Running title: Strategies for promoter regulation

Evolution of diverse strategies for promoter regulation

Vaclav Brazda[1], Martin Bartas[2] & Richard P. Bowater[3,*]

[1] Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, 612 65 Brno, Czech Republic

[2] Department of Biology and Ecology/Institute of Environmental Technologies, Faculty of Science, University of Ostrava, Ostrava, Czech Republic

[3] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, United Kingdom

*To whom correspondence should be addressed:

Tel.: 01603 592186

Fax: 01603 592250

e-mail: R.Bowater@uea.ac.uk

Keywords: DNA base sequence, DNA structure, epigenetics, evolution, promoters, transcription

Highlights

- Promoter sequences are crucial for the regulation of transcription and display sequence heterogeneity
- Classical sequence motifs have been characterized for many promoters, and the majority are expected to adopt standard double-helical structures, such as B-form DNA
- Some promoters are abundant in sequences that can form local DNA structures, such as intra-strand hairpins (cruciforms) and G-quadruplexes
- The accessibility of promoters to transcription machinery is dependent on epigenetic modifications of both DNA and proteins; these types of modifications on DNA impact on its structure in the vicinity of the promoter
- Diversification of promoter sequences and structure motifs allows additional complexity to be brought to transcriptional regulation associated with specific cells, tissues, or biological processes, such as stress response and tissue development
- Phylogenetic-based analyses of promoter sequences highlight trends to evolve similar sequence and structure motifs across different species and types of gene transcription machinery
- While most published studies have focused on sequence analyses and/or one feature of a promoter or class of promoter, we summarize data from a wide range of organisms from all kingdoms, with sequences analyzed from various sources and databases

Abstract

DNA is fundamentally important for all cellular organisms due to its role as a store of hereditary, genetic information. The precise and accurate regulation of gene transcription of the genes depends primarily on promoters, which vary significantly within and between genomes. Some promoters are rich in specific types of bases, while others have more varied, complex sequence characteristics. However, it is not only base sequence but also epigenetic modifications and altered DNA structure that regulate promoter activity. Significantly, many promoters across all organisms contain sequences that can form intra-strand hairpins (cruciforms) or 4-stranded structures (G- quadruplex or i-motif). In this review we integrate recent studies on promoter regulation that highlight the importance of DNA structure in the evolutionary adaptation of promoter sequences.

Comparison of promoters across the three domains of life

Transcription is a crucial biological process that allows genetic information to enact its cellular roles. Therefore, it is tightly regulated in all cells, and even a small imbalance in regulatory processes may have fatal consequences, leading to the premature death of the cell or organism, and severe genetic disorders in humans. Various steps are involved in the regulation of gene expression, such as chromatin domain organization, post-transcriptional modification, translation, and mRNA degradation [1-4]. Among the distinct regulated phases of gene expression, transcription initiation is a critical point of regulation [5, 6]. This aspect of transcription is regulated by adjacent gene sequences, especially before the **transcription start sites** (TSS) where the transcription complex forms. These regions of DNA, which are upstream of the 5' DNA sense strand, are called **promoters** (see Glossary). For more general details about promoters, TSS and regulatory elements, see (BOX1).

Here, we review the genetic and molecular details of promoters across all three domains of life that exist on earth: Archaea, Bacteria and Eukarya [1, 7, 8]. Characterization of their varied sequence and structural arrangements highlights diverse mechanisms of gene regulation, but does not identify the genetic mechanisms by which they have evolved as such details are beyond the remit of this review. Many similarities point to the universality of transcription from the earliest life-forms, but significant differences in promoters highlight evolutionary pressures within the different domains. Some studies of promoters focus on prokaryotes as a group that encompasses archaea and bacteria, but these have some major differences in relation to transcription [7]. Therefore, in general, we consider promoters as identified in the three domains of life, and only compare the situation in **prokaryotes** and **eukaryotes** when that is all the data allows.

BOX1 Transcription start site and distance of regulatory elements

Standard terminology defines the site at which transcription starts, the TSS, as + 1. This point is always located at the 5'-end of a gene sequence and its near surrounding is rich in various regulatory sequences, called "promoters". The size of these regulatory sequences, varies between species as well within individual genomes [9], although most promoters have a length, ranging from 100 to 1000 base pairs (bp) [10]. Transcription can also be affected by enhancers, which enhance the transcription of an associated gene, and repressors, which attenuate the transcription of an associated gene. Both enhancers and repressors bind remotely from the "core promoter" (up to 2-3 Mbp) [11], and likely have their effects due to complex organization of the DNA in three dimensional space [12-14].

Due to their importance for cell function, promoters have been intensively studied and many standard promoter motifs have been characterized, with some being shared across all domains of life and some being unique to certain domains (**Figure 1**). Consensus sequences can be defined for many of the regions – for details see "Promoter sequence motifs" below and **Table 1**. In bacteria there are four well-conserved promoter motifs and various other elements (**Figure 1**): the core recognition element (CRE) surrounding the TSS (+ 1), the -10 AT-rich element, the -35 element, and various upstream elements, such as **enhancers** and **silencers** [6]. In archaeal and eukaryotic organisms promoter regions can contain a TATA box, an initiator element (Inr), and a B recognition element (BRE) [15, 16]. Besides these sequences, eukaryotic transcription elements use an abundance of enhancers, silencers, and insulators that are distal to the core promoter region [17]. For example, in eukaryotes there is typically a downstream promoter element (DPE), and many other regions that can be regarded as promoter sequences, such as the CAAT box and the GC box, and other more distant regulatory elements. In addition to these "classical" promoter elements, more recent

observations have described specific sequence elements in promoters, such as “TATA-like”, “motif ten element (MTE)” and “downstream core element (DCE)”. While TATA boxes typically have a consensus sequence of TATAAAA, the TATA-like box sequence (TTTCAA) is more variable and is often located in RNA polymerase (RNAP) III promoters [18]. MTE (with the consensus sequence CSARCSSAACGS) is conserved from *Drosophila melanogaster* to humans, requires the presence of Inr and it is located at positions +18 to +27 in RNAPII promoters [19]. The DCE was discovered in the human β-globin promoter, and its sequence composition is distinct from that of the DPE and is presented with a high incidence in promoters containing a TATA motif [20, 21]. DCE motifs consist of three sub-elements, with consensus sequences as follows: for SI it is "CTTC", SII it is "CTGT", and SIII it is "AGC", which are located approximately at +7/9, +16 to +21 and +31/33 locations, respectively.

Comparison of the processes of transcription at prokaryotic and eukaryotic levels shows they have many more differences than are suggested from this simple analysis of promoters, as we summarize in **Table 1**. RNAPs carry out transcription in all three domains of life, and are regulated by interactions with transcription factors whose number and subunit complexity increase during evolution [8, 22]. Thus, comparisons between the RNAPs from the three domains of life show homology, but promoter architecture differs in complexity. Accurate function of the three eukaryotic RNAPs requires a complex set of transcription factors and a TATA-binding protein (TBP) [22]. Good awareness of the similarities and differences across all three domains of life is provided by focusing on eukaryotic RNAPII. The first high-resolution structure of RNAPII was determined at the turn of the 21st century, and has since been extended and improved to elucidate mechanisms of transcription on eukaryotic genomes [22, 23]. The archaeal transcription system appears to be an ancestral version of the eukaryotic RNAPII, requiring different accessory proteins to function

accurately [8, 24]. Notably, the complexity of promoter architecture increases when viewed in eukaryotes that are complex multicellular organisms compared to single-cell organisms (prokaryotes and eukaryotes), with tissue-specific promoter regulation becoming critical in the former category. Various examples have also been identified for TBP-independent transcription in eukaryotes (e.g. [25]). Such transcription processes clearly have sequence and structural requirements, but these differ from examples described here and are beyond the remit of this review.

Although beyond the focus of this review, it is useful to highlight other differences between transcription processes across the three domains of life. One obvious difference is the presence of introns in mRNA transcribed from eukaryotic genes, which must be removed ("spliced out") before protein synthesis occurs [26]. By contrast, introns are present only rarely in genes from prokaryotes [27]. Another significant difference is that mRNAs in prokaryotes tend to be **polycistronic**; by contrast, eukaryotic mRNAs are usually **monocistronic**, although this is not always the case [27, 28]. Importantly, transcription and translation often occur simultaneously in prokaryotes, but in eukaryotes the RNA is transcribed in the nucleus and translated in the cytoplasm [29].

