1	Jolma et al. 2015 [27]
STAT/STAT (GAS)	Pearse et al. 1993 [42], Seidel et al. 1995 [43]
STAT6/STAT6	Li et al. 2016 [44]
JUN/FOS (TRE)	Deppmann et al. 2006 [46], Amoutzias et al. 2007 [1], Cohen et al. 2018 [31]
CREB, ATF, JUN dimers (CRE)	Deppmann 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]
lMAF/lMAF (MARE)	Kataoka et al. 1994 [49], Newman et al. 2003 [30], Kurokawa et al. 2009 [48]
C/EBP/C/EBP	Cohen et al. 2018 [31]
C/EBP/ATF4 (CARE)	Cohen et al. 2018 [31]
JUNB/BATF/IRF4/8 (AICE)	Glasmacher et al. 2012 [51]
NR3C dimers (IR3)	Mangelsdorf et al. 1995 [2]
ER/ER (ERE)	Mangelsdorf et al. 1995 [2]
RAR/RXR (RARE, DR0-2,5,8, IR0)	Rastinejad et al. 1995 [59], Moutier et al. 2012 [53], Simandi et al. 2018 [52]
PPAR/RXR (PPRE)	lJpenberg et al. 1997 [8], Chandra et al. 2008 [55], Nagy et al. 2020 [67]
REV-ERB/REV-ERB (Rev-DR2)	lJpenberg et al. 1997 [8], Sierk et al. 2001 [54]
RXR/LXR (LXRE)	Feldmann et al. 2013 [58], Lou et al. 2014 [57]
RXR/VDR (VDRE)	Rastinejad et al. 1995 [59], Orlov et al. 2012 [60]
RXR/THR (THRE)	Rastinejad et al. 1995 [59], Grøntved et al. 2015 [61]
ROR (RORE)	lJpenberg et al. 1997 [8]
NR3B	Johnston et al. 1997 [65]
NR4A	Wilson et al. 1992 [64]
NR5A	Lala et al. 1992 [63]

MEIS1, myeloid ecotropic viral integration site 1; DLX3, distal-less homeobox 3; OCT4, octamer-binding transcription factor 4; SOX2, sex determining region Y (SRY)-box 2; IRF, interferon regulatory factor; ISRE, interferon-stimulated response element; PU.1, purine-rich nucleic acid binding protein 1; ETS, erythroblast transformation-specific; EICE/EIRE, ETS:IRF composite/response element; IECS, IRF:ETS composite sequence; GCM1, glial cells missing transcription factor 1; ELK1, ETS-like 1; STAT, signal transducer and activator of transcription; GAS, interferon γ-activated sequence; RARE, retinoic acid response element; further abbreviations for bZIP and NR proteins are listed in Figs. 2 and 3.

ancient bZIP motifs share the same half site and differ in spacer length [1]: 12-O-tetradecanoylphorbol-13-acetate (TPA) response element (TRE) contains a one-nucleotide long spacer (5′-TGA(C/G)TCA-3′), while cAMP-response element (CRE) has two spacer nucleotides (5′-TGACGTCA-3′). The former is optimal for members of the activator protein 1 (AP-1) families, primarily FOS/JUN heterodimers; and the latter can be bound by dimers of CRE binding protein (CREB)/activating transcription factor (ATF) family members (Fig. 2). Nevertheless, JUN proteins are capable of binding both TRE and CRE by forming heterodimers with different partners [46]. Certain bZIP dimers are specialized for longer sequences. Musculoaponeurotic fibrosarcoma (MAF) proteins, for example, recognize MAF response elements (MAREs) with 5′-TGCTGA(C/G)-3′ half-sites, which can be considered as upstream extended TRE/CRE half-sites [47,48]. Small MAFs form heterodimers basically with Cap'n'collar (CNC)-type bZIP proteins, which, in turn, also bind the short 5′-TGA(C/G)-3′ half-site, so their shared composite element is essentially a TRE, extended by three nucleotides to one direction. In contrast, dimers of large MAFs bind the inverted repeat of MARE half-sites with a single C/G nucleotide or both in the middle (Fig. 2) [49].

