

# Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes

Aditi Kanhere and Manju Bansal\*

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

Received February 23, 2005; Revised April 21, 2005; Accepted May 13, 2005

## ABSTRACT

**During the process of transcription, RNA polymerase can exactly locate a promoter sequence in the complex maze of a genome. Several experimental studies and computational analyses have shown that the promoter sequences apparently possess some special properties, such as unusual DNA structures and low stability, which make them distinct from the rest of the genome. But most of these studies have been carried out on a particular set of promoter sequences or on promoter sequences from similar organisms. To examine whether the promoters from a wide variety of organisms share these special properties, we have carried out an analysis of sets of promoters from bacteria, vertebrates and plants. These promoters were analyzed with respect to the prediction of three different properties, such as DNA curvature, bendability and stability, which are relevant to transcription. All the promoter sequences are predicted to share certain features, such as stability and bendability profiles, but there are significant differences in DNA curvature profiles and nucleotide composition between the different organisms. These similarities and differences are correlated with some of the known facts about transcription process in the promoters from the three groups of organisms.**

## INTRODUCTION

The process of transcription begins with the RNA polymerase (RNAP) binding to DNA in the promoter region, which is in the immediate vicinity of the transcription start site (TSS). Exactly, how RNAP locates this specific binding site in the large excess of non-promoter DNA remains a field of intense investigation. A typical promoter sequence is thought to comprise some sequence motifs positioned at specific sites relative to the TSS. For example, a prokaryotic promoter is observed to have two hexameric motifs centered at or near  $-10$  and  $-35$

positions relative to the TSS (1). The structure of eukaryotic promoters is generally more complex and they have several different sequence motifs, such as TATA box, INR box, BRE, CCAAT-box and GC-box (2). These sequence motifs were identified based on the analysis of a large number of promoters and they represent consensus sequences. In other words, each nucleotide in the consensus sequence motif represents the most frequently occurring nucleotide at that position and does not represent an actual sequence. It has been observed that a wide variety of sequences similar to these representative motifs are present in promoters. In fact, there are very few promoter sequences that exactly match the consensus sequence, and also each of these sequence motifs is found in only a few of the promoter sequences. In addition, because these sequence motifs comprise only 6–10 bp and are degenerate, the probability of finding similar sequences in regions other than promoters is quite high. Hence, it is difficult to believe that these sequence motifs alone are wholly responsible for RNAP–promoter interaction. It is possible that base sequences in the neighborhood of these specific motifs may also be involved in the identification process, and it is highly likely that, in addition to the actual sequence itself, the second-order properties of the promoter sequence can also play a role in transcriptional regulation. Experimental evidences indeed suggest that sequence-dependent secondary properties of promoters are important in their function. Three such properties that are often involved are stability, curvature and bendability of DNA in these promoter regions.

An important step during transcription is the open complex formation between RNAP and promoter sequence, which involves local separation of the two strands around the  $-10$  region (3–8). The transcription process takes place under conditions in which DNA melting is a thermodynamically unfavorable process and yet during open complex formation the two strands separate without the help of any external energy. It is thought that the low stability of promoter region may assist in initial melting (9–12).

Another property, often associated with upstream sequences, is the occurrence of unusual DNA structures, such as curved DNA, which can be defined as a double-stranded DNA with a curved helical axis. A number of

\*To whom correspondence should be addressed. Tel: +91 80 2293 2534; Fax: +91 80 2360 0535; Email: mb@mbu.iisc.ernet.in

examples, from eukaryotes and prokaryotes, have shown that many genes have curved regions upstream of the TSS (13–28). The experimental data also indicate a correlation between promoter functioning and sequence-dependent DNA curvature (16–20). Many computational studies also predict the presence of curved DNA regions in the promoters (29–33).

Depending on its sequence, curvature can be an inherent property of a DNA molecule or it can be induced by external factors, such as protein binding. Some DNA sequences, while being intrinsically straight, can readily undergo distortion, and hence bendability of a DNA molecule can be defined as the ease with which the molecule can be made to curve in any direction. It is known that DNA bendability is important for binding of transcription factors, such as TBP (34) and CAP (35). Many other transcription factors also facilitate the adoption of curved conformations by DNA molecules (36,37). In addition, there is compelling experimental evidence, which suggests that promoter DNA wraps around the RNA polymerase (22,38,39). Hence, it is essential to have a better understanding of DNA bendability in promoter regions.

We have, therefore, analyzed sets of promoters from different organisms with respect to the above three properties, such as their predicted stability, curvature and bendability. Although promoters from different origins share certain features, such as stability and bendability profiles, we also see significant differences in their curvature and nucleotide composition. With the availability of a large number of genome sequences, the task of gene identification has assumed more significance. The characterization of these structural properties, in addition to sequence motifs, can greatly help in improving the currently available promoter and gene prediction algorithms (40,41).

## METHODS

### Promoter sequence sets

All the promoter sequences used in this study are 1000 nt long, starting from 500 nt upstream (position –500) and extending up to 500 nt downstream (position +500) of the TSS. In order to avoid having multiple TSSs in a given 1000 nt sequence, we have excluded all the TSSs that are <500 nt apart. Our promoter set has 227 *Escherichia coli* promoters, 89 *Bacillus subtilis* promoters, 252 vertebrate promoters and 74 plant promoters.

*E.coli* promoter sequences. The *E.coli* promoters were taken from the PromEC dataset (42), which provides a compilation of 471 experimentally identified transcriptional start sites. As mentioned above, after excluding all the TSSs that are <500 nt apart, the dataset contains 227 promoters. With the help of TSS information, promoter sequences were extracted from the *E.coli* genome sequence (NCBI accession no: NC\_000913).

*B.subtilis* promoter sequences. The TSSs for *B.subtilis* promoters were obtained from the DBTBS database (43). The required length sequences around TSSs were extracted from the *Bacillus* genome sequence (NCBI accession no: NC\_000964). The DBTBS dataset has 97 *Bacillus* promoters with experimentally identified start site. Out of the 97 *Bacillus* promoters, 89 promoters were selected after excluding all the TSSs that are <500 nt apart.

*Eukaryotic promoter sequences.* The vertebrate and plant POL II promoter sequences were extracted from Eukaryotic Promoter Database (EPD) (44,45). The EPD dataset has 2540 vertebrate promoters and 198 plant promoters. Only those promoters that have single initiation site and <50% sequence similarity in the region between –79 and +20 positions (designated as +S in the FP line of EPD entry) were selected in the first round of screening (669 vertebrate and 124 plant promoters). In the second round of screening, only those sequences that extend 500 bp upstream and 500 bp downstream of the TSS were retained. Finally, 252 vertebrate promoters and 74 plant promoters were used for this study.

*Shuffled sequences (control set).* Each 1000 bp sequence in the four datasets was divided with respect to TSS into 500 bp upstream and 500 bp downstream regions. Each region was shuffled separately, such that its mononucleotide composition was maintained. Thus, shuffled sequences have the same nucleotide composition as the actual promoters, but not the characteristic sequence patterns, if any. This procedure was repeated five times, to produce different shuffled sequences, corresponding to each upstream and downstream region. The stability, bendability and curvature calculations were carried out (as described below) on these shuffled sequences, and the mean values for the five shuffled sequences are used for comparison with the original genome sequences.

### Free energy calculation

The stability of a DNA molecule can be expressed in terms of free energy. The stability of DNA depends on mononucleotide composition as well as dinucleotide composition and it is possible to predict the stability of a DNA duplex from its sequence if one knows the contribution of each nearest-neighbor interaction (46–48). The standard free energy change ( $\Delta G_{37}^0$ ) corresponding to the melting transition of an ‘*n*’ nucleotide (or ‘*n* – 1’ dinucleotides) long DNA molecule, from double strand to single strand, is calculated as follows (46):

$$\Delta G^0 = -\left(\Delta G_{\text{ini}}^0 + \Delta G_{\text{sym}}^0\right) + \sum_{i=1}^{n-1} \Delta G_{i,i+1}^0$$

where,

$G_{\text{ini}}^0$  is the initiation free energy for dinucleotide of type *ij*.

$\Delta G_{\text{sym}}^0$  equals +0.43 kcal/mol and is applicable if the duplex is self-complementary.

