Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Current Opinion in
Structural Biology

Role of DNA sequence based structural features of promoters in transcription initiation and gene expression

Manju Bansal, Aditya Kumar and Venkata Rajesh Yella

Regulatory information for transcription initiation is present in a stretch of genomic DNA, called the promoter region that is located upstream of the transcription start site (TSS) of the gene. The promoter region interacts with different transcription factors and RNA polymerase to initiate transcription and contains short stretches of transcription factor binding sites (TFBSs), as well as structurally unique elements. Recent experimental and computational analyses of promoter sequences show that they often have non-B-DNA structural motifs, as well as some conserved structural properties, such as stability, bendability, nucleosome positioning preference and curvature, across a class of organisms. Here, we briefly describe these structural features, the differences observed in various organisms and their possible role in regulation of gene expression.

Addresses
Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India

Corresponding author: Bansal, Manju (mb@mbu.iisc.ernet.in)

Current Opinion in Structural Biology 2014, **25**:77–85

This review comes from a themed issue on **Theory and simulation**

Edited by **Rommie E Amaro** and **Manju Bansal**

For a complete overview see the Issue and the Editorial

Available online 4th February 2014

0959-440X/$ – see front matter, © 2014 Elsevier Ltd. All rights reserved.

http://dx.doi.org/10.1016/j.sbi.2014.01.007

Introduction

Genomic DNA encompasses two kinds of information: Firstly, coding information stored in the gene body as triplet codons determines the amino acid sequence of proteins, as well as essential non-coding RNAs. Secondly, regulatory information controls essential events such as replication, recombination and transcription initiation, through DNA–protein interactions. The universal triplet code for protein sequence is very well established, but some understanding of a similar code for regulatory information, which is very complex and degenerate in nature, has only recently begun to emerge. Transcription initiation is the first step in the regulation of gene expression. The switch that plays a key role in this process is called the 'promoter' and is generally located upstream of the gene transcription start site (TSS). The identification of TSSs and promoters, within the much larger whole genome sequence is a challenging problem and is generally addressed by locating regions enriched with transcription factor binding sites (TFBSs) associated with protein coding genes and few other conserved sequences, such as TATA box and Initiator element [1–3]. However, recent whole genome 'transcriptome data' have revealed that a large number of transcripts are created which do not code for proteins, but which could have important regulatory roles [4,5]. Several analyses of sequence and structural motifs characterizing TSSs and promoters in various organisms have been reported in recent years and are briefly reviewed here.

TSS and promoter identification using sequence motifs

Several experimental studies have revealed the presence of about 300 transcription factors (TFs) in *E. coli* and that a single TF may bind to hundreds of promoters, while more than 30 TFs may be involved in regulating a single gene promoter [6] which makes it difficult to arrive at a consensus binding sequence for each TF and use the information to *ab initio* identify promoters. Similarly, several hundred octamer sequences have been identified as promoter constituents from *Arabidopsis* and rice [7] while more than 1000 TFs have been manually annotated in the human genome, which along with several other elements, such as CpG islands, are found to regulate transcription [8,9]. Hence it is not surprising that sequence motif search based computational methods have only been moderately successful in identifying the TSSs and TFBSs associated with individual, as well as co-regulated protein coding genes, in both prokaryotes and eukaryotes, even though the models are trained on a particular species [10–15]. A very large number of *in silico* methods have been proposed for eukaryotic promoter prediction, but a comparative analysis of human TSS data indicates that no program simultaneously achieves high sensitivity and positive predictive value and the performance is much improved if it is combined with gene prediction [16,17]. The frequency of occurrence of some well characterized 6-mer (or longer) sequence motifs, such as TATA box, G-quadruplex, oligo-A or G-tracts, in the proximal promoter regions of five model eukaryotic organisms are given in Table 1. It is clearly seen that promoter regions of yeast (*S. cerevisiae*) and *C. elegans* are AT rich and are highly enriched in oligo A-tracts, while being only moderately rich in TATA-box like sequences. A similar trend is observed in prokaryotes [18•] and to a lesser extent in plants [7,19•]. On the other hand, promoter regions of mammals are GC-rich and have

Table 1

Percentage of promoter regions with at least one occurrence of the consensus sequence elements (TATA-box, GGGCGG and GGCGGG) and few other commonly observed structural motifs such as A-tracts (A7 or T7), G-tracts (G7 or C7) and G-quadruplex favoring sequences. Three promoter regions, spanning −500 to +500, −150 to +50, and −50 to −1 (core promoter), with respect to TSS at 0 position have been considered. The total number of 1001-mer promoter sequences for the five eukaryotic model systems are: yeast (*S. cerevisiae*): 4912 [79], worm (*C. elegans*): 18 457 (http://www.modencode.org/), rice (*Oryza sativa*): 24 177 [80], mouse: 17 451 [81], and human: 29 456 [81]. In this table W, R and N refer to A/T, A/G and any nucleotide respectively