Promoter databases and in silico tools for promoter prediction

Several promoter databases collect accessible information about validated promoter sequences **(Table 2)**. One of the most comprehensive is the Eukaryotic Promoter Database (**EPDnew**)[i] [30], a collection of databases of experimentally validated promoters for selected eukaryotic model organisms. Currently, 15 organisms are in the database including 10 animals, 2 plants, 2 fungi and a unicellular protozoan parasite. From its most recent update for *Homo sapiens* (October 2019), it includes 29,598 promoters, covering 16,455 genes. Other databases are focused on promoters from plants [31], microbial organisms [32], and across all

organisms in the UCSC Genome Browser [33]. Important details about these databases are provided in **Table 2**.

Promoter sequence motifs

The main promoter sequence motifs have been reviewed repeatedly in various organisms, including archaea [15], bacteria [6] and eukaryotes [16] and are described in more detail below. A useful evaluation of the presence and relative significance of conserved promoter motifs has been done for a group of representative eukaryotes [34], as we highlight below in **Figure 3**. For this representative group of eukaryotic organisms, the AT content in the promoter regions varies in only some organisms in this group, such as plants and animals (**Figure 3A**).

The best characterized DNA sequence in the promoters of eukaryotes – the TATA box (**Figure 1 and Table 1**) – is usually located approximately 25 base pairs upstream of the TSS. Its importance in promoter function is highlighted by the association of **polymorphisms** within it with human hereditary pathologies [35]. The Inr motif is the simple core promoter element found in archaea and eukaryotes (**Figure 1 and Table 1**), and is more prevalent than TATA boxes in the representative examples of eukaryotes (in 53.3% promoters, compared to only 24.4% of promoters containing a TATA box, mean averages in both cases) (**Figure 2**). The CCAAT box is frequently found in eukaryotes, being found in a mean average of 23.2% of promoters (**Figure 2**). GC boxes are also typical for eukaryotic genes (**Figure 1 and Table 1**) with a mean average of 67.4 % of representative human promoters containing this promoter element and it is also often found in both birds and mammals. Further analysis of **Figure 2** confirms large differences in the presence of the standard promoter motifs between particular organisms. For example, while 78.1% of known human promoters have some of these motifs, only 21.1 % of promoters of *Drosophila melanogaster* have such motifs. In the

group of well-characterized model organisms referred to in **Figure 2**, the largest proportion of promoters with TATA motifs exist in *Arabidopsis thaliana* (47.8%), whilst in *Macaca mulatta* this motif was found only in 10.2% of promoters. Interestingly, all mammals and *Gallus gallus* (chicken) have TATA motifs in <20% of promoters. On average, the CCAAT motif is present in about one-quarter of eukaryotic promoters in EPD(new)[i] (with the minimum 13% in *Caenorhabditis elegans* and the maximum 35.1% in *Danio rerio*). The largest range of variation is the occurrence of GC boxes, which is present in 72.4% of promoters in *Canis familiaris*, but in only 6.9% promoters in *A. thaliana*. Interestingly, TATA and GC box proportions are inversely correlated in eukaryotes. Out of these standard promoter motifs, Inr sequences are the most abundant in all organisms on average, ranging from 38.5% in *C. familiaris* to 79.5% in *D. melanogaster* [34].

Special types of promoters are referred to as bidirectional promoters, which are located between TSS of two adjacent genes whose coding sequences are on opposite strands of DNA [36]. The length of the intergenic region is usually less than 1 kbp, and bidirectional promoters in humans are estimated to form up to 10 % of the whole promoter moiety. Significantly increased levels of bidirectional promoters are observed in genomes associated with some diseases, such as cancer [37]. In comparison to average values in human promoters, bidirectional promoters have a different distribution of sequence motifs, with a significant depletion of TATA motifs and enrichment of CpG islands [36], leading to the formation of characteristic chromatin structures [38].

Taken all together, the data from EPD shows that only about 60% of the genes of the model eukaryotic organisms contain experimentally verified promoters with at least one of the standard promoter motif sequences (TATA box, CCAAT box, GC box, Inr) (**Figure 2**). Moreover, the best characterized motif in mammalian promoters – the TATA box – is present in only about 15% of the promoters. Thus, it is apparent that there has been selection for other

features to designate promoter regions. Although some of these features are the additional sequence motifs discussed above (**Figure 1**), it is also clear that the structure of the DNA is a significant element, as we go on to describe in more detail below, including in BOX2.

Promoter structure motifs

The above analysis highlights that base sequences of promoters are only part of the molecular information that regulates transcription. Indeed, a large number of studies have identified a correlation between gene expression and structural properties of promoter DNA [39-41]. The typical structure adopted by double-stranded DNA is the right-handed helix, but DNA molecules can adopt many other conformations [42]. Non-B-DNA structures have historically been called "unusual DNA structures" (see BOX2), however it is clear that these can be readily adopted in DNA and they are proposed to play a variety of cellular regulation roles. Non-B-DNA structures can arise within specific types of base sequences, but different types of sequences can form similar structures, so their characterization is not as straightforward as for sequence-specific motifs [42-45].

As already described, the standard sequence motifs are absent (or variable) in many promoters, so other features must be recognized to initiate transcription. Changes to the structure of DNA are an important determinant of processes that act on it, and altered DNA binding affinity is critical for proteins involved in transcription, with DNA shape playing important roles [14, 46]. DNA structural features, such as protein-induced bendability and intrinsic curvature, are regularly detected in prokaryotic and eukaryotic promoters. The promoters from representative eukaryotes have been evaluated for the presence of various sequence and structural features [34], and we present the relative data in **Figure 3**.

BOX2 Non-B-DNA structures

Non-B-DNA structures are distinct from the classical right-handed, double helical structure (B-DNA) first described by Watson, Crick, Franklin, Wilkins and their co-workers in 1953 [47]. The non-B-DNA structures were initially called “unusual” as they were considered to be rare and physiologically irrelevant. However, in the last 40 years substantial progress in this topic has demonstrated that they play vital roles in regulation of basic molecular and biological processes [42]. Apart from their involvement in regulating gene expression, non-B-DNA structures are critical in DNA replication, telomere maintenance, recombination, and the immune response. Also, it is now very clear that non-B DNA structures are powerful determinants of mutagenesis [48]. There are an abundance of non-B-DNA structures [42], but those that are particularly relevant to gene promoter regions are G-quadruplexes [49, 50], i-motifs [51, 52], cruciforms [53], triplexes [54] and Z-DNA [55], as illustrated in the associated figure. These local, non-B-DNA structures are often targets of protein binding, including various transcription factors [56-58].

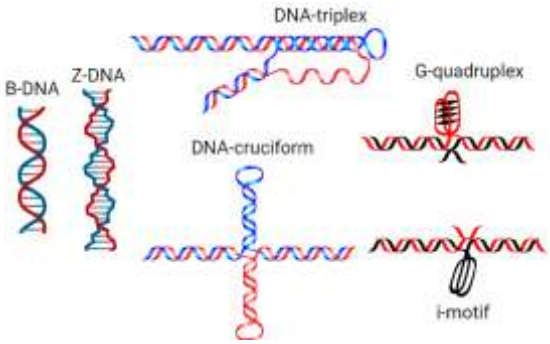


Figure I Schematic drawing of B-DNA and a range of non-B-DNA structures. All structures shown are formed by reorganisation of bases within a double-stranded DNA molecule.

Analysis of the NCBI Medline database finds many published studies that refer to promoters and non-B-DNA structures (**Figure 3F**), and we now focus on some of these examples. One of the simplest structural features that can influence promoter accessibility is

DNA curvature and bendability [14, 59]. Various DNA sequences are prone to bending, but one type that has been well characterized are A-tracts – runs of 4-6 adenine base pairs that are repeated with helical periodicity. A-tracts give rise to intrinsic DNA bending and are often enriched in regulatory regions of genes [60]. In bacteria, the presence of A-tracts upstream of promoters increases promoter activity due to optimization of interactions with the RNA polymerase α subunit [61]. However, these types of sequences can also be found in eukaryotic promoters in close proximity of TSS, especially for Worm and Yeast genomes (**Figure 3C**). A-tracts are comparatively more frequent than in TATA-less promoters and influence nucleosome positioning [62], with their impact identified by a range of high-resolution structures [16].