$\Delta G_{i,j}^0$  is the standard free energy change for the dinucleotide of type *ij*.

Because our analysis involves long continuous stretches of DNA molecules, in our calculation we did not consider the two terms,  $G_{\text{ini}}^0$  and  $\Delta G_{\text{sym}}^0$ , which are more relevant for short oligonucleotides. In the present calculation, each promoter sequence is divided into overlapping windows of 15 bp (or 14 dinucleotide steps) and for each window, the free energy is calculated as given in the above equation. The energy values corresponding to the 10 unique dinucleotide sequences are taken from the unified parameters proposed recently (47,48).

### Curvature prediction

All the curvature calculations on the promoter sequences studied in this analysis were carried out with the help of in-house

software NUCGEN (49). Our earlier analysis showed that a set of dinucleotide parameters (CS) based on crystal structure data of oligonucleotides (50,51) can correctly predict the curvature of synthetic and genomic DNA sequences. Hence, the CS parameters were used for the DNA structure generation. Additional analysis (A. Kanhere and M. Bansal, unpublished data) also showed that for a reliable curvature prediction, the window size should be at least 50 bp or larger. Hence, we chose a window size of 75 bp for all the curvature calculations. This not only allowed us to make a more reliable estimation of curvature, but also helped to reduce the noise. Thus, for a promoter sequence of length ' $n$ ' and with a window size ' $w$ ' = 75 bp, we obtained  $(n - w + 1)$  number of DNA fragments. The curvature of the predicted structure for each of these fragments was calculated in terms of (i) radius of curvature (LSC), (ii) ratio of maximum component ( $I_{\max}$ ) to minimum component ( $I_{\min}$ ) of moments of inertia ( $I_{\max}/I_{\min}$ ) and (iii) ratio of end-to-end distance ' $d$ ' to the contour length ' $l_{\max}$ ' along the path traced by the DNA molecule ( $d/l_{\max}$ ). Because similar trends were observed for all three parameters, only the parameter  $d/l_{\max}$  is discussed in detail.

#### DNase I and nucleosomal positioning preference of DNA sequences

Two different trinucleotide models, based on nucleosomal positioning preferences (52) and DNase I sensitivity (53), have been suggested for bendability predictions of DNA sequences. We followed the procedure used previously for the analysis of a set of human promoter sequences (54), whereby the bendability profiles are calculated by looking up the values of trinucleotide parameters corresponding to each consecutive overlapping trinucleotide in the sequence.

The purpose of this study is to analyze general characteristics of each set of promoter sequences. Hence, an average profile is obtained for each group of promoters, by taking the mean value at each position, over all the promoter sequences in any given group. For this purpose, all the sequences were aligned such that all the TSS are in identical position, one below the other, and no gaps were introduced in order to maximize the sequence similarity. The mean and standard errors were calculated by bootstrap method using 100 runs. The average properties were compared with the corresponding properties of shuffled sequences.

## RESULTS

The promoter sequence dataset, used in this study, comprises only experimentally proven TSS from different organisms, ranging from prokaryotes to eukaryotes. The promoter sequences of prokaryotes belong to two classes: those from *E.coli*, a well studied bacteria of gram-negative class and those from *B.subtilis*, a representative of gram-positive bacterial class. Eukaryotic promoters are also grouped into two classes, depending on whether they are from vertebrates or from plants. The choice of our dataset permits us to compare the properties of promoters from different classes of organisms and find out the similarity and differences among them. Another important feature of the present analysis is the comparison of the properties of promoter sequences with

calculations on the shuffled sequences, which have the same nucleotide composition as the actual promoters but lack their sequence patterns, if any. The properties of shuffled sequences thus provide a baseline for comparative analysis of the actual promoter sequences.

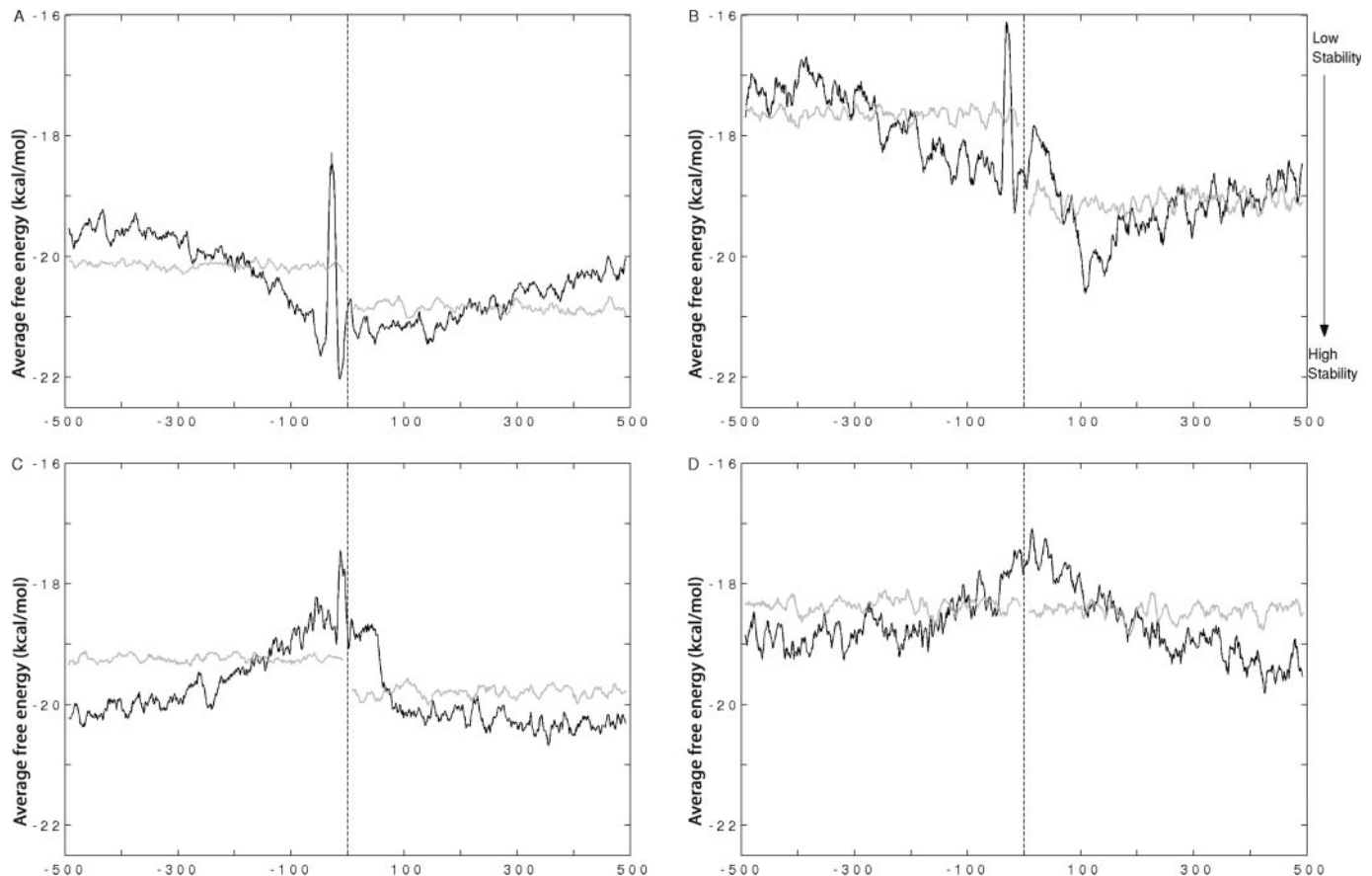
#### Promoter sequences are less stable than coding sequences

It is well known that DNA stability depends primarily on the sum of the interactions between the constituent dinucleotides (46). The mean stability profiles of different groups of promoter sequences were calculated based on this principle and are shown in Figure 1A–D. The most striking feature across all groups of promoters is the absence of any strong features in the shuffled sequences. Another prominent feature of the analysis (Figure 1A–D) is the difference in stabilities between the upstream and downstream regions. In all four groups of promoter sequences, the average stability of upstream region is predicted to be lower than the average stability of downstream region. The lower stability of upstream region probably arises owing to the higher AT content in this region (Table 1).