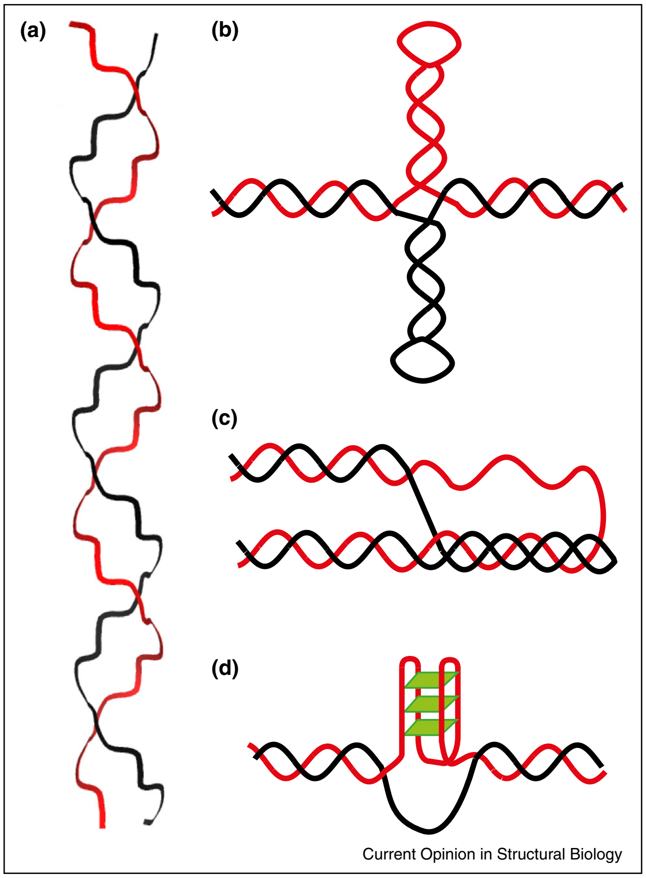
	Yeast	Worm	Rice	Mouse	Human
Percent AT content					
Whole genome	61.9	64.6	56.3	58.0	59.0
−500 to +500	61.5	64.3	48.7	45.0	46.6
−150 to +50	64.1	65.3	49.0	40.9	44.3
TATA box (TATAWAWR)					
−500 to +500	45.9	31.1	34.4	11.0	14.2
−150 to +50	20.1	9.4	14.0	3.1	3.6
−50 to −1	5.5	4.6	10.1	2.1	1.6
A-tracts (A6 or T6)					
−500 to +500	93.7	98.3	80.1	48.6	56.3
−150 to +50	64.2	67.9	30.3	9.2	15.6
−50 to −1	17.3	33.2	7.5	2.1	4.5
G-tracts (G6 or C6)					
−500 to +500	7.8	7.8	39.2	45.2	40.0
−150 to +50	1.8	2.2	13.7	13.9	11.7
−50 to −1	0.2	0.5	5.0	4.5	3.5
G-quadruplex (G{3–5}N{1–7}G{3–5}N{1–7}G{3–5}N{1–7}G{3–5})					
−500 to +500	0.4	1.6	23.3	44.7	41.5
−150 to +50	0.1	0.4	7.5	17.3	15.8
−50 to −1	0.0	0.1	1.9	4.4	4.0
GGGCGG or CCGCCC					
−500 to +500	12.8	16.5	57.3	61.7	53.2
−150 to +50	2.9	4.6	13.1	41.8	32.9
−50 to −1	0.2	1.0	3.5	19.7	14.7
GGCGGG or CCCGCC					
−500 to +500	11.5	15.2	58.7	62.2	55.7
−150 to +50	3.4	3.9	13.5	41.2	32.9
−50 to −1	0.2	0.7	3.6	18.7	14.2

higher prevalence of G-quadruplex sequence motifs (16–17%) and GGGCGG and GGCGGG hexamers which, along with their complementary sequences, are the most abundant 6-mers, with at least one copy being present in ~42% and 33% of proximal promoter sequences in mouse and human promoters respectively (Table 1). These sequences are the conserved part of GC-box, which is recognized by Sp1 transcription factors in mammals [20]. The differences in sequence preferences of promoters, between the various organisms, have important implications for the structural properties of promoter DNA and its function.

Structural features of promoter regions in prokaryotes and eukaryotes:

The right-handed double-helical structure for B-form DNA is the most commonly occurring structure *in vivo* and it was earlier considered to be intrinsically uniform. DNA–protein recognition was hence considered to be the interaction between the DNA-binding structural motifs in proteins and a well-defined spatial arrangement of the hydrogen bond acceptors and donors in the major and minor grooves, as well as the sugar-phosphate backbone of right-handed B-DNA helix [21]. Recent studies however indicate that several proteins interact with DNA through the minor groove, due to its characteristic electrostatic potential [22••]. This suggests that DNA readout by proteins can be broadly classified as direct readout (chemical signatures of the nucleotide bases of DNA) and indirect readout (where the proteins read the 3D-structure of DNA). Indirect readout can be further classified into local shape readout (minor groove size and shape) and global structure readout [23]. An analysis of sequence specific binding of 151 human full length DNA-transcription factors and 303 DNA binding domains has provided considerable information about the role of secondary structural features in TF-DNA binding [24]. Most classes of the TFs have common DNA structure-based-binding motifs, which are characterized by a core-binding sequence flanked by a stretch of A-tracts and their local DNA structural features are important in genome function [25]. Hydroxyl radical cleavage maps of 36 different mammalian systems also showed that, the local topography of DNA is more conserved than primary sequences, and the shape of regulatory regions are conserved through evolution [26]. It should be mentioned that very long stretches of A-tracts (extending to more than 20 nucleotides) are often found in non-promoter DNA and play a

Figure 1



Schematic illustrations showing sequence specific non-B DNA structures that can modulate gene expression. **(a)** left handed, zigzag Z-DNA, **(b)** extruded, four-arm cruciform, **(c)** intra-molecular triplex DNA and **(d)** G-quadruplex structure.