There are additional bZIP proteins that recognize TRE/CRE half-sites and are capable of interacting with bZIP proteins with

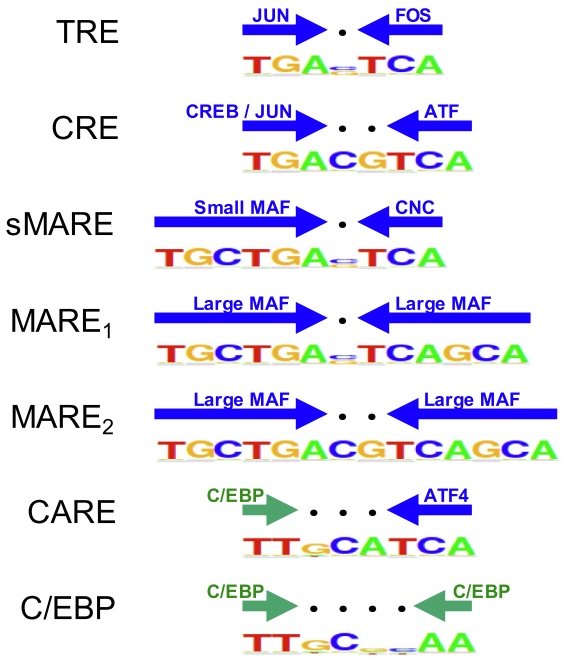

Fig. 2. Schematic representation of major bZIP motifs. TRE/CRE (5′-TGA(C/G)-3′) and MARE (5′-TGCTGA(C/G)-3′) half-sites are marked by blue arrows, C/EBP half sites are marked by green arrows, and spacer nucleotides are marked by black dots. Schematic motif logos and motif/protein names are indicated (TRE, TPA response element; CRE, cAMP response element; CREB, CRE binding protein; ATF, activating transcription factor; MAF, musculoaponeurotic fibrosarcoma protein; CNC, cap′n′collar-type bZIP protein; C/EBP, CCAAT/enhancer-binding protein; (s)MARE, (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.)

different DNA binding features [46]. Several ATF proteins form heterodimers with CCAAT/enhancer-binding proteins (C/EBPs) and can have a composite element other than the merge of the two canonical half-sites. Within the ATF4-specific C/EBP:ATF response element (CARE), ATF4 binds a half-site with a unique spacer (5′-TGA(T)-3′), while its partner binds the canonical C/EBP half-site (5′-TT(GC)-3′) (Fig. 2) [31]. Interestingly, the C/EBP-bound genomic regions contain a large number of motifs built up from a strong and a weak C/EBP half-site, but this motif degeneracy is compensated by shape features created by a directly upstream flanking nucleotide of the half-sites, which can be any nucleotides other than T. This nucleotide preference is characteristic of other bZIP motifs, such as TRE/CRE half-sites, as well [31]. The binding of weak motifs can also be supported by the neighboring elements, as in the case of the promoter-proximal enhancer of inferferon-β (IFNB1) [50]. In the TRE of this enhancer, ATF2 has extended interactions within the major groove (5′-TGA(C)-3′), while JUN has a suboptimal – essentially unrecognizable and barely bound – half-site (5′-TAT(G)-3′), which, in contrast, is optimal for minor groove binding by the neighboring IRF3. Interestingly, the downstream interferon-stimulated response element (ISRE) is also unusual, but this does not affect negatively the DNA-protein interactions and the formation of the so-called enhanceosome [50]. Interaction between AP-1 and IRF proteins is not restricted to this single site. Several AP-1:IRF composite elements (AICEs) could be identified in white blood cells, in which JUNB/BATF and IRF4/8 showed interaction at important immune response genes [51].