The bacterial promoter sequences (Figure 1C and D) show lowered stability around the  $-10$  region, while the eukaryotic promoter sequences (Figure 1A and B) show a peak lying between  $-25$  and  $-35$  region. The slight shift in the peak in eukaryotic promoter sequences as compared with prokaryotic sequences also suggests that the peak corresponds to the  $-10$  promoter element in bacteria and to the TATA box (at  $-30$  position) in the eukaryotic promoter sequences. This peak vanishes in the case of the TATA-less promoters in plants (data not shown) as well as in the case of shuffled sequences, thus confirming that the peak is owing to the characteristic TATA box sequence in this region. Similar stability calculation on *E.coli* promoter sequences, using a slightly smaller window size and different free energy parameters, had also reported a low stability peak around  $-10$  region (11). Our analysis on a diverse set of promoters confirms the universal nature of this characteristic peak.

#### Curvature prediction for promoter sequences

It is found that even in the absence of any external force, some DNA molecules can adopt a stable curved structure. Presence of such intrinsic curved DNA, upstream of promoter sequences, has been shown experimentally for eukaryotic and prokaryotic systems (13–28). To examine whether the presence of such altered DNA structure can be predicted from the promoter sequences, we obtained curvature profiles for each group of promoters. The  $d/l_{\max}$  profiles for all the four groups of promoters are shown in Figure 2A–D. Pronounced curvature is predicted for DNA regions in the vicinity of TSS of both the sets of bacterial promoter sequences (Figure 2C and D). An additional curved region around  $-300$  position is predicted for the *B.subtilis* promoters. In the same region, i.e. around  $-300$  position, a curved region is predicted in both plant and vertebrate promoters (Figure 2A and B). The magnitude of the mean curvature predicted (around  $-300$  position) for the vertebrate promoters is much smaller when compared with bacterial and plant promoters.



**Figure 1.** Distribution of free energy of duplex formation, near the TSSs. The figure shows the average free energy profiles (black) with respect to the relative base position ( $x$ -axis), in the case of (A) vertebrate (B) plant (C) *E.coli* and (D) *B.subtilis* promoters. More negative values indicate greater stability (indicated by black arrow on the top right hand corner of the figure). The profiles in this, and in subsequent Figures 2–4, extend from 500 nt upstream to 500 nt downstream of TSS (shown as dashed vertical line at 0 position). The profiles calculated for the shuffled sequences in the upstream and downstream regions are shown (gray) in each case.

**Table 1.** The average frequency of mononucleotides A+T in different groups of promoter sequences

Group of promoters	Complete genome	Present dataset			
		Upstream (–500 to TSS)	Downstream (TSS to +500)	Upstream (–150 to –50)	Downstream (100 to 200)
Vertebrate	—	0.47 (0.11)	0.43 (0.11)	0.44 (0.14)	0.40 (0.13)
Plant	—	0.63 (0.09)	0.54 (0.12)	0.58 (0.13)	0.50 (0.13)
<i>E.coli</i>	0.49	0.53 (0.06)	0.49 (0.04)	0.56 (0.08)	0.49 (0.06)
<i>B.subtilis</i>	0.56	0.60 (0.04)	0.57 (0.04)	0.60 (0.08)	0.59 (0.06)

The standard deviation values are given in parenthesis.

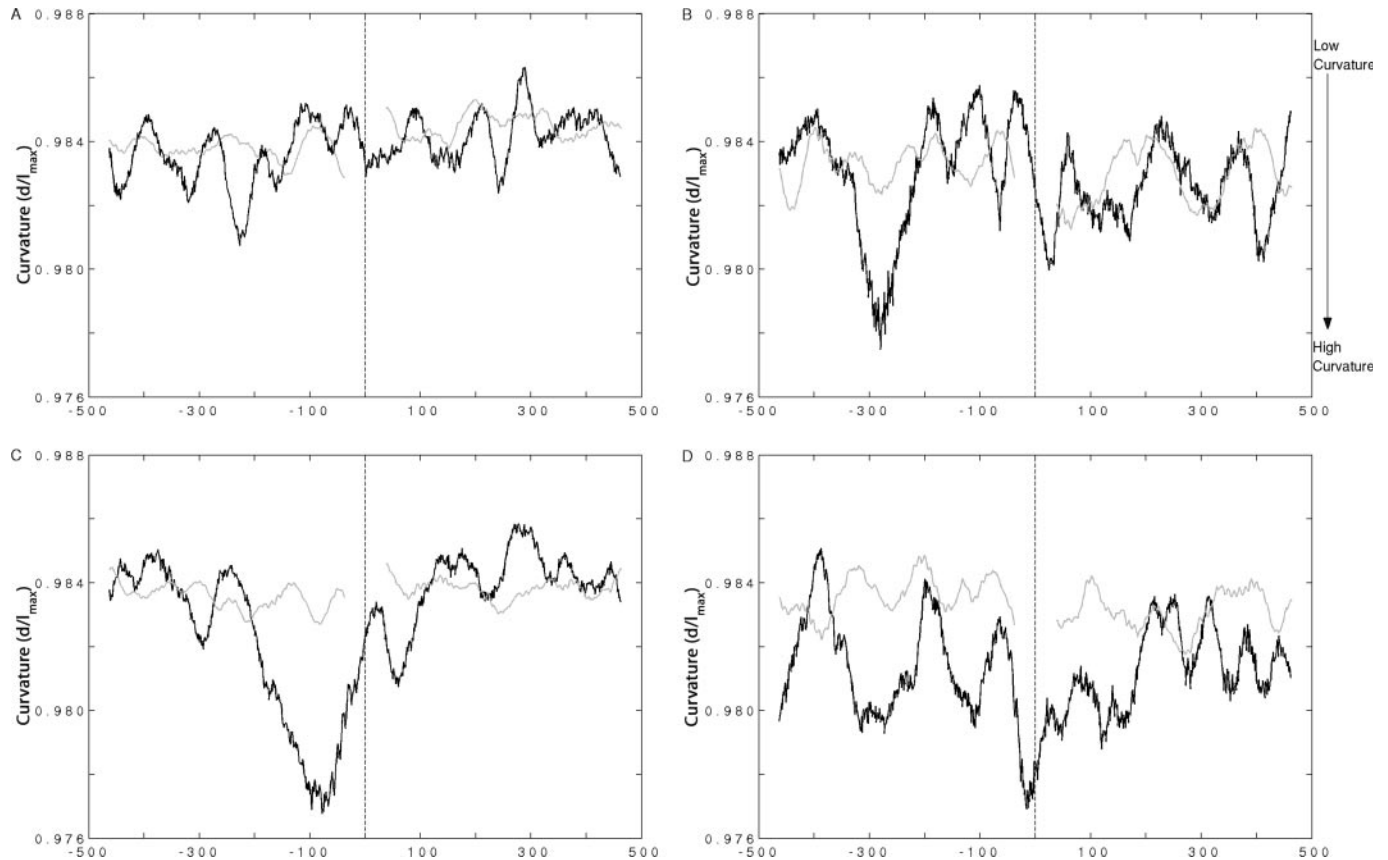
### Bendability prediction for promoter sequences

As mentioned above, bendability plays an important role in gene expression. Hence we predicted bendability profiles for different classes of promoters. The bendability profiles obtained using the nucleosomal positioning preference (Figure 3A–D) and DNase I based-bendability measure (Figure 4A–D) match well. Observation of the two profiles, for all four groups of promoters (Figure 3 and 4), reveals a common pattern in the bendability of the upstream and downstream region, i.e. upstream promoter region is predicted to be less bendable than the downstream coding region. Although a similar characteristic bendability pattern is seen for all four types of promoters, in the case of vertebrate promoters the difference is less

prominent as compared with that in the other three groups of promoters. The difference between the predicted bendability of the shuffled and genomic sequences is also less significant in the case of vertebrate promoters.

### Compositional analysis of promoter sequences

The characteristic differences observed in the upstream and downstream regions can be a consequence of base composition. Hence, we compared the composition of upstream and downstream regions of promoter sequences in terms of the mononucleotide, dinucleotide and trinucleotide frequencies. For this calculation, we considered 100 nt fragments in the upstream region (–150 to –50 position) and downstream region (+100 to +200 position). These regions were selected



**Figure 2.** Distribution of curvature around TSSs. The figure shows the average predicted curvature ( $d/l_{\max}$ ) profile (black) against the relative base position ( $x$ -axis), in the case of (A) vertebrate (B) plant (C) *E.coli* and (D) *B.subtilis* promoters. Smaller values indicate higher curvature (indicated by black arrow on the top right hand corner of the figure).