Completely different structures are formed by G-tracts, which are genomic loci enriched by repeated guanine bases [42]. G-tracts containing at least four copies of the G bases can form DNA secondary structures called **G-quadruplexes** and tracts of six G bases are enriched in the promoter regions, particularly plants and mammals (**Figure 3D**). G-quadruplex motifs are significantly enriched in promoter regions of human genes (within 1 kb upstream of the TSS) [34], with >40% of human gene promoters containing one or more G-quadruplex motifs (**Figure 3E**). Several types of G-quadruplex structures have been demonstrated in eukaryotic genomes, including the human genome [49, 50, 63-67].

Although potential G-quadruplex sequences are more abundant in complex eukaryotic genomes, the presence of G-quadruplex motifs has been demonstrated in many archaeal [68] and bacterial [69] species. For example, the presence of potential G quadruplex sequences have been mapped to promoter regions of *Mycobacterium tuberculosis* [70] and direct regulation of transcription due to G-quadruplex stabilization was shown in the soil bacterium *Paracoccus denitrificans* [71]. Furthermore, G-quadruplexes in promoter regions have been

suggested as a novel therapeutic approach for multi-drug resistant *Klebsiella pneumoniae* [72].

Several studies demonstrate that G-quadruplex prone sequences are abundant in human promoters [49, 73]. Notably, the distinct abundance of G-prone sequences was detected in the promoters of oncogenes, suggesting the potential to use G-quadruplex regulation in cancer treatment [50, 74]. Probably the best explored G-quadruplex in promoters is a sequence from the human *c-Myc* oncogene [75, 76]. Regulatory sequences of this gene contain G-rich regions that can adopt a non-canonical DNA structure, as demonstrated both *in vitro* and *in vivo*, where its formation serves as an important regulator of *c-Myc* expression [41]. Promoter G-quadruplexes can be specifically recognized by small specific ligands [64] and therefore could be targets for gene transcription regulation and targeted medicine [77, 78].

Cytosine-rich i-motifs are four-stranded local DNA structures that are often located on the opposite strand of DNA that contains G-quadruplex motif(s) [79]. There is ongoing debate about whether G-quadruplexes and i-motifs can coexist in the promoter regions, or if they are mutually exclusive *in vivo* [52]. Generally, it is accepted, that i-motifs require a lower pH (below 6) [79], although recent studies such that other cations and ligands can stabilise them at neutral pH [51, 80, 81].

Other non-B-DNA structures shown to be important in promoters are hairpins and the **cruciforms,** which can be formed by inverted repeat sequences that are abundant in many genomes [40, 42, 53]. Hairpins and cruciforms serve as effective targets for several regulatory proteins, including transcription factors such as p53 and p73 [82-84]. Inverted repeats with cruciform-forming potential are non-randomly present in the regulatory regions for initiation and termination of transcription in *Escherichia coli* [85]. Recently, inverted repeat sequences have been shown to be non-randomly distributed in human promoters, with longer inverted repeats enriched in the promoters of genes with specific biological functions, related mainly

to developmental processes, interferon and cancer signaling pathways [40]. Moreover, it has been shown that hairpins could serve as promoter switches in *in vitro* transcription networks [86].

The formation of G-quadruplexes, cruciforms and other non-B-DNA structures is influenced by protein binding and other thermodynamic parameters, such as **DNA supercoiling** [42]. Since changes to chromatin organization associate with changes in gene expression [14], interactions between these different features facilitates complex regulation of promoters. It has been demonstrated that constitutive promoters are less stable, less bendable, and also have lower nucleosome occupancy compared to promoters that have higher variability in expression of the genes they associate with [16, 87]. Cross-talking between these regulatory stimuli has to be taken in account because dynamic changes of local DNA structures in promoter regions have been proposed to facilitate their targeting with therapeutic applications [72, 88].

In summary, there is abundant evidence that non-B-DNA structures are important regulators of gene expression. These types of structures appear to be additional selective markers that can help regulate transcription according to tissue and stress conditions, providing flexible environmental regulatory feedback.

Evolution pressure for promoter diversity

Studies referred to throughout this review demonstrate that promoter sequences are often identified based on an analyses of sequence data and verified by laboratory experiments. A single set of sequence and/or structural promoter features are unlikely to be able to form a promoter for genes under all conditions as these would not be flexible to changing conditions and, therefore, would be likely to be eliminated during evolution. By contrast, genes with flexible promoter features that are capable of sensing and regulating transcription, based on

different external signals, would be likely preserved as genomes evolve. Interestingly, experimental analyses of promoter shapes in *D. melanogaster* genomes show functional properties of natural promoters are an important factor in promoter evolution [89]. Alteration of promoters are also a mechanism that is critical for cell type differentiation [90].

Although all living organisms share the same basic types of genes to support fundamental basic cellular processes, there are important differences in their regulatory sequences. Such sequences are an integral and essential part of the originality of each species and group of organisms. In eukaryotes the genes that encode proteins that function together in metabolism or specific signaling pathways are often located in different regions of the genome, but they are still able to be synchronized for transcription [40]. There is limited data about the synchronization of such promoters being coordinated, but it is clear that they can significantly contribute to differences between individuals within one species. Indeed, examples show that promoter polymorphism and methylation are associated with diseases, such as *PARK2* and colorectal cancer [91], or association of *TNFa* promoter polymorphisms with multiple sclerosis [92]. Similarly, the *hTERT* promoter forms three parallel G-quadruplexes [93], and its mutation and methylation influence human tumorigenesis [94, 95]. Methylation is a type of epigenetic modification that we consider in more detail below.

Analyses of local structures in human promoters show their specific occurrence associates with gene families of the same signaling pathways. This has been demonstrated for cruciform-forming inverted repeats by principal component analysis [40]. Another related example is the influence of G-quadruplexes in promoters, where the size of the G tract influences the stability of the non-B-DNA structure. Although four repeats of the G-tract are enough to form a G-quadruplex, five G-tracts are present in several crucial regulatory regions, which likely guarantees the formation of a G-quadruplex even if one of the G-tracts will be mutated or lost [96]. For these types of sequences, another component of regulation is the

number of Gs in the G-repeat. Whilst the G-quadruplex can be formed by four G bases in the tracts, with an increased number of G bases the stability of the G-quadruplex increases. Interestingly, five and six G bases in the tract are quite rare and are associated predominantly with gene regulatory regions that are species-specific [97]. These facts point to the importance of potential quadruplex-forming sequences being retained during the evolution of processes that regulate transcription.

Evolutionary pressures usually impact on genomes over very long timescales, but viruses and transposable elements offer relatively quick opportunities to alter genomes and have a substantial impact on their evolution. These processes could develop altered (or new) regulatory features in transcription, and have been shown to influence gene expression in eukaryotes [98, 99]. Significantly, active transposable elements can increase the size of the genome and facilitate rapid genetic adaptation to stressful environments [100]. Transposition activity can have positive and negative influences on gene expression and strong promoters from viruses and transposable elements contribute to the rapid genetic diversity of primates [101] and many plant species [102, 103]. It is also of note that stable and conserved G-quadruplexes are located in the LTR promoter of retroviruses [104] and human herpesviruses immediate-early promoters [73]. Repeat sequences with potential to form non-B-DNA structures have a global impact on vertebrate **gene regulatory networks**, with transposable elements being a major catalyst of rearrangements within them throughout evolution [100, 105].

An important element in the regulation of promoters is the accessibility of their TSS, which is critically impacted by epigenetic modifications of both DNA and proteins [106, 107]. In-depth analysis of epigenetic processes is beyond the scope of this review, but their significance is demonstrated by association of epigenetic modifications with several human diseases [106]. Although such modifications are generally considered as a signal for gene

silencing, they can also lead to gene activation [108]. Moreover there is communication between epigenetic modifications and non B-DNA structures [109]. It has been demonstrated that formation of cruciform structures and G-quadruplexes protect DNA from methylation, whilst DNA that is already methylated influences the stability of G-quadruplexes [110]. Epigenetic modifications also influence the stability of i-motifs that can form in some promoter regions and methylation patterns influence their formation *in vivo* [51, 111]. Notably, i-motif forming sequences that can be stabilised at relatively neutral pH [80, 81] were more likely to be epigenetically modified than traditional acidic i-motif forming sequences [51]. Clearly epigenetic modifications of DNA impact on its structure and how it interacts with proteins, but specific details about how these effects influence promoter activity are yet to be disentangled.