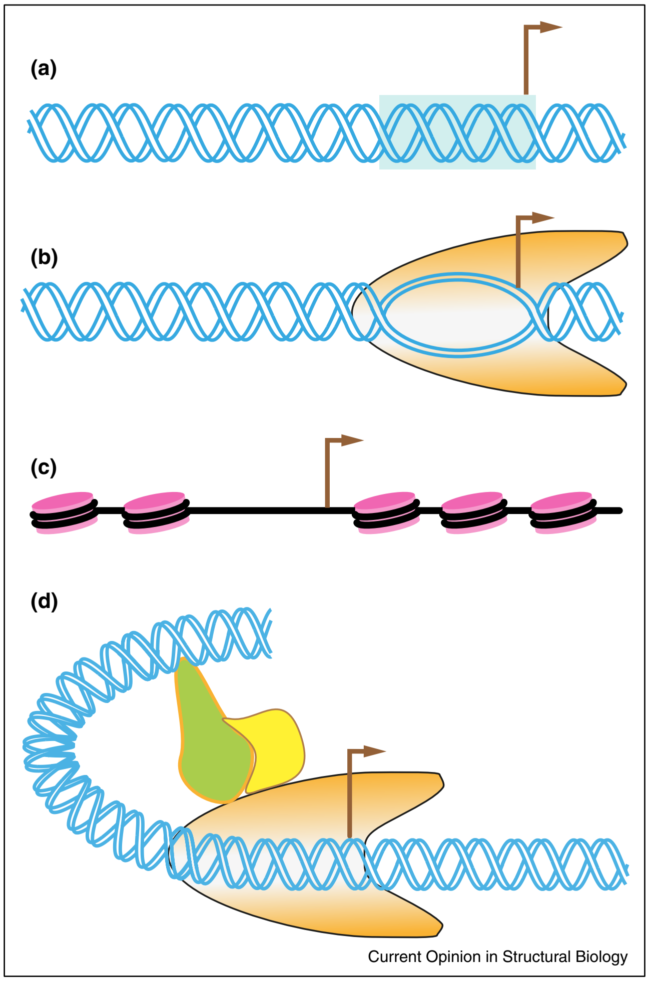
role in chromatin organization [27]. In addition to A-tracts, there are other sequence motifs that can lead to unusual non-B-DNA structures, such as Z-DNA, cruciform structure, triplex DNA and G-quadruplexes, shown in Figure 1(a–d) respectively. These have been shown, in recent years, to exist under physiological conditions and can have functional roles *in vivo*, such as regulating transcription and being structural hotspots for genomic instability. Details of potential non-B-DNA structures in *E coli* genome are available in a recent review [28••].

In higher eukaryotes, GC rich sequences are located near the genes and have been identified as being the 'punctuation marks' for transcription through the formation of left-handed Z-DNA (Figure 1a) at alternating purine–pyrimidine stretches [29,30]. The G-quadruplex (G4) structures (Figure 1d) were first identified for telomeric repeat sequences of chromosomes [31], but such sequences have subsequently been found to be widely prevalent, particularly in the promoters of human genome [32]. The G4 structures have the potential to influence transcription in both positive and negative ways [33,34]. The four-armed cruciform secondary structure (Figure 1b) can be formed when an inverted repeat sequence, more than 6 nucleotide long (and generally AT rich) is present. Such sequences have been found near replication origins and promoters in several organisms [35]. Triplex DNA can form when the pyrimidine strand, in a long stretch of homopyrimidine sequence, loops out and binds to the purine rich strand in the major groove of double helical B-DNA, by forming Hoogsteen type hydrogen bonds. Sequences favoring these structures have been found in human promoters and postulated to be involved in feedback based gene regulation [36].

Structural 'signals' within B-DNA that define promoters

While the effect of non-B-DNA structures on transcription regulation can be readily understood, the role played by small variations in local structure of B-DNA, such as minor groove width, low stability regions and longer range structural features, namely bendability and curvature, in regulation of gene expression, is more subtle and complex. The sequence dependent secondary structural properties of promoter proximal regions have been the subject of intense experimental and computational analysis in recent years. While more than two dozen properties have been examined in some cases [37,38,39•] most are found to be redundant or not significant. Only five or six structural features, such as superhelix induced DNA destabilization [40], intrinsically low stability or stacking energy in prokaryotes [41] higher rigidity as predicted from DNase I cutting sensitivity [42] and nucleosomal positioning preference [43] as well as higher intrinsic curvature [44,45] are consistently observed in promoters of prokaryotes [18•,46–48] as well as lower eukaryotes [18•,37,49]. Mammals are unique in that their TSSs are characterized by flanking regions with higher melting temperature [50,51], The relatively higher AT content in the vicinity of transcription start sites, in all prokaryotes and lower eukaryotes (Table 1) leads to lower stability and easier melting of DNA, which facilitates formation of transcription bubbles (as shown in Figure 2b) and transcription initiation. Similarly, less flexibility or higher rigidity of promoter DNA disfavors formation of nucleoids in prokaryotes and nucleosomes in eukaryotes, making these regions 'nucleosome depleted' and more accessible to the transcription machinery (Figure 2c). DNA bendability can be calculated using several different di-, tri- or tetra-nucleotide based models. The trinucleotide models derived from large scale experimental data on DNase I sensitivity and Nucleosomal Positioning Preference (NPP) are most reliable and only these are discussed here. The DNase I model [42] provides a bendability scale related to the ease of bending towards the major groove, with AT rich trinucleotides assigned a high negative value, which corresponds to lower bendability, while GC rich sequences are more bendable. The NPP