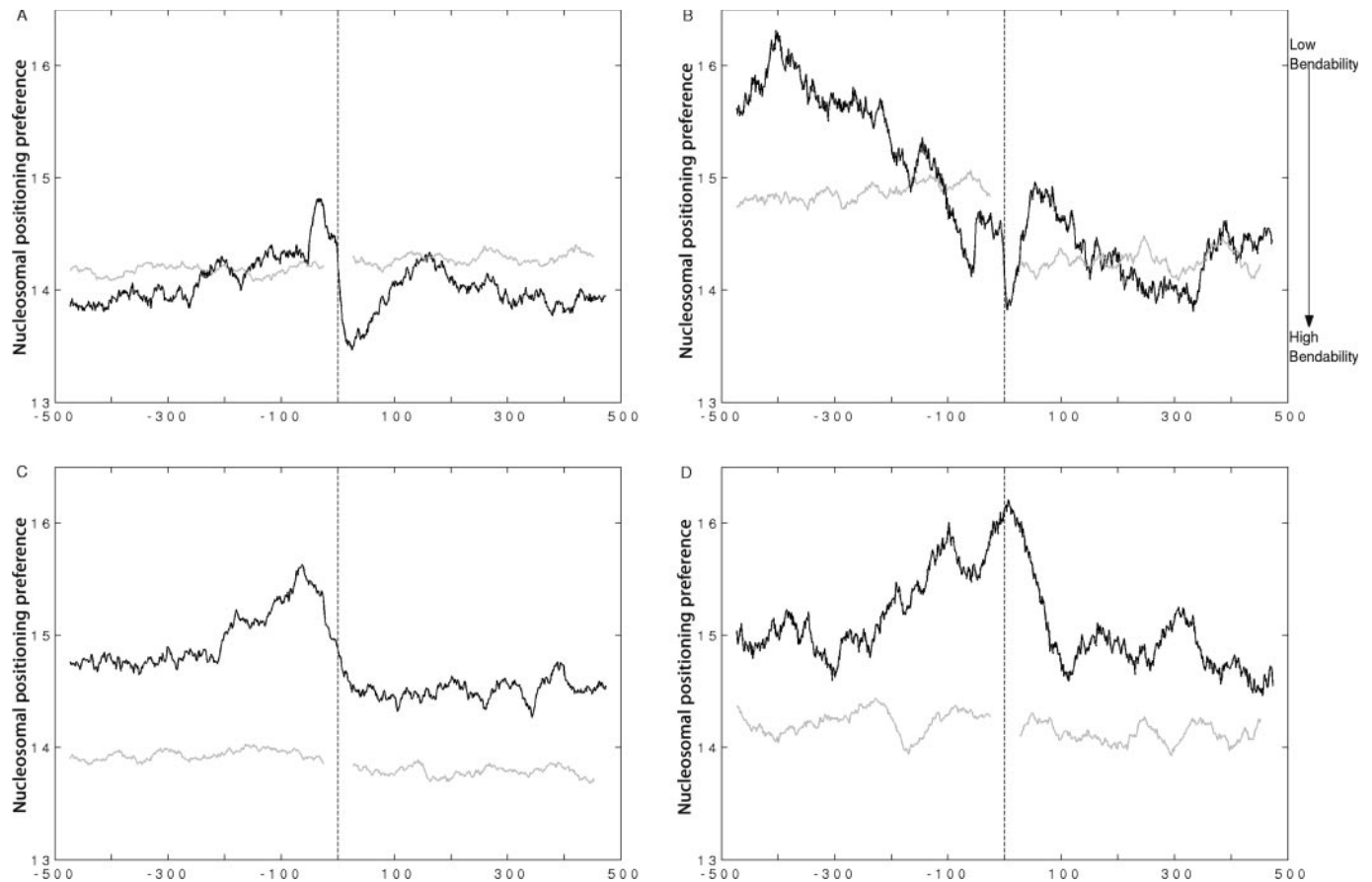
because they are the best representatives of upstream and downstream regions. We avoided the middle region encompassing the low stability peak corresponding to the TATA box, known to have a composition biased toward high T+A content. All the groups of promoters show a high occurrence of AA, AT, TA and TT dinucleotides in the upstream regions as compared with the downstream region, while the dinucleotides GC, GG, CG, CC, AG, GA and TG are over-represented in the downstream region as compared with the upstream region (Figure 5). Interestingly, the trinucleotides AAA, TTT, AAT, ATT, ATA, TTA, TAA, TTC and TCA containing the above identified dinucleotides are generally over-represented in the upstream region, while the trinucleotides CCG, CGG, GCC, GGC, AGC, GAG, CAG, GTG, TGC, TGG, CTG and GCT are over-represented in the downstream region (Figure 6). This is also reflected in the mononucleotide composition of all the promoters, the upstream region being more A+T-rich than the downstream region (Table 1).

The calculation of dinucleotide and trinucleotide frequencies along the sequence length also shows sharp transitions near TSS. Though compositional differences are seen in the upstream and downstream region of all the four groups of promoters, the magnitude of the difference varies between the different groups. The plant promoters and *E.coli* promoters (Figure 5B and C) are quite distinct in showing very large differences in the dinucleotide composition, between upstream

and downstream region as compared with the other two promoter groups. The Bacillus and plant sequences are AT-rich while vertebrate sequences are GC-rich (Table 1). The dinucleotide frequencies of the vertebrate and Bacillus sequences (Figure 5A and D) span a wide range, extending from the lower ( $\sim 3\%$ ) to the higher ( $\sim 12\%$ ) end of the scale, whereas those for plant and *E.coli* sequences (Figure 5B and C and Figure 6B and C) are clustered in the middle region of the scale ( $\sim 4\text{--}9\%$ ). In the case of trinucleotide frequencies, only the Bacillus sequences span a wide range ( $\sim 0.5\text{--}5\%$ ) while for the other three classes they are clustered in the middle region ( $\sim 0.5\text{--}3.5\%$ ).

## DISCUSSION

We have compared various structural properties as predicted for the promoter sequences from organisms belonging to three different kingdoms, such as bacteria, animal and plant. The study indicates that there are certain properties, which may be shared by promoters, independent of the organism that they belong to or the gene that they control. In general, the promoter regions are less stable and less bendable but contain DNA elements with enhanced curvature, when compared with the downstream coding regions. However, there are also striking differences between the structural profiles of prokaryotic and



**Figure 3.** Bendability distribution around TSSs calculated using trinucleotide parameters based on nucleosomal positioning preferences. The figure shows the bendability profiles (black) with respect to the relative base position ( $x$ -axis), in the case of (A) vertebrate (B) plant (C) *E.coli* and (D) *B.subtilis* promoters. For the sake of clarity, the profiles are smoothed using a 50 nt window. Smaller values indicate greater bendability (indicated by black arrow on the top right hand corner of the figure).

eukaryotic promoters. Here, we discuss the possible role of these observations and their implications in the process of transcription.