In summary, the current view of promoters and their regulation involve several sequence motifs and also complex interactions with structure and epigenetic characteristics. All of these promoter "features" are important parts of the full set of cellular processes that allow effective and dynamic transcriptional regulation (**Figure 4**). It is crucial to think of promoter regulatory sequences as a synergy of sequence and structure-based promoter elements and, importantly, protein-DNA interactions influence localized chromatin structure and promoter activity. Such interactions have been well characterized for nucleosomes in eukaryotes [12, 59], but other proteins have similar effects in prokaryotes. Thus, the dynamics and reversibility of structure formation in promoters is essential to provide transcriptional flexibility in response to cellular stress and environmental conditions. Indeed, recent measurements suggest that intrinsic DNA flexibility is functionally important and must have applied selective pressure throughout the evolution of genomes [14].

Concluding Remarks

Promoter regions are crucial for transcription regulation and, in this review, we summarized their diversity and mode of action. We collated contemporary information about sequences and structures formed in promoter regions, using various sources and databases to highlight important characteristics across all types of promoters from all kingdoms. By focusing on recently emerged data, we highlighted the importance in promoters of non-B-DNA structures, such as G-quadruplexes and cruciforms, demonstrating how the number and frequencies of individual promoter sequences and their structural motifs varies across evolutionary groups and species. Several important questions remain, such as why is it that sequences with potential to form the same type of non-B-DNA structures have a large influence on some promoters but less on others? It is also necessary to examine how these non-B-DNA structures are impacted by other factors, including epigenetic modifications. In attempting to understand how conserved these different structural features are within evolutionary groups and individual species, this questions what is the evolutionary origin of non-B-DNA structures that influence promoter activity? Ultimately, a combination of genetics, biochemistry and molecular studies will lead to answer these questions.

Resources

[i]https://epd.epfl.ch//index.php

[ii]http://www.softberry.com/berry.phtml?topic=plantprom&group=data&subgroup=plantprom

[iii]https://operondb.jp/

[iv]www.phisite.org/main/

[v]http://ppdb.agr.gifu-u.ac.jp/ppdb/cgi-bin/index.cgi

[vi]https://dbtss.hgc.jp/

[vii]http://nucleix.mbu.iisc.ernet.in/prombase/

[viii]www.prodoric.de/

[ix]https://genome.ucsc.edu/

Acknowledgements

This work was supported by The Czech Science Foundation (18-15548S) to VB.

Declaration of Interests

The authors disclose that there are no interests to declare.

References

1. Mack, K.L. and Nachman, M.W. (2017) Gene Regulation and Speciation. Trends Genet 33 (1), 68–80.

2. Holoch, D. and Moazed, D. (2015) RNA-mediated epigenetic regulation of gene expression. Nat Rev Genet 16 (2), 71-84.

3. Talbert, P.B. et al. (2019) Old cogs, new tricks: the evolution of gene expression in a chromatin context. Nat Rev Genet 20 (5), 283-297.

4. Song, J. and Yi, C. (2017) Chemical Modifications to RNA: A New Layer of Gene Expression Regulation. ACS Chem Biol 12 (2), 316-325.

5. Haberle, V. and Stark, A. (2018) Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol 19 (10), 621-637.

6. Chen, J. et al. (2021) Diverse and unified mechanisms of transcription initiation in bacteria. Nat Rev Microbiol 19, 95–109

7. Baker, B.J. et al. (2020) Diversity, ecology and evolution of Archaea. Nat Microbiol 5 (7), 887-900.

8. Abril, A.G. et al. (2020) Prokaryotic sigma factors and their transcriptional counterparts in Archaea and Eukarya. Appl Microbiol Biotechnol 104 (10), 4289-4302.

9. Zheng, W. et al. (2011) Regulatory Variation Within and Between Species. Annu Rev Genomics Hum Genet 12 (1), 327–346.

10. Danino, Y.M. et al. (2015) The core promoter: At the heart of gene expression. Biochim Biophys Acta 1849 (8), 1116–1131.

11. Schoenfelder, S. and Fraser, P. (2019) Long-range enhancer-promoter contacts in gene expression control. Nat Rev Genet 20 (8), 437-455.

12. Kornberg, R.D. and Lorch, Y. (2020) Primary Role of the Nucleosome. Mol Cell 79 (3), 371–375.

13. Andersson, R. and Sandelin, A. (2020) Determinants of enhancer and promoter activities of regulatory elements. Nat Rev Genet 21 (2), 71-87.

14. Basu, A. et al. (2021) Measuring DNA mechanics on the genome scale. Nature 589 (7842), 462-467.

15. Wenck, B.R. and Santangelo, T.J. (2020) Archaeal transcription. Transcription 11 (5), 199-210.

16. Yella, V.R. and Bansal, M. (2017) DNA structural features of eukaryotic TATA-containing and TATA-less promoters. FEBS Open Bio 7 (3), 324-334.

17. Roy, A.L. and Singer, D.S. (2015) Core promoters in transcription: old problem, new insights. Trends Biochem Sci 40 (3), 165–171.

18. Tatosyan, K.A. et al. (2020) TATA-Like Boxes in RNA Polymerase III Promoters: Requirements for Nucleotide Sequences. Int J Mol Sci 21 (10).

19. Shir-Shapira, H. et al. (2019) Identification of evolutionarily conserved downstream core promoter elements required for the transcriptional regulation of Fushi tarazu target genes. PLoS One 14 (4), e0215695.

20. Haberle, V. et al. (2019) Transcriptional cofactors display specificity for distinct types of core promoters. Nature 570 (7759), 122-126.

21. Vo Ngoc, L. et al. (2019) The RNA Polymerase II Core Promoter in Drosophila. Genetics 212 (1), 13-24.

22. Kramm, K. et al. (2019) Transcription initiation factor TBP: old friend new questions. Biochem Soc Trans 47 (1), 411-423.

23. Osman, S. and Cramer, P. (2020) Structural Biology of RNA Polymerase II Transcription: 20 Years On. Annu Rev Cell Dev Biol 36, 1-34.

24. Blombach, F. et al. (2019) Key Concepts and Challenges in Archaeal Transcription. J Mol Biol 431 (20), 4184-4201.

25. Gazdag, E. et al. (2016) Activation of a T-box-Otx2-Gsc gene network independent of TBP and TBP-related factors. Development 143 (8), 1340-50.

26. Gehring, N.H. and Roignant, J.Y. (2021) Anything but Ordinary - Emerging Splicing Mechanisms in Eukaryotic Gene Regulation. Trends Genet 37, 355-372.

27. Setubal, J.C. et al. (2018) Comparative Genomics for Prokaryotes. Methods Mol Biol 1704, 55-78.

28. Cardon, T. et al. (2020) Shedding Light on the Ghost Proteome. Trends Biochem Sci.

29. Yao, R.-W. et al. (2019) Linking RNA Processing and Function. Cold Spring Harb Symp Quant Biol 84, 67–82.

30. Dreos, R. et al. (2014) The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. Nucleic Acids Res 43 (D1), D92–D96.

31. Kusunoki, K. and Yamamoto, Y.Y. (2017) Plant Promoter Database (PPDB). Methods Mol Biol 1533, 299–314.

32. Cao, H. et al. (2019) DOOR: a prokaryotic operon database for genome analyses and functional inference. Brief Bioinform 20 (4), 1568-1577.

33. Navarro Gonzalez, J. et al. (2020) The UCSC Genome Browser database: 2021 update. Nucleic Acids Res 49(D1), D1046-D1057.

34. Bansal, M. et al. (2014) Role of DNA sequence based structural features of promoters in transcription initiation and gene expression. Curr Opin Struct Biol 25, 77–85.

35. Bae, S.H. et al. (2015) Functional analysis of the molecular interactions of TATA box-containing genes and essential genes. PLoS One 10 (3), e0120848.

36. Bagchi, D.N. and Iyer, V.R. (2016) The Determinants of Directionality in Transcriptional Initiation. Trends Genet 32 (6), 322-333.

37. Tu, J. et al. (2019) Characterization of bidirectional gene pairs in The Cancer Genome Atlas (TCGA) dataset. PeerJ 7, e7107.