Figure 2

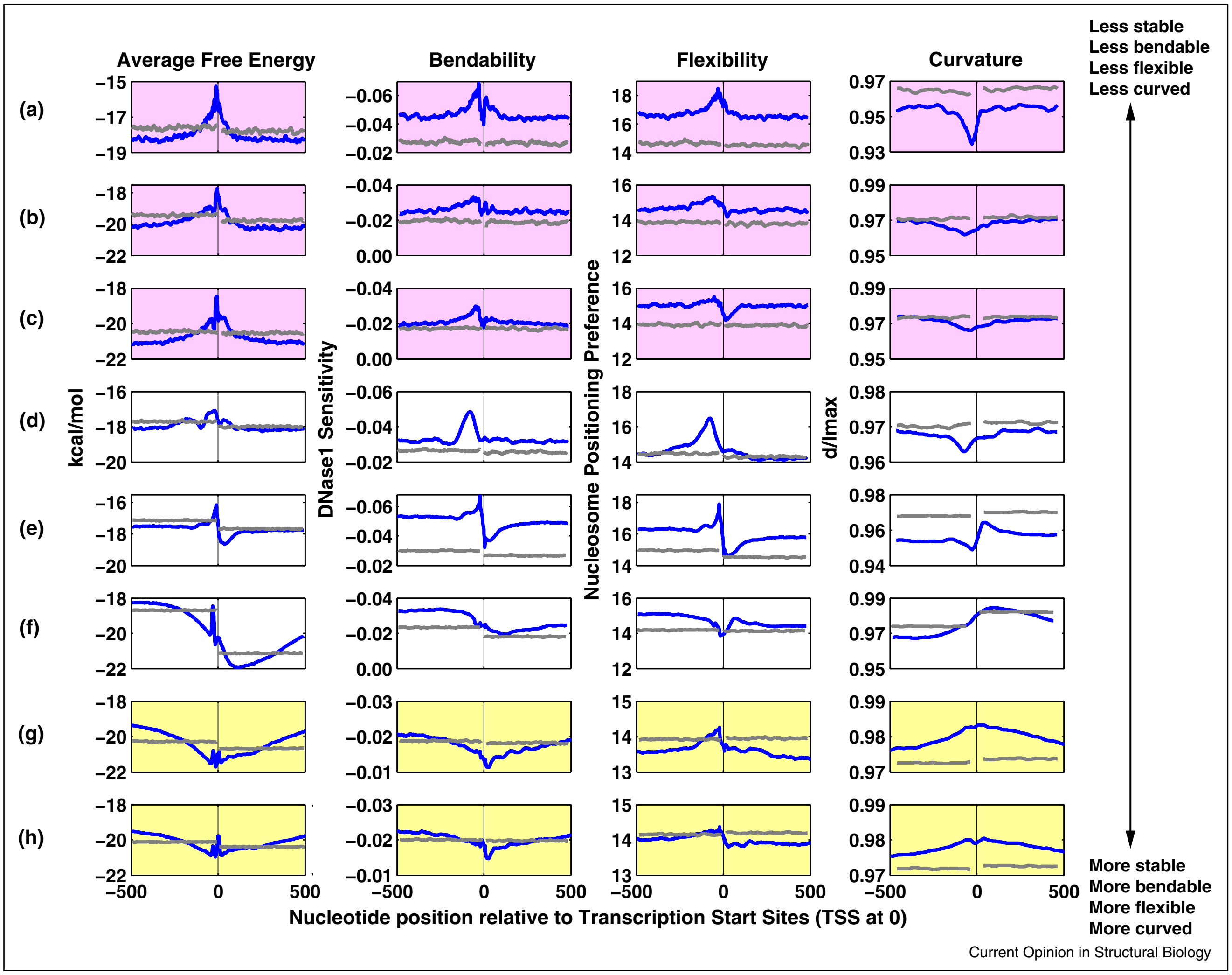


Schematic illustrations showing **(a)** canonical B-DNA along with structural elements representing **(b)** meltable, low stability regions **(c)** nucleosome depleted/free region (NDR/NFR) and **(d)** curved DNA, that are associated with promoter sequences located upstream of transcription start sites (indicated by brown arrows).

model [43] classifies each trinucleotide (and its complementary sequence) on the basis of its minor or major groove to preferentially face the histone core or show no preference. Trinucleotides comprising of only A/T or G/C bases generally show strong preferences in their orientation. Hence a sequence containing clusters of these trinucleotides (i.e. an AT or GC rich region) will make the DNA rigid and lead to a nucleosome depleted region (NDR), while regions rich in trinucleotides with weak preference will be flexible. In addition, long range secondary structural features, such as DNA curvature or looping (shown in Figure 2d) can facilitate interaction between distant regions of DNA, leading to an indirect readout mechanism.

The profiles of the four properties discussed above, in promoter regions of three prokaryotic organisms (*H. pylori*, *E. coli* and *K. pneumoniae*, with AT content of 61%, 49% and 43% respectively) and the five eukaryotic organisms listed in Table 1, are shown in Figure 3, as representative of various domains of life, with varying AT/GC composition. In prokaryotes and lower eukaryotes, such as yeast (*S. cerevisiae*) and worm *(C. elegans*), as well as rice, the promoter regions have AT rich sequences leading to lower average free energy values (Figure 3a–f). This property has been successfully used for promoter identification in bacterial [41,48] and plant genomes [19•]. The structural features of promoter regions in mouse and human (Figure 3g,h) differ from those seen in prokaryotes and lower eukaryotes, since their core promoter regions are GC-rich with about half of the genes containing CpG islands. Hence the promoters have higher free energy/stability as compared to the flanking regions (Figure 3g,h) and this feature has been used for promoter identification [37,50,51]. They also have significant amount of non-B-DNA forming structural motifs such as G-quadruplexes and other GC-rich TFB motifs