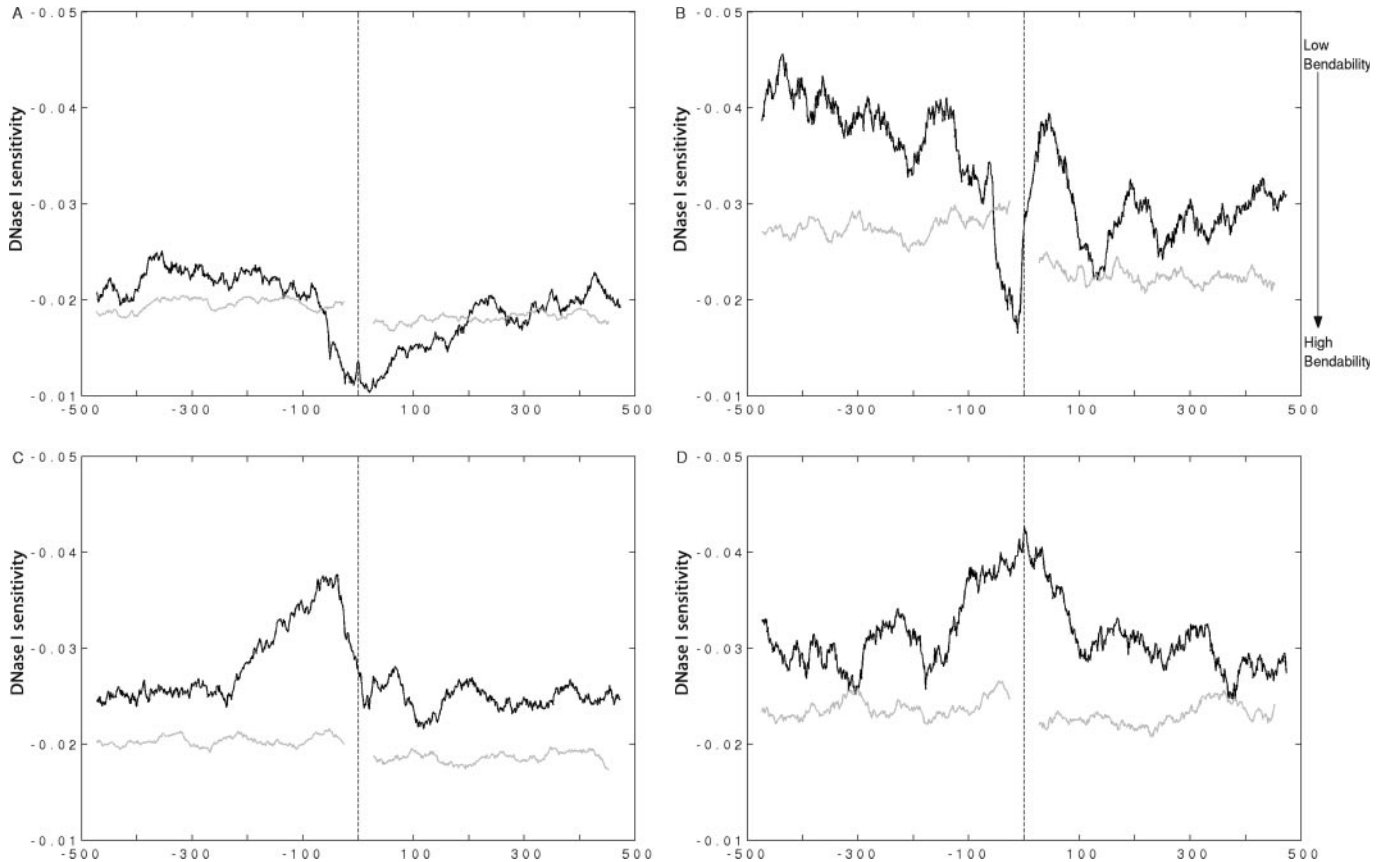
#### Low stability of promoter regions as compared with the non-promoter regions

The RNA polymerase movement during transcription leads to the induction of positive supercoils ahead and negative supercoils behind, leading to torsional stresses. Opening of any base pair under this stress changes the denaturing probability of every other base pair. It has been reported earlier that the susceptibility to duplex destabilization induced by superhelical stress is closely associated with the boundaries of genes and transcription regulatory sites (55). The low stability predicted for promoter regions as compared with the non-promoter region can explain the stress-induced promoter-specific opening of the DNA. It is interesting that the feature of lower stability of upstream region is common to all promoters, independent of their overall mononucleotide composition (Table 1). In plants and *E.coli* promoters, the difference in the dinucleotide composition (Figure 5B and C) is more prominent and this is reflected in the greater difference in stability between the upstream and downstream region.

#### Possible roles of DNA curvature and bendability

The presence of sequence-dependent DNA curvature in promoter region, independent of any external factors such as proteins, has been experimentally observed in many cases [reviewed in (19,21,24–25)]; furthermore, transcriptional regulation by curved DNA stretches has been demonstrated in a number of cases (17,18,56–58). Our analysis clearly shows that a significant number of promoters in all the groups may have curved DNA elements upstream of the TSSs, thus facilitating transcription. The difference in the location of predicted curved DNA from one group of promoter to another correlates with differences in their transcription regulation (as discussed below).

In addition to the sequence-dependent intrinsic curvature, DNA bendability also plays an important role in transcription. Based on various experimental studies, it has been suggested that during transcription initiation, the promoter DNA of length  $\sim 300$  Å wraps around the polymerase (22,39). It has also been proposed that the energy cost of DNA bending may play a role in modulating the open complex formation, as well as in facilitating promoter clearance and that, without this energy cost, the energy of RNAP–DNA complex would probably be too high to permit the escape of the polymerase from the promoter. In this context, our observation of the distinct



**Figure 4.** Bendability distribution in the vicinity of TSSs calculated using DNase I sensitivity parameters. The figure shows the bendability profiles (black) with respect to the relative base position (x-axis), in the case of (A) vertebrate (B) plant (C) *E.coli* and (D) *B.subtilis* promoters. For the sake of clarity, the profiles are smoothed using a 50 nt window. Less negative values indicate higher bendability (indicated by black arrow on the top right hand corner of the figure).

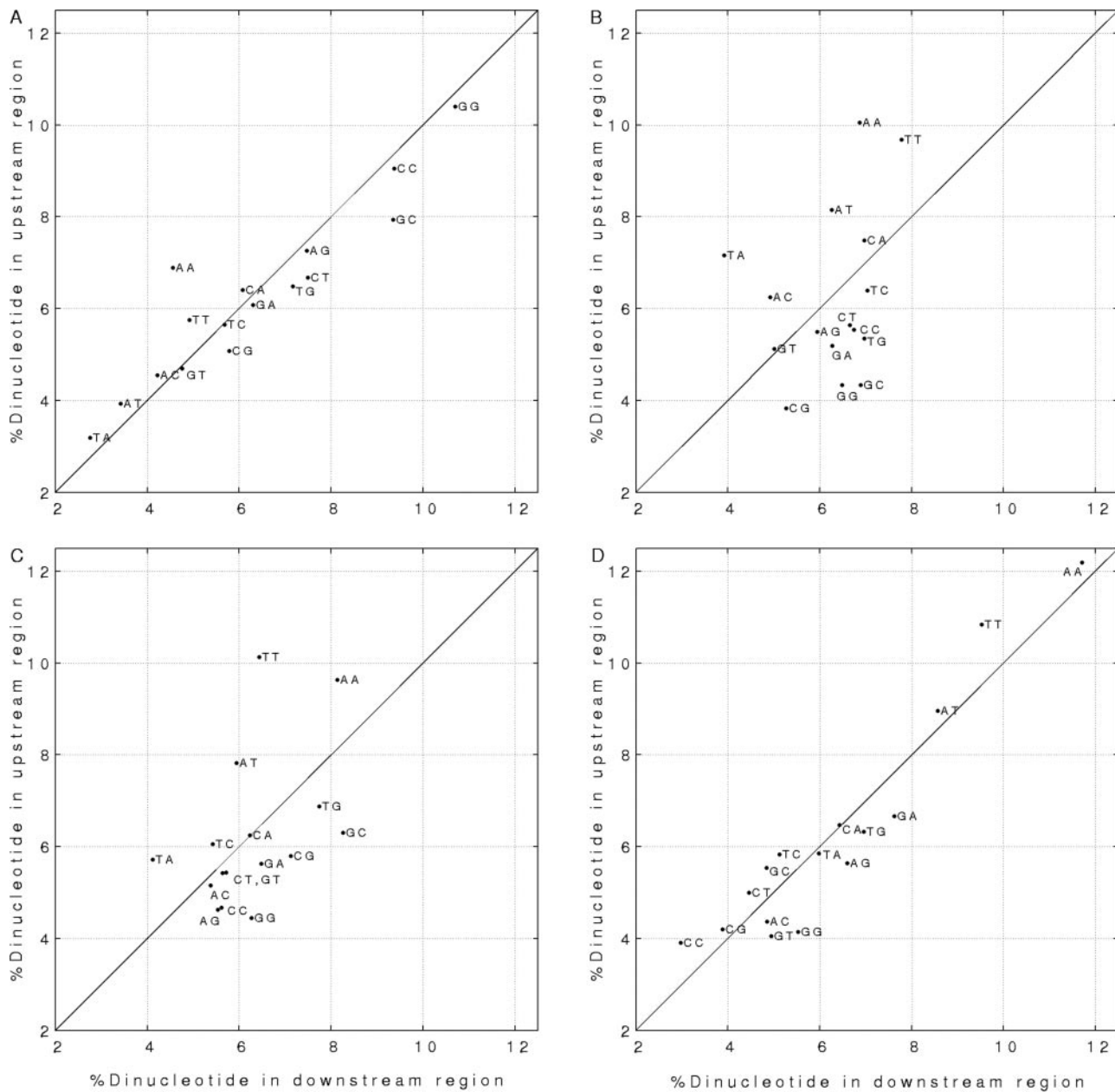
presence of a low bendability region in the proximity of the TSS is significant.