38. Jangid, R.K. et al. (2018) Bidirectional promoters exhibit characteristic chromatin modification signature associated with transcription elongation in both sense and antisense directions. BMC Genomics 19 (1), 313.
39. Szlachta, K. et al. (2018) Alternative DNA secondary structure formation affects RNA polymerase II promoter-proximal pausing in human. Genome Biol 19.
40. Brázda, V. et al. (2020) Global analysis of inverted repeat sequences in human gene promoters reveals their non-random distribution and association with specific biological pathways. Genomics 112 (4), 2772–2777.
41. Del Mundo, I.M.A. et al. (2017) Alternative DNA structure formation in the mutagenic human c-MYC promoter. Nucleic Acids Res 45 (8), 4929–4943.
42. Brazda, V. et al. (2020) Structures and stability of simple DNA repeats from bacteria. Biochem J 477 (2), 325-339.
43. Cer, R.Z. et al. (2013) Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. Nucleic Acids Res 41 (Database issue), D94-d100.
44. Brázda, V. et al. (2016) Palindrome analyser – A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. Biochem Biophys Res Commun 478 (4), 1739–1745.
45. Brázda, V. et al. (2019) G4Hunter web application: a web server for G-quadruplex prediction. Bioinformatics 35 (18), 3493–3495.
46. Yesudhas, D. et al. (2017) Proteins Recognizing DNA: Structural Uniqueness and Versatility of DNA-Binding Domains in Stem Cell Transcription Factors. Genes 8 (8).
47. Bowater, R.P. and Waller, Z.A.E. (2014) DNA Structure. eLS, John Wiley & Sons, Ltd (Ed.).

48. Guiblet, W.M. et al. (2021) Non-B DNA: a major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. Nucleic Acids Res 49 (3), 1497-1516.

49. Huppert, J.L. and Balasubramanian, S. (2006) G-quadruplexes in promoters throughout the human genome. Nucleic Acids Res 35 (2), 406–413.

50. Balasubramanian, S. et al. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? Nat Rev Drug Discov 10 (4), 261.

51. Wright, E.P. et al. (2020) Epigenetic modification of cytosines fine tunes the stability of i-motif DNA. Nucleic Acids Res 48 (1), 55-62.

52. King, J.J. et al. (2020) DNA G-Quadruplex and i-Motif Structure Formation Is Interdependent in Human Cells. J Am Chem Soc 142 (49), 20600-20604.

53. Fleming, A.M. et al. (2020) Cruciform DNA Sequences in Gene Promoters Can Impact Transcription upon Oxidative Modification of 2'-Deoxyguanosine. Biochemistry 59 (28), 2616–2626.

54. Ou, M. et al. (2020) Long non-coding RNA CDKN2B-AS1 contributes to atherosclerotic plaque formation by forming RNA-DNA triplex in the CDKN2B promoter. EBioMedicine 55, 102694.

55. Ravichandran, S. et al. (2019) Z-DNA in the genome: from structure to disease. Biophys Rev 11 (3), 383-387.

56. Brázda, V. et al. (2014) DNA and RNA quadruplex-binding proteins. Int J Mol Sci 15 (10), 17493–17517.

57. Mishra, S.K. et al. (2016) G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. Sci Rep 6, 38144.

58. Huang, Z.-L. et al. (2018) Identification of G-Quadruplex-Binding Protein from the Exploration of RGG Motif/G-Quadruplex Interactions. J Am Chem Soc 140 (51), 17945–17955.

59. Freeman, G.S. et al. (2014) DNA shape dominates sequence affinity in nucleosome formation. Phys Rev Lett 113 (16), 168101.

60. Bacolla, A. et al. (2015) Local DNA dynamics shape mutational patterns of mononucleotide repeats in human genomes. Nucleic Acids Res 43 (10), 5065-80.

61. Lara-Gonzalez, S. et al. (2020) The RNA Polymerase α Subunit Recognizes the DNA Shape of the Upstream Promoter Element. Biochemistry 59 (48), 4523-4532.

62. Dršata, T. et al. (2014) Mechanical properties of symmetric and asymmetric DNA A-tracts: implications for looping and nucleosome positioning. Nucleic Acids Res 42 (11), 7383-94.

63. Yella, V.R. et al. (2018) Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy. Sci Rep 8 (1), 1–13.

64. Dhamodharan, V. and Pradeepkumar, P.I. (2019) Specific Recognition of Promoter G-Quadruplex DNAs by Small Molecule Ligands and Light-up Probes. ACS Chem Biol 14 (10), 2102–2114.

65. Summers, P.A. et al. (2021) Visualising G-quadruplex DNA dynamics in live cells by fluorescence lifetime imaging microscopy. Nat Commun 12 (1), 162.

66. Hänsel-Hertsch, R. et al. (2020) Landscape of G-quadruplex DNA structural regions in breast cancer. Nat Genet 52 (9), 878-883.

67. Marquevielle, J. et al. (2020) Structure of two G-quadruplexes in equilibrium in the KRAS promoter. Nucleic Acids Res 48 (16), 9336-9345.

68. Brázda, V. et al. (2020) G-Quadruplexes in the Archaea Domain. Biomolecules 10 (9), 1349.

69. Bartas, M. et al. (2019) The Presence and Localization of G-Quadruplex Forming Sequences in the Domain of Bacteria. Molecules 24 (9), 1711.

70. Perrone, R. et al. (2017) Mapping and characterization of G-quadruplexes in Mycobacterium tuberculosis gene promoter regions. Sci Rep 7 (1), 5743.

71. Waller, Z.A. et al. (2016) Control of bacterial nitrate assimilation by stabilization of G-quadruplex DNA. Chem Commun (Camb) 52 (92), 13511–13514.

72. Shankar, U. et al. (2020) Conserved G-Quadruplex Motifs in Gene Promoter Region Reveals a Novel Therapeutic Approach to Target Multi-Drug Resistance Klebsiella pneumoniae. Front Microbiol 11, 1269.

73. Frasson, I. et al. (2019) Conserved G-Quadruplexes Regulate the Immediate Early Promoters of Human Alphaherpesviruses. Molecules 24 (13), 2375.

74. Kawauchi, K. et al. (2020) Photosensitizers Based on G-Quadruplex Ligand for Cancer Photodynamic Therapy. Genes 11 (11).

75. Zheng, B.X. et al. (2020) A small-sized benzothiazole-indolium fluorescent probe: the study of interaction specificity targeting c-MYC promoter G-quadruplex structures and live cell imaging. Chem Commun (Camb) 56 (95), 15016-15019.

76. Rodríguez, J. et al. (2016) Ruthenation of Non-stacked Guanines in DNA G-Quadruplex Structures: Enhancement of c-MYC Expression. Angew Chem Int Ed Engl 55 (50), 15615-15618.

77. Cimino-Reale, G. et al. (2016) Emerging Role of G-quadruplex DNA as Target in Anticancer Therapy. Curr Pharm Des 22 (44), 6612–6624.

78. Asamitsu, S. et al. (2019) Recent Progress of Targeted G-Quadruplex-Preferred Ligands Toward Cancer Therapy. Molecules 24 (3), 429.

79. Sengupta, P. et al. (2021) The Molecular Tête-à-Tête between G-Quadruplexes and the i-motif in the Human Genome. ChemBioChem 22, 1–22.

80. Abdelhamid, M.A. et al. (2018) Redox-dependent control of i-Motif DNA structure using copper cations. Nucleic Acids Res 46 (12), 5886-5893.

81. Abdelhamid, M.A.S. et al. (2019) Destabilization of i-Motif DNA at Neutral pH by G-Quadruplex Ligands. Biochemistry 58 (4), 245-249.

82. Brazda, V. et al. (2017) The structure formed by inverted repeats in p53 response elements determines the transactivation activity of p53 protein. Biochem Biophys Res Commun 483 (1), 516–521.

83. Brázda, V. and Coufal, J. (2017) Recognition of local DNA structures by p53 protein. Int J Mol Sci 18 (2), 375.

84. Čechová, J. et al. (2018) p73, like its p53 homolog, shows preference for inverted repeats forming cruciforms. PLoS one 13 (4), e0195835.

85. Miura, O. et al. (2018) Requirement or exclusion of inverted repeat sequences with cruciform-forming potential in Escherichia coli revealed by genome-wide analyses. Curr Genet 64 (4), 945-958.

86. Kar, S. and Ellington, A.D. (2018) In Vitro Transcription Networks Based on Hairpin Promoter Switches. ACS Synth Biol 7 (8), 1937-1945.