7. Motif grammar of nuclear receptors

Specific and selective DNA binding by NRs is determined by most aspects of motif grammar beyond the orientation and spac-

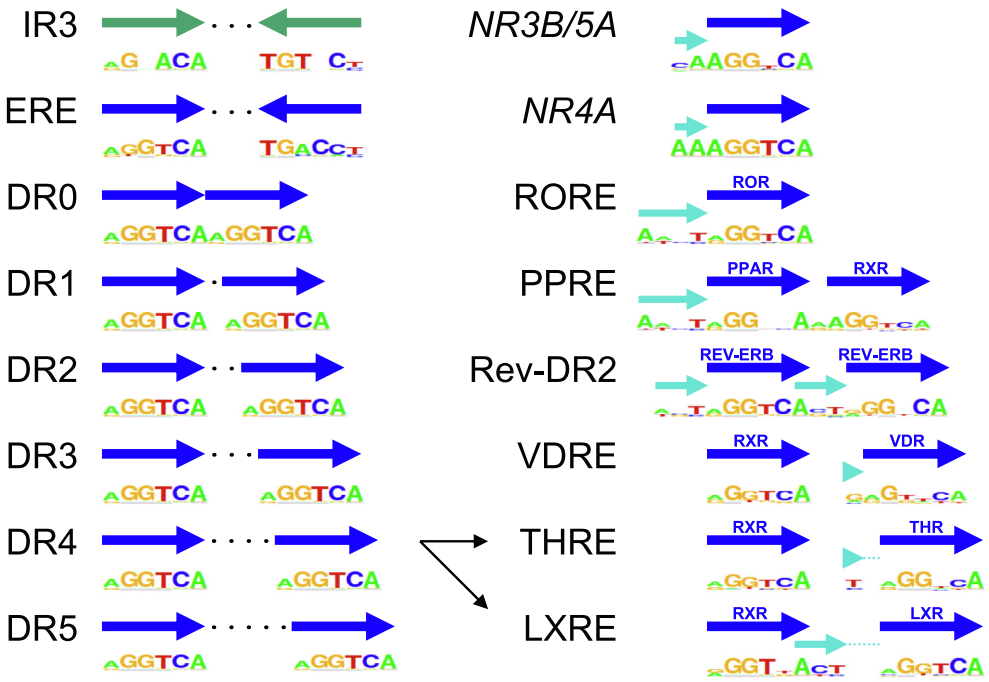

Fig. 3. Schematic representation of nuclear receptor motifs. The general (5′-(A/G)GGTCA-3′) and the NR3C steroid hormone receptor-specific (5′-AGAACA-3′) half-sites are marked by blue or green arrows, respectively. 5′ extensions are marked by cyan arrows, and spacer nucleotides are marked by black dots. 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.)

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′-(A/G)GGTCA-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 type-specific 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)A(C/G)T(A/G)GGTCA-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′-CTNN-3′ sequence, which is bound by the amino-terminal extension (NTE) 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)TCA-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-

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 PPREs, 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 cistromic (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 (KKP129909, K124298) and to G.N. (PD124843). 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 (R01DK115924).

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;43:1–8.
[5] Huber EM, Scharf DH, Hortschansky P, Groll M, Brakhage AA. DNA minor groove sensing and widening by the CCAAT-binding complex. Structure 2012;20:1757–68.
[6] Dailey L, Basilico 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, Gordân R. Protein-DNA binding: complexities and multi-protein codes. Nucleic Acids Res 2014;42:2099–111.
[8] Ijpenberg A, Jeannin E, Wahli 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 gene PPAR response element. J Biol Chem 1997;272:20108–17.
[9] Fujii 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] Nikolov 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;8:27–42.e6.
[12] Rube HT, Rastogi C, Kribelbauer JF, Bussemaker HJ. A unified approach for quantifying and interpreting DNA shape readout by transcription factors. Mol Syst Biol 2018;14:e7902.
[13] Chiu T-P, Xin B, Markarian N, Wang Y, Rohs R. TFBSshape: an expanded motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res 2019;48:D246–55.
[14] Rohs R et al. The role of DNA shape in protein-DNA recognition. Nature 2009;461:1248.
[15] O'Malley RC et al. Cistrome and epicistrome 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.