Another explanation for the typical bendability profile (Figures 3 and 4), with upstream region having lower bendability than the downstream region, was given by Pedersen *et al.* (54) from their analysis of a set of human promoters. They suggested that the characteristic bendability pattern in the promoter sequences is possibly connected with the formation of nucleosomes. Because nucleosomes have a preference for more flexible DNA (59,60), any elements that destabilize nucleosomes can activate transcription by facilitating access to transcription factors (61,62). This fact is supported by the observation of low bendability in the upstream region and high bendability in the downstream region of eukaryotic promoters (Figure 3A and B and Figure 4A and B). In bacterial genomes, proteins, such as H-NS (63), are analogous in function to the histones and the HMG box proteins of eukaryotes and this may explain the low bendability in the upstream region in bacteria, even though the genome is not organized as nucleosomes. Another interesting point about the bendability profiles is that although the two trinucleotide bending parameters are not highly correlated (64), the bendability profiles of the promoters, derived using the two parameters, show very similar features, suggesting that the nucleotide composition in promoters is such that some characteristic properties, such as bendability, are conserved.

### Comparison between the four groups of promoters

Although the overall function seems to be conserved across the different groups of promoters, they do differ in finer details. In eukaryotes, the DNA is packed into nucleosomes, which blocks the recognition of the core promoters by the basic transcription machinery (65–68). In comparison, the prokaryotic DNA is essentially naked, i.e. the RNA polymerase is not greatly hindered in its ability to gain access to the DNA and initiate RNA synthesis (65). DNA flexibility is also known to play a role in nucleosome formation, and perhaps overall higher flexibility of downstream regions in eukaryotic promoters (matching that of shuffled DNA) is important in this regard. In contrast, the prokaryotic promoters, where the DNA is not packaged into nucleosomes, are overall more rigid than the shuffled sequences.

Another noticeable feature of eukaryotic promoters is the presence of regulatory sites hundreds of base pairs upstream from TSS, while the regulatory elements in bacterial promoters tend to be located in the vicinity of the TSS. Our analysis also indicates that the special upstream features seem to extend at least up to  $-500$  position in the case of eukaryotic promoter sequences (Figures 1–4 A, B), but seem to be confined up to  $-300$  position in the case of prokaryotic promoters (Figures 1–4 C, D). The observation that in eukaryotes, transcription factors can bind hundreds of base pair upstream seems to be reflected in the position of the predicted



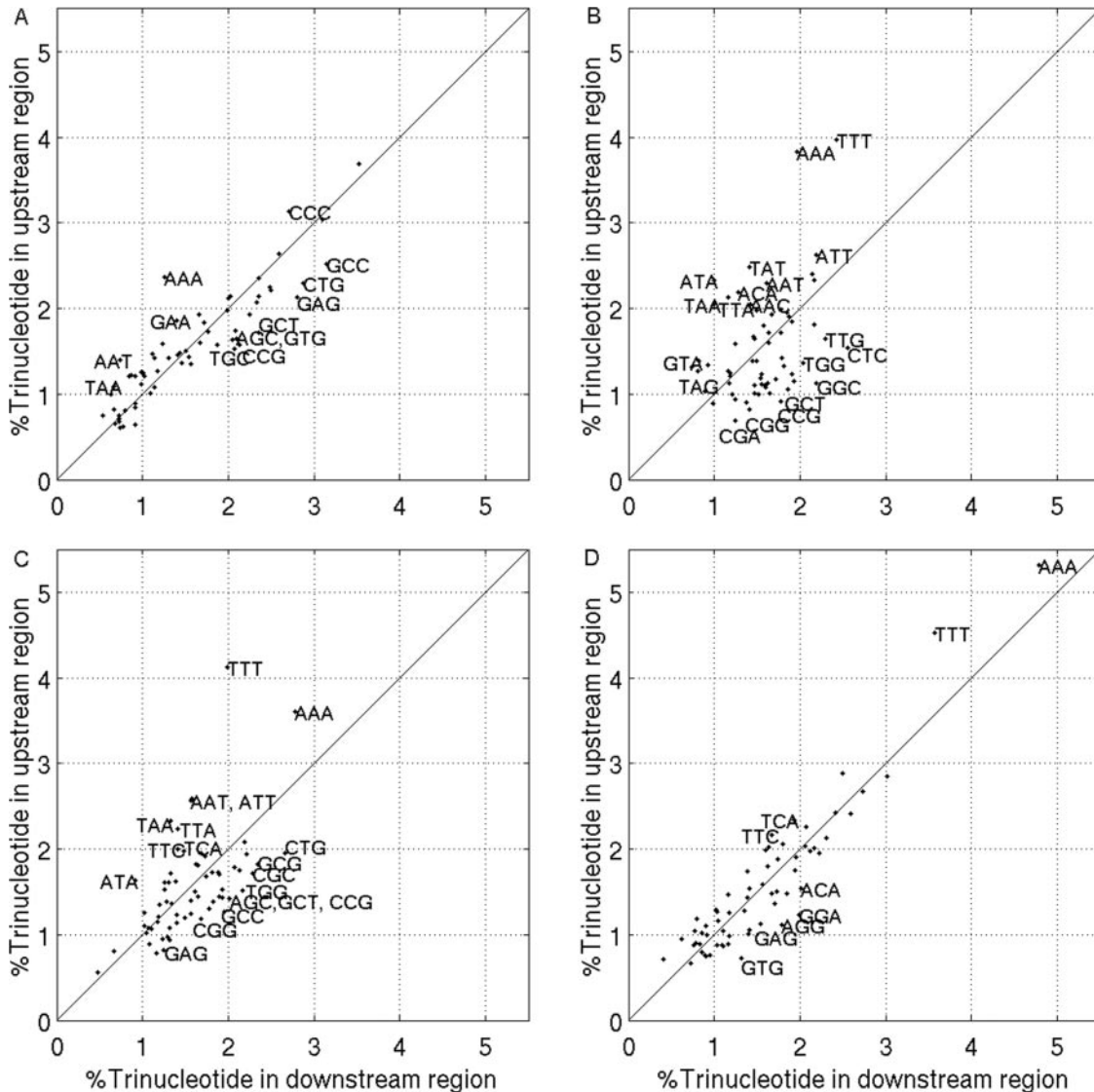
**Figure 5.** The percentage occurrence of each dinucleotide in upstream (y-axis) versus downstream (x-axis) region in the near vicinity of TSS, i.e.  $-150$  to  $-50$  and  $100$  to  $200$  in the case of (A) vertebrate, (B) plant, (C) *E. coli* (D) *B. subtilis*. The dinucleotides, which are present more often in downstream region than in the upstream region, appear below the diagonal and *vice versa*.

curved region. Both groups of eukaryotic promoters show the presence of a curved region considerably upstream of the TSS ( $>-200$  bp); however, the bacterial promoters show the presence of a curved region nearer to the TSS.