87. Engel, C. et al. (2017) Structural Basis of RNA Polymerase I Transcription Initiation. Cell 169 (1), 120-131.e22.

88. Tateishi-Karimata, H. and Sugimoto, N. (2020) Chemical biology of non-canonical structures of nucleic acids for therapeutic applications. Chem Commun (Camb) 56 (16), 2379–2390.

89. Schor, I.E. et al. (2017) Promoter shape varies across populations and affects promoter evolution and expression noise. Nat Genet 49 (4), 550–558.

90. Ray, M. and Larschan, E. (2020) Getting started: altering promoter choice as a mechanism for cell type differentiation. Genes Dev 34 (9-10), 619–620.

91. Bhat, Z.I. et al. (2019) Association of PARK2 promoter polymorphisms and methylation with colorectal cancer in North Indian population. Gene 682, 25–32.
92. Grigorova, A.A. et al. (2021) Association of polymorphism -308G/A in tumor necrosis factor-alpha gene (TNF-α) and TNF-α serum levels in patients with relapsing-remitting multiple sclerosis. Neurol Res 43, 291-298.
93. Monsen, R.C. et al. (2020) The hTERT core promoter forms three parallel G-quadruplexes. Nucleic Acids Res 48 (10), 5720–5734.
94. Lee, D.D. et al. (2020) DNA methylation of the TERT promoter and its impact on human cancer. Curr Opin Genet Dev 60, 17–24.
95. Lorbeer, F.K. and Hockemeyer, D. (2020) TERT promoter mutations and telomeres during tumorigenesis. Curr Opin Genet Dev 60, 56–62.
96. Fleming, A.M. et al. (2015) A Role for the Fifth G-Track in G-Quadruplex Forming Oncogene Promoter Sequences during Oxidative Stress: Do These "Spare Tires" Have an Evolved Function? ACS Cent Sci 1 (5), 226–233.
97. Bartas, M. et al. (2018) Bioinformatics analyses and in vitro evidence for five and six stacked G-quadruplex forming sequences. Biochimie 150, 70–75.
98. Platt, R.N., 2nd et al. (2018) Mammalian transposable elements and their impacts on genome evolution. Chromosome Res 26 (1-2), 25-43.
99. Stapley, J. et al. (2015) Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Mol Ecol 24 (9), 2241-52.
100. Diehl, A.G. et al. (2020) Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. Nat Commun 11 (1), 1796.
101. Daub, J.T. et al. (2017) Detection of Pathways Affected by Positive Selection in Primate Lineages Ancestral to Humans. Mol Biol Evol 34 (6), 1391–1402.

102. Bariah, I. et al. (2020) Where the Wild Things Are: Transposable Elements as Drivers of Structural and Functional Variations in the Wheat Genome. Front Plant Sci 11, 585515.
103. Fambrini, M. et al. (2020) The plastic genome: The impact of transposable elements on gene functionality and genomic structural variations. Genesis (New York, N.Y.: 2000), e23399.
104. Ruggiero, E. et al. (2019) Stable and Conserved G-Quadruplexes in the Long Terminal Repeat Promoter of Retroviruses. ACS Infect Dis 5 (7), 1150–1159.
105. Miao, B. et al. (2020) Tissue-specific usage of transposable element-derived promoters in mouse development. Genome Biol 21 (1), 255.
106. Cavalli, G. and Heard, E. (2019) Advances in epigenetics link genetics to the environment and disease. Nature 571 (7766), 489-499.
107. Ludwig, A.K. et al. (2016) Modifiers and Readers of DNA Modifications and Their Impact on Genome Structure, Expression, and Stability in Disease. Front Genet 7, 115.
108. Smith, J. et al. (2020) Promoter DNA Hypermethylation and Paradoxical Gene Activation. Trends Cancer 6 (5), 392–406.
109. Mao, S.Q. et al. (2018) DNA G-quadruplex structures mold the DNA methylome. Nat Struct Mol Biol 25 (10), 951-957.
110. Jara-Espejo, M. and Line, S.R. (2020) DNA G-quadruplex stability, position and chromatin accessibility are associated with CpG island methylation. FEBS J 287 (3), 483-495.
111. Školáková, P. et al. (2020) Composite 5-methylations of cytosines modulate i-motif stability in a sequence-specific manner: Implications for DNA nanotechnology and epigenetic regulation of plant telomeric DNA. Biochim Biophys Acta Gen Subj 1864 (9), 129651.
112. Shahmuradov, I.A. et al. (2003) PlantProm: a database of plant promoter sequences. Nucleic Acids Res 31 (1), 114–117.

113. Okuda, S. and Yoshizawa, A.C. (2011) ODB: a database for operon organizations, 2011 update. Nucleic Acids Res 39 (Database issue), D552-5.

114. Klucar, L. et al. (2010) phiSITE: database of gene regulation in bacteriophages. Nucleic Acids Res 38 (Database issue), D366–370.

115. Yamashita, R. et al. (2012) DBTSS: DataBase of Transcriptional Start Sites progress report in 2012. Nucleic Acids Res 40 (Database issue), D150–154.

116. Rangannan, V. and Bansal, M. (2011) PromBase: a web resource for various genomic features and predicted promoters in prokaryotic genomes. BMC research notes 4, 257.

117. Grote, A. et al. (2009) PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. Nucleic Acids Res 37 (Database issue), D61–65.

Tables

Table 1. Major differences and conserved details in the organization of promoter sequences in prokaryotes and eukaryotes.

Promoter feature	Prokaryotes	Eukaryotes	Conserved Consensus Sequence
Pribnow box / TATA box	Yes (-10 motif), Pribnow box in bacteria	Yes, TATA box	TATAAT (Bac) TATAWAWR (Arch/Euk)
TATA-like	No	Yes	TTTCAA (but variable)
-35 motif	Yes	No	TTGACA
B-recognition Element (BRE)	Only for Archaea	Yes	SSRCGCC (BREu) RTDKKKK (BREd)
Initiator Element (Inr)	Only for Archaea	Yes	BBCABW (human)
Downstream Promote Element (DPE)	No	Yes	Several conserved motifs
CCAAT box	No	Yes	CCAAT
GC box	No	Yes	GGGCGG
Motif Ten Element (MTE)	No	Yes	CSARCSSAACGS
Downstream Core Element (DCE)	No	Yes	SI = CTTC, SII = CTGT, SIII = AGC
Enhancers	Rare, varied distances relative to TSS	Yes (multiple), varied distances relative to TSS	No conserved sequence
Silencers	Repressor proteins, varied distances relative to TSS	Multiple, varied distances relative to TSS	No conserved sequence

Notes. Numbers referred to are relative to the transcription start site, TSS, defined as +1. Note that SI, SII and SIII are "sub-elements" of the Downstream Core Element. Sequence details follow standard nomenclature for bases, as follows: "B" = C, G or T; "D" = A, G or T; "K" = G or T; "R" = purines; "S" = C or G; "V" = A, C or G; "W" = A or T; "Y" = pyrimidines. "Arch", "Bac" and "Euk" refers to Archaea, Bacteria and Eukarya, respectively. "BREu" and

"BREd" refer to B-recognition element "upstream" and "downstream", respectively.

Conserved sequences for eukaryotic promoters are from [5].

Table 2. Promoter databases: representative examples of databases that provide details about promoters in different groups of organisms.

Database name	Abbreviated name	Content	URL	Reference
Eukaryotic Promoter Database	EPDnew	15 Reference Eukaryotic organisms	https://epd.epfl.ch//index.php[i]	[30]
Plant Promoter Database	PlantProm	305 entries from monocot, dicot and other plants	http://www.softberry.com/berry.phtml?topic=plantprom&group=data&subgroup=plantprom[ii]	[112]
Database of Prokaryotic Operons	ODB	Database covering 9479 operons	https://operondb.jp[iii]	[113]
Database of Gene Regulation in Bacteriophages	phiSITE	contains more than 700 regulatory elements from 32 bacteriophages from *Siphoviridae*, *Myoviridae* and *Podoviridae* families	http://www.phisite.org/main/ [iv]	[114]
ppdb: Plant Promoter Database ver 3.0	ppdb	573 entries from *Arabidopsis thaliana*, *Oryza sativa*, and *Physcomitrella* patens	http://ppdb.agr.gifu-u.ac.jp/ppdb/cgi-bin/index.cgi[v]	[31]
Database of Transcriptional Start Sites	DBTSS	*Homo sapiens* and *Mus musculus*	https://dbtss.hgc.jp/ [vi]	[115]
The Microbial Promoter Database	PromBase	913 microbial genomes	http://nucleix.mbu.iisc.ernet.in/prombase/[vii]	[116]
Prokaryotic database of gene regulation	PRODORIC	696 prokaryotic genomes	www.prodoric.de/[viii]	[117]
UCSC Genome Browser		*Homo sapiens* and *Mus musculus*	https://genome.ucsc.edu/[ix]	[23]

Figure Legends

Figure 1: Schematic comparison of promoter elements in archaea, bacteria and eukaryotes. The colours indicate different elements that are standard across many promoters for each domain of life, with abbreviations defined in the text. Note that additional non-standard elements have been characterized for promoters in organisms in each domain.