[26] Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS. Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. Annu Rev Cell Dev Biol 2019;35:357–79.
[27] Jolma A et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature 2015;527:384–8.
[28] Jankowski A, Prabhakar S, Tiuryn J. TACO: a general-purpose tool for predicting cell-type-specific transcription factor dimers. BMC Genom 2014;15:208.
[29] Levitsky V et al. A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package. Nucleic Acids Res 2019;47:e139.
[30] Newman, J. R. S. & Keating, A. E. Comprehensive Identification of Human bZIP Interactions with Coiled-Coil Arrays. Science (80-.). 300, 2097–2101 (2003).
[31] Cohen DM, Lim H-W, Won K-J, Steger DJ. Shared nucleotide flanks confer transcriptional competency to bZip core motifs. Nucleic Acids Res 2018;46:8371–84.
[32] Landschulz, W., Johnson, P. & McKnight, S. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. Science (80-.). 240, 1759–1764 (1988).
[33] Evans RM, Mangelsdorf DJ. Nuclear receptors, RXR, and the big bang. Cell 2014;157:255–66.
[34] Moskow JJ, Bullrich F, Huebner K, Daar IO, Buchberg AM. Meis1, a PBX1-related homeobox gene involved in myeloid leukemia in BXH-2 mice. Mol Cell Biol 1995;15:5434–43.
[35] Rodda DJ et al. Transcriptional regulation of nanog by OCT4 and SOX2. J Biol Chem 2005;280:24731–7.
[36] Eklund EA, Jalava A, Kakar R. PU.1, interferon regulatory factor 1, and interferon consensus sequence-binding protein cooperate to increase gp91 (phox) expression. J Biol Chem 1998;273:13957–65.
[37] Meraro D et al. Protein-protein and DNA-protein interactions affect the activity of lymphoid-specific IFN regulatory factors. J Immunol 1999;163:6468–78.
[38] Meraro D, Gleit-Kielmanowicz M, Hauser H, Levi B-Z. IFN-stimulated gene 15 is synergistically activated through interactions between the myelocyte/lymphocyte-specific transcription factors, PU.1, IFN regulatory factor-8/IFN consensus sequence binding protein, and IFN regulatory factor-4: characterization of a new subtype of IFN-stimulated response element. J Immunol 2002;168:6224–31.
[39] Tamura T, Thotakura P, Tanaka TS, Ko MSH, Ozato K. Identification of target genes and a unique cis element regulated by IRF-8 in developing macrophages. Blood 2005;106:1938–47.
[40] Kurotaki D et al. Essential role of the IRF8-KLF4 transcription factor cascade in murine monocyte differentiation. Blood 2013;121:1839–49.
[41] Brass AL, Kehrli E, Eisenbeis CF, Storb U, Pip Singh H. A lymphoid-restricted IRF, contains a regulatory domain that is important for autoinhibition and ternary complex formation with the Ets factor PU.1. Genes Dev 1996;10:2335–47.
[42] Pearse RN, Feinman R, Shuai K, Darnell JE, Ravetch JV. Interferon gamma-induced transcription of the high-affinity Fc receptor for IgG requires assembly of a complex that includes the 91-kDa subunit of transcription factor ISGF3. Proc Natl Acad Sci 1993;90:4314–8.
[43] Seidel HM et al. Spacing of palindromic half sites as a determinant of selective STAT (signal transducers and activators of transcription) DNA binding and transcriptional activity. Proc Natl Acad Sci USA 1995;92:3041–5.
[44] Li J et al. Structural basis for DNA recognition by STAT6. Proc Natl Acad Sci USA 2016;113:13015–20.
[45] Long HK, Prescott SL, Wysocka J. Ever-changing landscapes: transcriptional enhancers in development and evolution. Cell 2016;167:1170–87.
[46] Deppmann CD, Alvania RS, Taparowsky EJ. Cross-species annotation of basic leucine zipper factor interactions: insight into the evolution of closed interaction networks. Mol Biol Evol 2006;23:1480–92.