The observed differences between the promoters from the two bacterial origins may be attributed to the differences in their mode of binding to their respective RNA polymerases (69–71). On the other hand, differences between the two eukaryotic promoter sequences, such as vertebrate and plant, may be a consequence of basic differences in their transcription regulation mechanisms, as well as due to their distinct composition (with plants being overall AT-rich and more so in the upstream region). The recent releases of plant genome sequences have revealed that plants have much higher gene

density when compared with animal genomes; however, the increase in gene density comes at the cost of greater logistic problem in transcriptional regulation of the genes (72). One solution to this problem would be to have a larger number of regulatory proteins and plant genomes are in fact known to have a very high percentage of genes coding for transcription factors (73). We would like to suggest that the sharp delineation in the various properties, such as stability, bendability, etc., of intergenic and coding region of plant genomes may be one more way of identifying the transcriptional regulatory regions. In line with this hypothesis, it is to be noted that the vertebrate class has an average gene density much smaller than the members of other three groups (Gene densities: Human  $\sim 1/100\,000$ , *Arabidopsis*  $\sim 1/4000$ ,





**Figure 6.** The percentage occurrence of trinucleotides in upstream (y-axis) versus downstream (x-axis) region in the near vicinity of TSS, i.e.  $-150$  to  $-50$  and  $100$  to  $200$  in the case of (A) vertebrate, (B) plant, (C) *E. coli* (D) *B. subtilis*. The trinucleotides, which are found more often in the downstream region than in the upstream region, appear below the diagonal and *vice versa*. For the sake of clarity, only some trinucleotides, which show significantly large differences in their upstream and downstream frequencies, are labeled.

*E. coli*  $\sim 1/1000$  and *B. subtilis*  $\sim 1/1000$ ), and accordingly vertebrate promoters do not show large differences in their upstream and downstream region, as compared with the other three groups of promoters.

## CONCLUSIONS

Promoter regions in prokaryotic and eukaryotic genomes are predicted to have several common structural features, such as lower stability, higher curvature and lesser bendability as compared with their neighboring regions. All the four groups of promoters considered here are also distinctly different from non-promoter regions in their mononucleotide, dinucleotide and trinucleotide composition. However, there are also some important differences among the various groups of promoters. In the case of prokaryotic sequences, the distinct structural features are confined to relatively short upstream region as

compared with eukaryotic sequences, where they seem to extend over a significantly larger upstream region. In addition, the prokaryotic sequences are predicted to be overall less bendable when compared with eukaryotic promoters. The differences in prokaryotic and eukaryotic promoter sequences match well with their distinct patterns of transcription regulation. We have also observed some distinct features in the two prokaryotic promoter sets as well as in the eukaryotic promoter sets. In general, these similarities and differences between the promoters can provide a rationale for some of the known facts about the transcription process in the various organisms.

## ACKNOWLEDGEMENTS

This work of supported by Department of Biotechnology, India. During the study, A.K. was supported by University Grants Commission and Council of Scientific and Industrial

Research, India. The Open Access publication charges for this article were waived by Oxford University Press.

*Conflict of interest statement.* None declared.

## REFERENCES

- Harley, C.B. and Reynolds, R.P. (1987) Analysis of *E. coli* promoter sequences. *Nucleic Acids Res.*, **15**, 2343–2361.
- Bucher, P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.
- Buckle, M. and Buc, H. (1989) Fine mapping of DNA single-stranded regions using base-specific chemical probes: study of an open complex formed between RNA polymerase and the lac UV5 promoter. *Biochemistry*, **28**, 4388–4396.
- Chen, Y.F. and Helmann, J.D. (1997) DNA-melting at the *Bacillus subtilis* flagellin promoter nucleates near –10 and expands unidirectionally. *J. Mol. Biol.*, **267**, 47–59.
- Craig, M.L., Suh, W.C. and Record, M.T., Jr (1995) HO. and DNase I probing of E sigma 70 RNA polymerase–lambda PR promoter open complexes: Mg<sup>2+</sup> binding and its structural consequences at the transcription start site. *Biochemistry*, **34**, 15624–15632.
- Sasse-Dwight, S. and Gralla, J.D. (1989) KMnO<sub>4</sub> as a probe for lac promoter DNA melting and mechanism *in vivo*. *J. Biol. Chem.*, **264**, 8074–8081.
- Siebenlist, U., Simpson, R.B. and Gilbert, W. (1980) *E. coli* RNA polymerase interacts homologously with two different promoters. *Cell*, **20**, 269–281.
- Suh, W.C., Ross, W. and Record, M.T., Jr (1993) Two open complexes and a requirement for Mg<sup>2+</sup> to open the lambda PR transcription start site. *Science*, **259**, 358–361.
- Nakata, K., Kanehisa, M. and Maizel, J.V., Jr (1988) Discriminant analysis of promoter regions in *Escherichia coli* sequences. *Comput. Appl. Biosci.*, **4**, 367–371.
- Vollenweider, H.J., Fiandt, M. and Szybalski, W. (1979) A relationship between DNA helix stability and recognition sites for RNA polymerase. *Science*, **205**, 508–511.
- Margalit, H., Shapiro, B.A., Nussinov, R., Owens, J. and Jernigan, R.L. (1988) Helix stability in prokaryotic promoter regions. *Biochemistry*, **27**, 5179–5188.
- Pedersen, A.G., Jensen, L.J., Brunak, S., Staerfeldt, H.H. and Ussery, D.W. (2000) A DNA structural atlas for *Escherichia coli*. *J. Mol. Biol.*, **299**, 907–930.
- Bossi, L. and Smith, D.M. (1984) Conformational change in the DNA associated with an unusual promoter mutation in a tRNA operon of *Salmonella*. *Cell*, **39**, 643–652.
- Galas, D.J., Eggert, M. and Waterman, M.S. (1985) Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from *Escherichia coli*. *J. Mol. Biol.*, **186**, 117–128.
- Hsu, L.M., Giannini, J.K., Leung, T.W. and Crosthwaite, J.C. (1991) Upstream sequence activation of *Escherichia coli* argT promoter *in vivo* and *in vitro*. *Biochemistry*, **30**, 813–822.
- Kuhnke, G., Fritz, H.J. and Ehrling, R. (1987) Unusual properties of promoter-up mutations in the *Escherichia coli* galactose operon and evidence suggesting RNA polymerase-induced DNA bending. *EMBO J.*, **6**, 507–513.
- Lamond, A.I. and Travers, A.A. (1983) Requirement for an upstream element for optimal transcription of a bacterial tRNA gene. *Nature*, **305**, 248–250.
- McAllister, C.F. and Achberger, E.C. (1988) Effect of polyadenine-containing curved DNA on promoter utilization in *Bacillus subtilis*. *J. Biol. Chem.*, **263**, 11743–11749.
- Perez-Martin, J., Rojo, F. and de Lorenzo, V. (1994) Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. *Microbiol. Rev.*, **58**, 268–290.
- Plaskon, R.R. and Wartell, R.M. (1987) Sequence distributions associated with DNA curvature are found upstream of strong *E. coli* promoters. *Nucleic Acids Res.*, **15**, 785–796.
- Ohyama, T. (2001) Intrinsic DNA bends: an organizer of local chromatin structure for transcription. *Bioessays*, **23**, 708–715.
- Rees, W.A., Keller, R.W., Vesenka, J.P., Yang, G. and Bustamante, C. (1993) Evidence of DNA bending in transcription complexes imaged by scanning force microscopy. *Science*, **260**, 1646–1649.
- Tanaka, K., Muramatsu, S., Yamada, H. and Mizuno, T. (1991) Systematic characterization of curved DNA segments randomly cloned from *Escherichia coli* and their functional significance. *Mol. Gen. Genet.*, **226**, 367–376.
- Hagerman, P.J. (1990) Sequence-directed curvature of DNA. *Annu. Rev. Biochem.*, **59**, 755–781.
- Kanhere, A. and Bansal, M. (2004) DNA bending and curvature: A ‘turning’ point in DNA function? *PINSA*, **B70**, 239–255.
- Prosseda, G., Falconi, M., Giangrossi, M., Gualerzi, C.O., Micheli, G. and Colonna, B. (2004) The virF promoter in *Shigella*: more than just a curved DNA stretch. *Mol. Microbiol.*, **51**, 523–537.
- Kaji, M., Matsushita, O., Tamai, E., Miyata, S., Taniguchi, Y., Shimamoto, S., Katayama, S., Morita, S. and Okabe, A. (2003) A novel type of DNA curvature present in a *Clostridium perfringens* ferredoxin gene: characterization and role in gene expression. *Microbiology*, **149**, 3083–3091.
- Agrawal, G.K., Asayama, M. and Shirai, M. (2003) Two distinct curved DNAs upstream of the light-responsive psbA gene in a cyanobacterium. *Biosci. Biotechnol. Biochem.*, **67**, 1817–1821.
- Kozobay-Avraham, L., Hosid, S. and Bolshoy, A. (2004) Curvature distribution in prokaryotic genomes. *In Silico Biol.*, **4**, 29.
- Jauregui, R., Abreu-Goodger, C., Moreno-Hagelsieb, G., Collado-Vides, J. and Merino, E. (2003) Conservation of DNA curvature signals in regulatory regions of prokaryotic genes. *Nucleic Acids Res.*, **31**, 6770–6777.
- Kalate, R.N., Kulkarni, B.D. and Nagaraja, V. (2002) Analysis of DNA curvature distribution in mycobacterial promoters using theoretical models. *Biophys. Chem.*, **99**, 77–97.
- Gabrielian, A.E., Landsman, D. and Bolshoy, A. (2000) Curved DNA in promoter sequences. *In Silico Biol.*, **1**, 183–196.
- Tosato, V., Gjuracic, K., Vlahovick, K., Pongor, S., Danchin, A. and Bruschi, C.V. (2003) The DNA secondary structure of the *Bacillus subtilis* genome. *FEMS Microbiol. Lett.*, **218**, 23–30.
- Nikolov, D.B., Chen, H., Halay, E.D., Hoffman, A., Roeder, R.G. and Burley, S.K. (1996) Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc. Natl. Acad. Sci. USA*, **93**, 4862–4867.
- Schultz, S.C., Shields, G.C. and Steitz, T.A. (1991) Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science*, **253**, 1001–1007.
- Nagaich, A.K., Appella, E. and Harrington, R.E. (1997) DNA bending is essential for the site-specific recognition of DNA response elements by the DNA binding domain of the tumor suppressor protein p53. *J. Biol. Chem.*, **272**, 14842–14849.
- Konig, P. and Richmond, T.J. (1993) The X-ray structure of the GCN4-bZIP bound to ATF/CREB site DNA shows the complex depends on DNA flexibility. *J. Mol. Biol.*, **233**, 139–154.
- Rivetti, C., Guthold, M. and Bustamante, C. (1999) Wrapping of DNA around the *E. coli* RNA polymerase open promoter complex. *EMBO J.*, **18**, 4464–4475.
- Cheatham, G.M., Jeruzalmi, D. and Steitz, T.A. (1999) Structural basis for initiation of transcription from an RNA polymerase–promoter complex. *Nature*, **399**, 80–83.
- Ohler, U., Niemann, H., Liao, G. and Rubin, G.M. (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, **17**, S199–S206.
- Kanhere, A. and Bansal, M. (2005) A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinformatics*, **6**, 1.
- Hershberg, R., Bejerano, G., Santos-Zavaleta, A. and Margalit, H. (2001) PromEC: an updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. *Nucleic Acids Res.*, **29**, 277.
- Makita, Y., Nakao, M., Ogasawara, N. and Nakai, K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res.*, **32**, D75–D77.
- Praz, V., Perier, R., Bonnard, C. and Bucher, P. (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res.*, **30**, 322–324.
- Schmid, C.D., Praz, V., Delorenzi, M., Perier, R. and Bucher, P. (2004) The Eukaryotic Promoter Database EPD: the impact of *in silico* primer extension. *Nucleic Acids Res.*, **32**, D82–D85.