Figure 2. Motifs in representative promoters of selected model eukaryotic organisms. The information was collected in the Eukaryotic Promoter Database (https://epd.epfl.ch/EPDnew_select.php[i]) [30] and subsequently analyzed and edited to this summary table. Nomenclature for motifs follows the definitions given in Table 1. Red colours highlight values below the average, and blue above the average, with more extreme values shown in more intensive colours.

Figure 3: Comparison of promoter regions across a range of eukaryotes. Various elements characterized in eukaryotic promoters were assessed in *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (worm), *Oryza glaberrima* (rice), *Mus musculus* (mouse), and *Homo sapiens* (human), with each panel showing the percentage containing the named element: (A) comparison of AT content; (B) presence of TATA box; (C) presence of A-tract; (D) presence of G-tract; (E) comparison of regions with G-quadruplex motif (percentage with at least one occurrence of the element); (F) NCBI Medline search for "promoter" and "structural features" (IR – inverted repeat, G4 – G-quadruplex, December 1st, 2020). For panels (A) - (E) the colours indicate the distance relative to the transcription start site, TSS (whole genome, blue; -500 bp to +150bp, orange; -150 to +50, grey). Data is adapted from [34].

Figure 4: Interplay of processes that regulate promoter activity. Features included in relation to both sequence (white background) and structure (grey background) are important in the effective usage of promoters and their careful regulation. These regulatory sites are located mainly in the promoter sequence immediately before the TSS, but could also be located at relatively far distances from the TSS, especially in eukaryotes (lower panel). For illustrative purposes the lower panel includes nucleosomes that are present in eukaryotes, but protein-DNA interactions also influence chromatin structure in prokaryotes and are likely to have similar effects on promoter activity. Created with BioRender.com

Figures

Figure 1

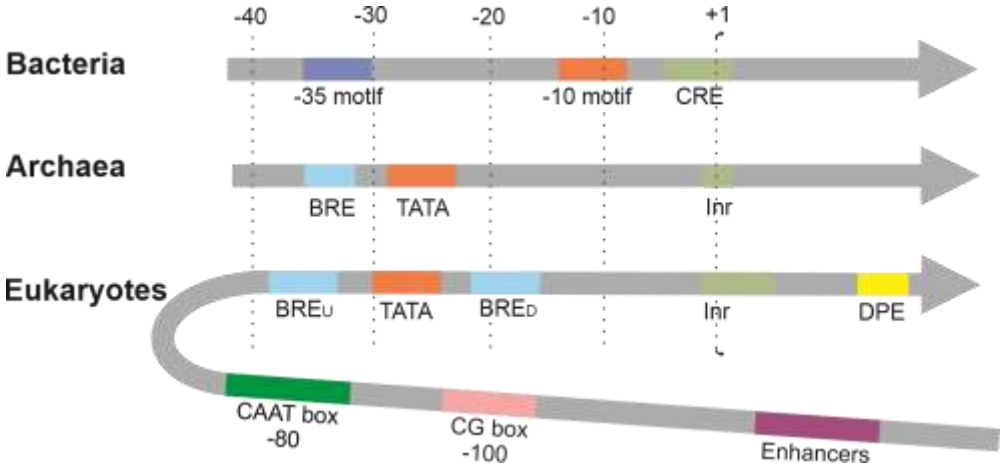


Figure 2

Organism name/Higher taxonomical unit	% of genes with characterized promoters	% of promoters with particular motifs			
		TATA	CCAAT	GC	Inr
Human (*Homo sapiens*)/Mammalia	78.1%	16.1%	24.2%	67.4%	46.1%
Rhesus macaque (*Macaca mulatta*)/Mammalia	66.4%	10.2%	26.0%	69.2%	47.0%
House mouse (*Mus musculus*)/Mammalia	77.7%	18.5%	25.7%	61.8%	48.6%
Brown rat (*Rattus norvegicus*)/Mammalia	76.8%	17.7%	26.7%	64.4%	45.4%
Dog (*Canis familiaris*)/Mammalia	59.6%	14.3%	22.3%	72.4%	38.5%
Red junglefowl (*Gallus gallus*)/Aves	69.3%	16.3%	22.8%	66.4%	40.1%
Zebrafish (*Danio rerio*)/ Actinopterygii	60.4%	32.9%	35.1%	28.1%	47.0%
Roundworm (*Caenorhabditis elegans*)/Chromadorea	66.0%	22.2%	13.0%	21.0%	75.1%
Honey bee (*Apis mellifera*)/Insecta	49.5%	25.5%	18.3%	7.6%	72.1%
Fly (*Drosophila melanogaster*)/Insecta	21.1%	24.7%	17.8%	8.9%	79.5%
Fission yeast (*Schizosaccharomyces pombe*)/Schizosaccharomycetes	42.3%	35.2%	19.4%	8.7%	56.3%
Maize (Zea *mays*)/Monocots	52.8%	35.5%	24.3%	25.3%	45.4%
Thale cress (*Arabidopsis thaliana*)/Eudicots	59.1%	47.8%	25.4%	6.9%	51.7%
Average	**59.9%**	**24.4%**	**23.2%**	**39.1%**	**53.3%**