Figure 3



Sequence dependent structural properties of the promoter regions (−500 to +500 w.r.t. TSS at 0 position): Profiles of four structural properties are shown, for promoter regions of eight model systems which represent various domains of life: Average Free Energy for a 15-mer window, calculated using dinucleotide free energy values for the 16 dinucleotides [75], DNase1 Sensitivity [42] and Nucleosome Positioning Preference [43] calculated over 30-mer windows and curvature represented by d/l_{max} for a 75-mer fragment generated using 'wedge angles' for the 16 dinucleotides [76]. The rows (a–h) correspond to *H. pylori*, *E. coli*, *K. pneumoniae*, *S. cerevisiae*, *C. elegans*, rice, mouse and human, with the total number of 1001-mer promoter sequences being 714 [4], 1350 [77], 2170 [78], 4912 [79], 18 457 (http://www.modencode.org/), 24 177 [80], 17 451 [81], and 29 456 [81] respectively. The plots with pink and yellow background correspond to prokaryotes and mammals respectively, which show similar characteristic features within their domain, but differ considerably from each other. Gray lines in each subplot correspond to the feature values of up- and down-stream shuffled sequences.

(Table 1). Interestingly, in the AT rich promoters of bacteria, yeast and *C. elegans*, the TATA box sequences are found in only 10–20% promoters, while there is a very high occurrence (>60%) of A-tract sequences. These have the potential to form local structures with narrow minor groove and when occurring in a phased manner in longer DNA stretches, they are also expected to show enhanced curvature [44,45,52], as seen in last column of Figure 3(a–f) for promoter regions of prokaryotes, yeast, worm (*C. elegans*) and even rice. Site-directed mutagenesis and electrophoresis studies show that curved DNA sequences are prominently present in the regulatory regions of operons in *E. coli* highlighting the role of curvature in gene regulation [53]. Also in *E. coli*, the nucleoid associated protein Fis binds preferentially to adjacent major grooves interfaces, the correct spacing being achieved by compression of the central minor groove which contains a conserved AT-rich region [54]. These act as upstream activating sequences (UAS) of promoter regions and form a toroidal micro-loop attached to RNA polymerase [55]. On whole genome scale, thermodynamic stability and superhelicity has been found to be associated with the polarity of the chromosome and gene expression in bacterial growth cycle [56]. The mouse and human promoter regions, being GC rich, are predicted as being less curved than the flanking regions as well as the shuffled sequences and their transcription initiation is regulated by a different mechanism.

The high occurrence of A-tracts in bacterial and lower eukaryotes also suggests that their promoter regions are less bendable or flexible (as seen in columns 2 and 3 of Figure 3). The whole region upstream of TSS in rice promoters is relatively rigid, but they do not seem to have any strong bendability features in the core promoter region. Interestingly, mouse and human promoters, being GC rich, are predicted to be quite bendable by the DNase I sensitivity criteria, while the NPP model indicates lower flexibility for the promoter regions. This is explained by the fact mentioned earlier, *viz.* the NPP model classifies both AT and GC rich trinucleotides as being less flexible and hence disfavoring nucleosome occupancy. The barrier posed by the rigid promoter regions has also been invoked to explain the promoter nucleosome occupancy, on the basis of statistical positioning signals [57]. The high GC content and width of CpG islands in mammalian promoters has been correlated with nucleosome depletion, as well as its positioning and interestingly, it is found that the nucleosome exclusion is independent of transcription [58•]. Several recent *in silico* studies have attempted to characterize the transcription regulation elements, at the genomic scale, by analyses of one or more unique structural properties of DNA in the promoter regions, which seem to be conserved from prokaryotes to lower eukaryotes [19•,25,46,48,59–61]. In mammalian genomes, the greater stability of promoter regions has been used to identify TSSs [37,50], but it is the higher order structure of DNA that plays a significant role in transcription initiation and regulation.

DNA structure and regulation of gene expression

Gene expression and its variability has been extensively studied in yeast and human [62•,63] and nucleosome organization is understood to be a major player in regulating transcription. It has been suggested by Jiang and Pugh [64] that for eukaryotes 'the entire genome can be thought of as a continuous thermodynamic landscape, in which NFRs represent the least favorable regions and the +1 or −1 nucleosome positions represent the most favorable regions'. The −1 nucleosome, NFR (or NDR) and +1 nucleosome arrangement seems to be common among eukaryotes, regardless of the GC composition of the genome. NDR in yeast possess the anti-nucleosomal sequences (such as oligo A-tracts) along with the binding sites for general transcription machinery, while, the −1 and +1 nucleosome associated regions prefer sequences that favor bending of DNA as well as interaction with the histone proteins and manipulating these can regulate gene expression [65]. The structural features of sequences responsible for nucleosome arrangement may also be expected to differ for genes with different expression levels [57,62•]. Two classes of yeast genes, growth genes (ribosomal genes) and stress genes (many cell wall proteins) have been shown to have different promoter nucleosome occupancy [66••]. Growth genes are regulated by TFIID, have TATA-less promoters and their promoter regions have been shown to be highly rigid compared to those of stress genes. They have also been shown to have two well positioned nucleosomes flanking the NDR, which is located immediately upstream of TSS. On the other hand, stress genes have delocalized nucleosomes and promoter regions are often occupied by nucleosomes [66••]. They have to depend on signal induced nucleosome eviction at promoter regions to control their expression. In fact genes with high expression variability have been found to have non-canonical chromatin environment and are controlled by chromatin regulators and bound transcription factors [62•]. A recent study on variability of gene expression has attempted to establish the relationship between DNA structural features and gene expression. This study on *S. cerevisiae* shows that promoters of genes with low variability in expression plasticity (responsiveness) have lower stability, more rigidity, and lower nucleosome occupancy, as compared to genes with high expression plasticity [67].