[47] Inamdar NM, Ahn YI, Alam J. The heme-responsive element of the mouse heme oxygenase-1 gene is an extended AP-1 binding site that resembles the recognition sequences for MAF and NF-E2 transcription factors. Biochem Biophys Res Commun 1996;221:570–6.
[48] Kurokawa H et al. Structural basis of alternative DNA recognition by maf transcription factors. Mol Cell Biol 2009;29:6232.
[49] Kataoka K, Noda M, Nishizawa M. Maf nuclear oncoprotein recognizes sequences related to an AP-1 site and forms heterodimers with both Fos and Jun. Mol Cell Biol 1994;14:700–12.
[50] Panne D, Maniatis T, Harrison SC. An atomic model of the interferon-beta enhanceosome. Cell 2007;129:1111–23.
[51] Glasmacher E et al. A genomic regulatory element that directs assembly and function of immune-specific AP-1-IRF complexes. Science 2012;338:975–80.
[52] Simandi Z et al. RXR heterodimers orchestrate transcriptional control of neurogenesis and cell fate specification. Mol Cell Endocrinol 2018;471:51–62.
[53] Moutier E et al. Retinoic acid receptors recognize the mouse genome through binding elements with diverse spacing and topology. J Biol Chem 2012;287:26328–41.
[54] Sierk ML, Zhao Q, Rastinejad F. DNA deformability as a recognition feature in the RevErb response element ‡. Biochemistry 2001;40:12833–43.
[55] Chandra V et al. Structure of the intact PPAR-gamma-RXR- nuclear receptor complex on DNA. Nature 2008;456:350–6.
[56] Marciano DP et al. The therapeutic potential of nuclear receptor modulators for treatment of metabolic disorders: PPARγ, RORs, and Rev-erbs. Cell Metab 2014;19:193–208.
[57] Lou X et al. Structure of the retinoid X receptor α-liver X receptor β (RXRα-LXRβ) heterodimer on DNA. Nat Struct Mol Biol 2014;21:277–81.
[58] Feldmann R et al. Genome-wide analysis of LXRα activation reveals new transcriptional networks in human atherosclerotic foam cells. Nucleic Acids Res 2013;41:3518–31.
[59] Rastinejad F, Perlmann T, Evans RM, Sigler PB. Structural determinants of nuclear receptor assembly on DNA direct repeats. Nature 1995;375:203–11.
[60] Orlov I, Rochel N, Moras D, Klaholz BP. Structure of the full human RXR/VDR nuclear receptor heterodimer complex with its DR3 target DNA. EMBO J 2012;31:291–300.
[61] Grøntved L et al. Transcriptional activation by the thyroid hormone receptor through ligand-dependent receptor recruitment and chromatin remodelling. Nat Commun 2015;6:7048.
[62] Tuoresmäki P, Väisänen S, Neme A, Heikkinen S, Carlberg C. Patterns of genome-wide VDR locations. PLoS ONE 2014;9:e96105.
[63] Lala DS, Rice DA, Parker KL. Steroidogenic factor I, a key regulator of steroidogenic enzyme expression, is the mouse homolog of fushi tarazu-factor I. Mol Endocrinol 1992;6:1249–58.
[64] Wilson, T., Paulsen, R., Padgett, K. & Milbrandt, J. Participation of non-zinc finger residues in DNA binding by two nuclear orphan receptors. Science (80-). 256, 107–110 (1992).
[65] Johnston SD et al. Estrogen-related receptor alpha 1 functionally binds as a monomer to extended half-site sequences including ones contained within estrogen-response elements. Mol Endocrinol 1997;11:342–52.
[66] Penvose A, Keenan JL, Bray D, Ramlall V, Siggers T. Comprehensive study of nuclear receptor DNA binding provides a revised framework for understanding receptor specificity. Nat Commun 2019;10:2514.
[67] Nagy G, Daniel B, Cuaranta-Monroy I, Nagy L. Unraveling the hierarchy of *cis* and *trans* factors that determine the DNA binding by PPARγ. Mol Cell Biol 2020. https://doi.org/10.1128/MCB.00547-19.