Figure 3

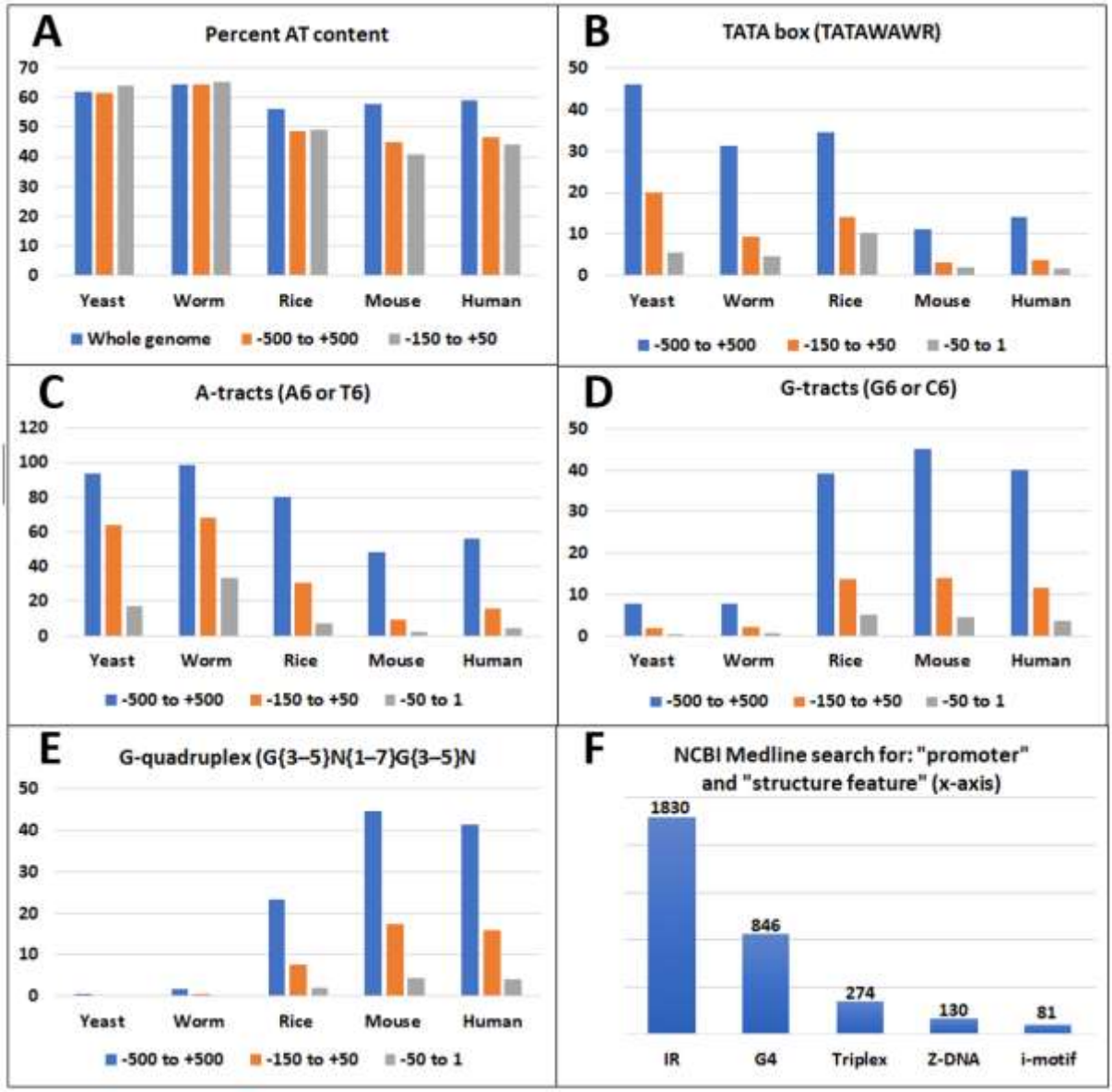
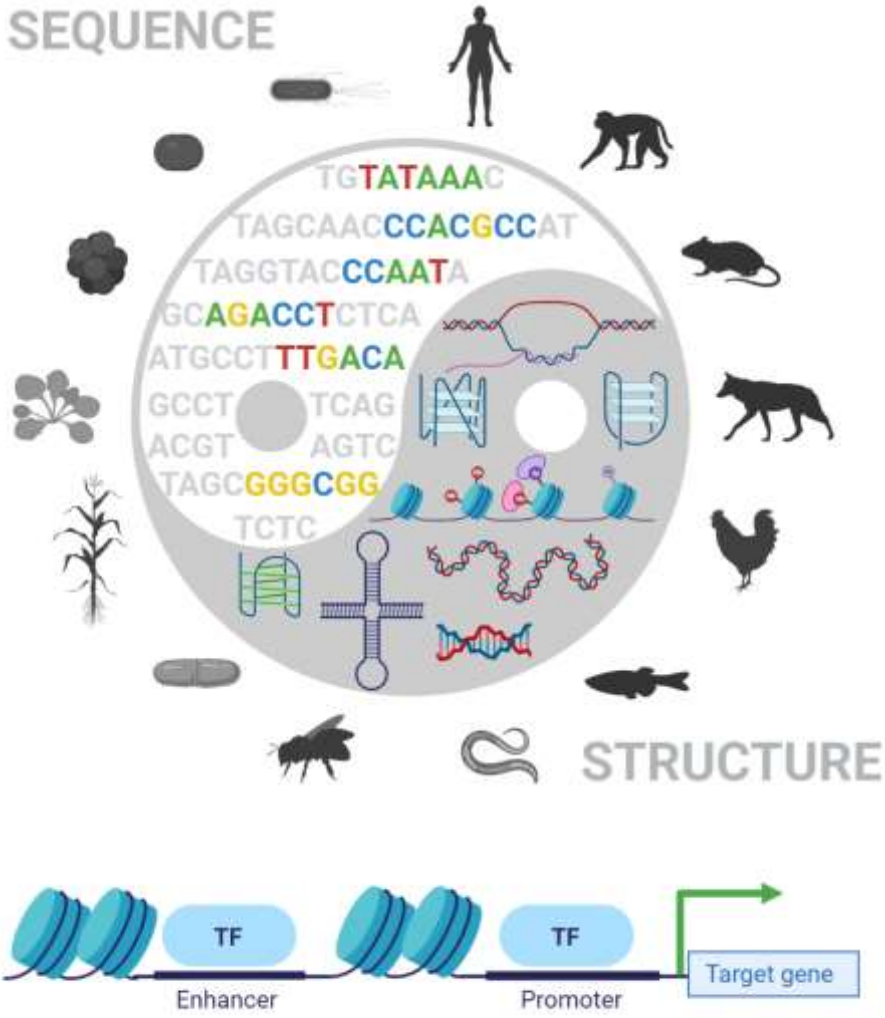

A
Percent AT content
70
60
50
40
30
20
10
0
Yeast
Worm
Rice
Mouse
Human
Whole genome
-500 to +500
-150 to +50
B
TATA box (TATAWAWR)
50
40
30
20
10
0
Yeast
Worm
Rice
Mouse
Human
-500 to +500
-150 to +50
-50 to 1
C
A-tracts (A6 or T6)
120
100
80
60
40
20
0
Yeast
Worm
Rice
Mouse
Human
-500 to +500
-150 to +50
-50 to 1
D
G-tracts (G6 or C6)
50
40
30
20
10
0
Yeast
Worm
Rice
Mouse
Human
-500 to +500
-150 to +50
-50 to 1
E
G-quadruplex (G{3–5}N{1–7}G{3–5}N
50
40
30
20
10
0
Yeast
Worm
Rice
Mouse
Human
-500 to +500
-150 to +50
-50 to 1
F
NCBI Medline search for: "promoter" and "structure feature" (x-axis)
1830
846
274
130
81
IR
G4
Triplex
Z-DNA
i-motif

Figure 4


SEQUENCE
TGTATAAAC
TAGCAACCCACGCCAT
TAGGTACCCAATA
GCAGACCTCTCA
ATGCCTTTGACA
GCCT TCAG
ACGT AGTC
TAGCGGGCGG
TCTC
STRUCTURE
TF
TF
Target gene
Enhancer
Promoter

Outstanding Questions Box

- What is the evolutionary origin of non-B-DNA structures influencing promoter activity?
- Can the different non-B-DNA structures be classified to have different levels of influence on promoter activity? For example, why is it that sequences with potential to form the same type of non-B-DNA structure have a large influence on some promoters but less on others? What other factors impact on these effects?
- How conserved are these different structural features within evolutionary groups and within individual species? Do the different structures have variable impacts in homologous promoters in organisms from different kingdoms?
- G-quadruplexes are common to eukaryotic promoters, but it is not yet clear if they are functionally equivalent across different promoter types. Are they only necessary for the initiation of transcription, or do they have additional, alternative functions in different promoters?
- When epigenetic modifications of DNA impact on non-B-DNA structures in promoter regions, is this a feature that regulates promoter activity in cells?

Glossary

Cruciform: Non-B-DNA structure formed within inverted repeat loci. There are proteins that prefer binding to these structures and modulate molecular processes.

DNA supercoiling: Supercoiling stands for over- or under-winding of DNA strands compared to its preferred helical structure, altering its conformation in three-dimensional space. DNA supercoiling influences many biological processes, including replication and transcription.

Enhancer: DNA region that can be bound by specific proteins to increase the likelihood that transcription of a particular gene will occur.

EPDnew: High precision and high coverage collection of databases of experimentally validated promoters for selected model organisms.

Eukaryotes: Cellular organisms that contain a clearly defined nucleus with its own envelope (or membrane) to separate it from other components of the cell.

Gene regulatory networks: A set of genes that interact with each other to control a specific cell function, such as differentiation, or stress responses.

G-quadruplex: Four-stranded nucleic acid structure dynamically formed in a variety of loci that are rich in guanine nucleotides.

Monocistronic mRNA: A messenger RNA coding for a single protein, which is typical for eukaryotic cells. Contrast with polycistronic mRNA (below).

Non-B-DNA structures: DNA structures differing from double stranded B-DNA form, such as G-quadruplex, cruciform, Z-DNA etc.

Polycistronic mRNA: A messenger RNA that encodes two or more proteins. Polycistronic mRNAs are common in prokaryotes.

Polymorphisms: DNA polymorphisms are variations in the base sequence or length of a gene among a particular sub-set of the gene. The vast majority of DNA polymorphisms are referred to as “silent” because they do not affect the protein sequence and, therefore, have no (or little) biological impact.

Prokaryotes: Cellular organisms lacking envelope-closed nucleus. Comprising the domains bacteria and archaea.

Promoter: DNA region that defines where transcription of a gene by RNA polymerase begins. Promoter sequences are typically located upstream of the transcription start site.

Silencer: DNA region that can be bound by specific proteins to decrease the likelihood that transcription of a particular gene will occur.

Transcription start site: The base at which transcription of a gene starts, i.e. where DNA is transcribed into RNA.