Conclusion

Promoter regions in prokaryotes and lower eukaryotes are AT rich and their sequence dependent structural properties, such as free energy, bendability and curvature, differ significantly from those of the flanking sequences and play an important role in transcription initiation and regulation [28••,39•,56,68]. In eukaryotes, the promoters contain GC rich sequence motifs which disfavor nucleosome formation, a major factor in transcription regulation [62•,69,70]. After much debate, on whether there exists a precise 'genomic code' for nucleosome positioning, there seems to be a consensus that, a combination of DNA sequence, nucleosome remodeling enzymes and transcription factors, as well as few other factors, determine nucleosome positioning. These properties can differ between genes and interplay between them affects level of gene expression [71••]. However, some characteristic sequence features, unique to yeast and human core promoters seem to be good predictors of high promoter activity [72].

In addition, pervasive transcription leading to spurious transcripts and cryptic transcription has been postulated to modulate gene expression, as well as telomere silencing etc. [73,74]. Hence understanding the structural features of promoters can facilitate identification of TSSs associated with these less characterized transcripts.

Acknowledgement

MB is a recipient of JC Bose National Fellowship of DST, India. We thank Asmita Gupta for assistance in preparation of Figures 1 and 2.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Fickett JW, Hatzigeorgiou AG: **Eukaryotic promoter recognition**. *Genome Res* 1997, **7**:861-878.

2. Blanchette M, Tompa M: **Discovery of regulatory elements by a computational method for phylogenetic footprinting**. *Genome Res* 2002, **12**:739-748.

3. Maston GA, Evans SK, Green MR: **Transcriptional regulatory elements in the human genome**. *Annu Rev Genom Hum Genet* 2006, **7**:29-59.

4. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R *et al.*: **The primary transcriptome of the major human pathogen Helicobacter pylori**. *Nature* 2010, **464**:250-255.

5. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM: **A high-resolution map of transcription in the yeast genome**. *Proc Natl Acad Sci U S A* 2006, **103**:5320-5325.

6. Ishihama A: **Prokaryotic genome regulation: a revolutionary paradigm**. *Proc Jpn Acad Ser B Phys Biol Sci* 2012, **88**:485-508.

7. Yamamoto YY, Ichida H, Abe T, Suzuki Y, Sugano S, Obokata J: **Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis**. *Nucleic Acids Res* 2007, **35**:6219-6226.

8. Fulton DL, Sundararajan S, Badis G, Hughes TR, Wasserman WW, Roach JC, Sladek R: **TFCat: the curated catalog of mouse and human transcription factors**. *Genome Biol* 2009, **10**:R29.

9. Valen E, Sandelin A: **Genomic and chromatin signals underlying transcription start-site selection**. *Trends Genet* 2011, **27**:475-485.

10. Ohler U, Niemann H, Liao G, Rubin GM: **Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition**. *Bioinformatics* 2001, **17(Suppl 1)**:S199-S206.

11. Bulyk ML, McGuire AM, Masuda N, Church GM: **A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli***. *Genome Res* 2004, **14**:201-208.

12. Burden S, Lin YX, Zhang R: **Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences**. *Bioinformatics* 2005, **21**:601-607.

13. Jacques PE, Rodrigue S, Gaudreau L, Goulet J, Brzezinski R: **Detection of prokaryotic promoters from the genomic distribution of hexanucleotide pairs**. *BMC Bioinform* 2006, **7**:423.

14. Das MK, Dai HK: **A survey of DNA motif finding algorithms**. *BMC Bioinform* 2007, **8(Suppl 7)**:S21.

15. Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juarez K, Contreras-Moreira B *et al.*: **Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli***. *PLoS ONE* 2009, **4**:e7526.

16. Bajic VB, Tan SL, Suzuki Y, Sugano S: **Promoter prediction analysis on the whole human genome**. *Nat Biotechnol* 2004, **22**:1467-1473.

17. Bajic VB, Brent MR, Brown RH, Frankish A, Harrow J, Ohler U, Solovyev VV, Tan SL: **Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment**. *Genome Biol* 2006, **7(Suppl 1)**:1-13 S3.

18. • Kanhere A, Bansal M: **Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes**. *Nucleic Acids Res* 2005, **33**:3165-3175.

One of the first reports on the characteristic DNA structural features of promoters in prokaryotes, vertebrates and plants.

19. • Morey C, Mookherjee S, Rajasekaran G, Bansal M: **DNA free energy-based promoter prediction and comparative analysis of Arabidopsis and rice genomes**. *Plant Physiol* 2011, **156**:1300-1315.

This bioinformatic analysis used a DNA stability based promoter prediction algorithm to successfully annotate the promoter regions of *Arabidopsis* and rice genomes.

20. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: **A census of human transcription factors: function, expression and evolution**. *Nat Rev Genet* 2009, **10**:252-263.

21. Luscombe NM, Laskowski RA, Thornton JM: **Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level**. *Nucleic Acids Res* 2001, **29**:2860-2874.

22. •• Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B: **The role of DNA shape in protein–DNA recognition**. *Nature* 2009, **461**:1248-1253.

This comprehensive analysis of DNA–protein complex crystal structures, reported a novel protein–DNA recognition strategy.

23. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS: **Origins of specificity in protein–DNA recognition**. *Annu Rev Biochem* 2010, **79**:233-269.

24. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G *et al.*: **DNA-binding specificities of human transcription factors**. *Cell* 2013, **152**:327-339.

25. Greenbaum JA, Parker SC, Tullius TD: **Detection of DNA structural motifs in functional genomic elements**. *Genome Res* 2007, **17**:940-946.

26. Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH: **Local DNA topography correlates with functional noncoding regions of the human genome**. *Science* 2009, **324**:389-392.

27. Segal E, Widom J: **Poly(dA:dT) tracts: major determinants of nucleosome organization**. *Curr Opin Struct Biol* 2009, **19**:65-71.

28. •• Du X, Wojtowicz D, Bowers AA, Levens D, Benham CJ, Przytycka TM: **The genome-wide distribution of non-B DNA motifs is shaped by operon structure and suggests the transcriptional importance of non-B DNA structures in *Escherichia coli***. *Nucleic Acids Res* 2013, **41**:5965-5977.

An *in silico* analysis of putative B DNA–non-B DNA transition-susceptible motifs in regulatory regions of *E. coli*.

29. Rich A, Zhang S: **Timeline: Z-DNA: the long road to biological function**. *Nat Rev Genet* 2003, **4**:566-572.

30. Khuu P, Sandor M, DeYoung J, Ho PS: **Phylogenomic analysis of the emergence of GC-rich transcription elements**. *Proc Natl Acad Sci U S A* 2007, **104**:16528-16533.

31. Williamson JR, Raghuraman MK, Cech TR: **Monovalent cation-induced structure of telomeric DNA: the G-quartet model**. *Cell* 1989, **59**:871-880.

32. Huppert JL, Balasubramanian S: **G-quadruplexes in promoters throughout the human genome**. *Nucleic Acids Res* 2007, **35**:406-413.

33. Qin Y, Hurley LH: **Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions**. *Biochimie* 2008, **90**:1149-1171.

34. Balasubramanian S, Hurley LH, Neidle S: **Targeting G-quadruplexes in gene promoters: a novel anticancer strategy?** *Nat Rev Drug Discov* 2011, **10**:261-275.

35. van Holde K, Zlatanova J: **Unusual DNA structures, chromatin and transcription**. *Bioessays* 1994, **16**:59-68.

36. Buske FA, Bauer DC, Mattick JS, Bailey TL: **Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data**. *Genome Res* 2012, **22**:1372-1381.

37. Abeel T, Saeys Y, Bonnet E, Rouze P, Van de Peer Y: **Generic eukaryotic core promoter prediction using structural features of DNA**. *Genome Res* 2008, **18**:310-323.

38. Brick K, Watanabe J, Pizzi E: **Core promoters are predicted by their distinct physicochemical properties in the genome of *Plasmodium falciparum***. *Genome Biol* 2008, **9**:R178.

39. • Gan Y, Guan J, Zhou S: **A comparison study on feature selection of DNA structural properties for promoter prediction**. *BMC Bioinform* 2012, **13**:4.
In this study, authors have examined various DNA structural features in promoter regions of human genome. Out of a dozen features, energy related features are found to be highly correlated with promoter regions.

40. Wang H, Noordewier M, Benham CJ: **Stress-induced DNA duplex destabilization (SIDD) in the *E. coli* genome: SIDD sites are closely associated with promoters**. *Genome Res* 2004, **14**:1575-1584.

41. Kanhere A, Bansal M: **A novel method for prokaryotic promoter prediction based on DNA stability**. *BMC Bioinform* 2005, **6**:1.

42. Brukner I, Sanchez R, Suck D, Pongor S: **Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides**. *EMBO J* 1995, **14**:1812-1818.

43. Satchwell SC, Drew HR, Travers AA: **Sequence periodicities in chicken nucleosome core DNA**. *J Mol Biol* 1986, **191**:659-675.

44. Bansal M: **Structural variations observed in DNA crystal structures and their implications for protein–DNA interaction**. In *Biological Structure and Dynamics*, vol 1. Edited by Sharma RH, Sharma M. Adenine Press; 1996.

45. Nov Klaiman T, Hosid S, Bolshoy A: **Upstream curved sequences in *E. coli* are related to the regulation of transcription initiation**. *Comput Biol Chem* 2009, **33**:275-282.

46. Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW: **A DNA structural atlas for *Escherichia coli***. *J Mol Biol* 2000, **299**:907-930.

47. Rangannan V, Bansal M: **Relative stability of DNA as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition**. *Mol Biosyst* 2009, **5**:1758-1769.

48. Rangannan V, Bansal M: **High-quality annotation of promoter regions for 913 bacterial genomes**. *Bioinformatics* 2010, **26**:3043-3050.

49. Florquin K, Saeys Y, Degroeve S, Rouze P, Van de Peer Y: **Large-scale structural analysis of the core promoter in mammalian and plant genomes**. *Nucleic Acids Res* 2005, **33**:4255-4264.

50. Abeel T, Van de Peer Y, Saeys Y: **Toward a gold standard for promoter prediction evaluation**. *Bioinformatics* 2009, **25**:i313-i320.

51. Dineen DG, Wilm A, Cunningham P, Higgins DG: **High DNA melting temperature predicts transcription start site location in human and mouse**. *Nucleic Acids Res* 2009, **37**:7360-7367.

52. Haran TE, Kahn JD, Crothers DM: **Sequence elements responsible for DNA curvature**. *J Mol Biol* 1994, **244**:135-143.