46. Breslauer, K.J., Frank, R., Blocker, H. and Marky, L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.
47. SantaLucia, J., Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
48. Allawi, H.T. and SantaLucia, J., Jr (1997) Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry*, **36**, 10581–10594.
49. Bansal, M., Bhattacharyya, D. and Ravi, B. (1995) NUPARM and NUCGEN: software for analysis and generation of sequence dependent nucleic acid structures. *Comput. Appl. Biosci.*, **11**, 281–287.
50. Kanhere, A. and Bansal, M. (2003) An assessment of three dinucleotide parameters to predict DNA curvature by quantitative comparison with experimental data. *Nucleic Acids Res.*, **31**, 2647–2658.
51. Bansal, M. (1996) Structural variation observed in DNA crystal structures and their implications for protein–DNA interaction. In Sarma, R.H. and Sarma, M.H. (eds), *Biological Structure and Dynamics, Proceedings of ninth convention.*, Adenine press, New York, NY, Vol. 1, pp. 121–134.
52. Satchwell, S.C., Drew, H.R. and Travers, A.A. (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, **191**, 659–675.
53. Brukner, I., Sanchez, R., Suck, D. and Pongor, S. (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.*, **14**, 1812–1818.
54. Pedersen, A.G., Baldi, P., Chauvin, Y. and Brunak, S. (1998) DNA structure in human RNA polymerase II promoters. *J. Mol. Biol.*, **281**, 663–673.
55. Benham, C.J. (1996) Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions. *J. Mol. Biol.*, **255**, 425–434.
56. Brahms, G., Brahms, S. and Magasanik, B. (1995) A sequence-induced superhelical DNA segment serves as transcriptional enhancer. *J. Mol. Biol.*, **246**, 35–42.
57. Kim, J., Klooster, S. and Shapiro, D.J. (1995) Intrinsically bent DNA in a eukaryotic transcription factor recognition sequence potentiates transcription activation. *J. Biol. Chem.*, **270**, 1282–1288.
58. Ellinger, T., Behnke, D., Knaus, R., Bujard, H. and Gralla, J.D. (1994) Context-dependent effects of upstream A-tracts. Stimulation or inhibition of *Escherichia coli* promoter function. *J. Mol. Biol.*, **239**, 466–475.
59. Fitzgerald, D.J. and Anderson, J.N. (1998) Unique translational positioning of nucleosomes on synthetic DNAs. *Nucleic Acids Res.*, **26**, 2526–2535.
60. Widlund, H.R., Kuduvalli, P.N., Bengtsson, M., Cao, H., Tullius, T.D. and Kubista, M. (1999) Nucleosome structural features and intrinsic properties of the TATAAACGCC repeat sequence. *J. Biol. Chem.*, **274**, 31847–31852.
61. Zhu, Z. and Thiele, D.J. (1996) A specialized nucleosome modulates transcription factor access to a *C. glabrata* metal responsive promoter. *Cell*, **87**, 459–470.
62. Iyer, V. and Struhl, K. (1995) Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.*, **14**, 2570–2579.
63. Hommais, F., Krin, E., Laurent-Winter, C., Soutourina, O., Malpertuy, A., Le Caer, J.P., Danchin, A. and Bertin, P. (2001) Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, H-NS. *Mol. Microbiol.*, **40**, 20–36.
64. Brukner, I., Sanchez, R., Suck, D. and Pongor, S. (1995) Trinucleotide models for DNA bending propensity: comparison of models based on DNase I digestion and nucleosome packaging data. *J. Biomol. Struct. Dyn.*, **13**, 309–317.
65. Struhl, K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, **98**, 1–4.
66. Kornberg, R.D. and Lorch, Y. (2002) Chromatin and transcription: where do we go from here. *Curr. Opin. Genet. Dev.*, **12**, 249–251.
67. Beato, M. and Eisefeld, K. (1997) Transcription factor access to chromatin. *Nucleic Acids Res.*, **25**, 3559–3563.
68. Landick, R., Stewart, J. and Lee, D.N. (1990) Amino acid changes in conserved regions of the beta-subunit of *Escherichia coli* RNA polymerase alter transcription pausing and termination. *Genes. Dev.*, **4**, 1623–1636.
69. Weillbaecher, R., Hebron, C., Feng, G. and Landick, R. (1994) Termination-altering amino acid substitutions in the beta' subunit of *Escherichia coli* RNA polymerase identify regions involved in RNA chain elongation. *Genes. Dev.*, **8**, 2913–2927.
70. Zhou, Y.N. and Jin, D.J. (1998) The rpoB mutants destabilizing initiation complexes at stringently controlled promoters behave like 'stringent' RNA polymerases in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **95**, 2908–2913.
71. Dobinson, K.F. and Spiegelman, G.B. (1987) Effect of the delta subunit of *Bacillus subtilis* RNA polymerase on initiation of RNA synthesis at two bacteriophage phi 29 promoters. *Biochemistry*, **26**, 8206–8213.
72. Bird, A.P. (1995) Gene number, noise reduction and biological complexity. *Trends Genet.*, **11**, 94–100.
73. Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R. et al. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.