53. Olivares-Zavaleta N, Jauregui R, Merino E: **Genome analysis of *Escherichia coli* promoter sequences evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated**. *Genomics* 2006, **87**:329-337.

54. Stella S, Cascio D, Johnson RC: **The shape of the DNA minor groove directs binding by the DNA-bending protein Fis**. *Genes Dev* 2010, **24**:814-826.

55. Skoko D, Yoo D, Bai H, Schnurr B, Yan J, McLeod SM, Marko JF, Johnson RC: **Mechanism of chromosome compaction and looping by the *Escherichia coli* nucleoid protein Fis**. *J Mol Biol* 2006, **364**:777-798.

56. Sobetzko P, Glinkowska M, Travers A, Muskhelishvili G: **DNA thermodynamic stability and supercoil dynamics determine the gene expression program during the bacterial growth cycle**. *Mol Biosyst* 2013, **9**:1643-1651.

57. Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF: **A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome**. *Genome Res* 2008, **18**:1073-1083.

58. • Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I *et al.*: **CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters**. *Genome Res* 2012, **22**:2399-2408.
This study reported that, the CpG islands and GC content of mammalian promoters is correlated with nucleosome occupancy and positioning, in the context of transcription.

59. Wang H, Benham CJ: **Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress**. *BMC Bioinform* 2006, **7**:248.

60. Goni JR, Perez A, Torrents D, Orozco M: **Determining promoter location based on DNA structure first-principles calculations**. *Genome Biol* 2007, **8**:R263.

61. Meysman P, Marchal K, Engelen K: **DNA structural properties in the classification of genomic transcription regulation elements**. *Bioinform Biol Insights* 2012, **6**:155-168.

62. • Choi JK, Kim YJ: **Intrinsic variability of gene expression encoded in nucleosome positioning sequences**. *Nat Genet* 2009, **41**:498-503.
This comprehensive study on gene expression variability in yeast establishes its correlation with promoter nucleosome occupancy. Authors have shown that the promoter regions of genes with higher expression variability possess non-canonical nucleosome occupancy.

63. **A user's guide to the encyclopedia of DNA elements (ENCODE)**. *PLoS Biol* 2011, **9**:e1001046.

64. Jiang C, Pugh BF: **Nucleosome positioning and gene regulation: advances through genomics**. *Nat Rev Genet* 2009, **10**:161-172.

65. Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, Zeevi D, Sharon E, Weinberger A, Segal E: **Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast**. *Nat Genet* 2012, **44**:743-750.

66. •• Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ: **The role of nucleosome positioning in the evolution of gene regulation**. *PLoS Biol* 2010, **8**:e1000414.
The authors have studied the evolutionary changes in chromatin packing in 12 yeast species and found that it is linked to gene expression.

67. Yella VR, Bansal M: **DNA structural features and architecture of promoter regions play a role in gene responsiveness of *S. cerevisiae***. *J Bioinform Comput Biol* 2013, **11**:1343001.

68. Duran E, Djebali S, Gonzalez S, Flores O, Mercader JM, Guigo R, Torrents D, Soler-Lopez M, Orozco M: **Unravelling the hidden DNA structural/physical code provides novel insights on promoter location**. *Nucleic Acids Res* 2013, **41**:7220-7230.

69. Field Y, Fondufe-Mittendorf Y, Moore IK, Mieczkowski P, Kaplan N, Lubling Y, Lieb JD, Widom J, Segal E: **Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization**. *Nat Genet* 2009, **41**:438-445.

70. Rhee HS, Pugh BF: **Genome-wide structure and organization of eukaryotic pre-initiation complexes**. *Nature* 2012, **483**:295-301.

71. •• Struhl K, Segal E: **Determinants of nucleosome positioning**. *Nat Struct Mol Biol* 2013, **20**:267-273.
This review updates the knowledge of nucleosome positioning in yeast genome, emphasizing the role of DNA sequence and remodeling enzymes.

72. Lubliner S, Keren L, Segal E: **Sequence features of yeast and human core promoters that are predictive of maximal promoter activity**. *Nucleic Acids Res* 2013, **41**:5569-5581.

73. Berretta J, Morillon A: **Pervasive transcription constitutes a new level of eukaryotic genome regulation**. *EMBO Rep* 2009, **10**:973-982.

74. Jacquier A: **The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs**. *Nat Rev Genet* 2009, **10**:833-844.

75. SantaLucia J Jr: **A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics**. *Proc Natl Acad Sci U S A* 1998, **95**:1460-1465.

76. Bolshoy A, McNamara P, Harrington RE, Trifonov EN: **Curved DNA without A–A: experimental estimation of all 16 DNA**

wedge angles. *Proc Natl Acad Sci U S A* 1991, **88**:2312-2316.

77. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A *et al.*: **RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more**. *Nucleic Acids Res* 2013, **41**:D203-D213.

78. Kim D, Hong JS, Qiu Y, Nagarajan H, Seo JH, Cho BK, Tsai SF, Palsson BO: **Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling**. *PLoS Genet* 2012, **8**:e1002867.

79. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Munster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM: **Bidirectional promoters generate pervasive transcription in yeast**. *Nature* 2009, **457**:1033-1037.

80. Rice Annotation P, Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T *et al.*: **The Rice Annotation Project Database (RAP-DB): 2008 update**. *Nucleic Acids Res* 2008, **36**:D1028-D1033.

81. Yamashita R, Suzuki Y, Wakaguri H, Tsuritani K, Nakai K, Sugano S: **DBTSS: DataBase of Human Transcription Start Sites, progress report 2006**. *Nucleic Acids Res* 2006, **34**:D86-D89.