Curved DNA in promoter sequences

Andrei E. Gabrielian, David Landsman* and Alexander Bolshoy**

Computational Biology Branch,
National Center for Biotechnology Information,
National Library of Medicine,
National Institutes of Health,
Building 38A, Room 8N805,
Bethesda, MD 20894,
USA.


Running Title: Curved promoters

Keywords: DNA curvature, bendability, sequence analysis

* to whom all correspondence should be addressed
e-mail: landsman@ncbi.nlm.nih.gov    fax: 301-435-7794

** present address: Institute of Evolution, University of Haifa, Haifa, 31905 ISRAEL
ABSTRACT

Sequence-dependent DNA curvature is known to play an important role in initiation of transcription of many genes. We compared the distribution of predicted intrinsic curvature of *Escherichia coli* and human promoter sequences with the distribution of curvature of randomly selected coding and non-coding fragments from these organisms. Different methods of curvature calculation were found to yield mostly similar overall tendencies of DNA curvature in all groups of sequences. According to all methods of calculation, *E. coli* promoters were found to be more curved than coding sequences from the same genome and random sequences with the same nucleotide composition. By contrast, the average curvature of human promoter sequences was only marginally greater than the average curvature of human coding sequences. Non-coding intron sequences were found to be the most curved of the human sequences examined. Based on these observations, we hypothesize about the role of DNA curvature in promoter sequences.
INTRODUCTION

For more than 10 years, the biological importance of intrinsic DNA curvature in gene expression has intrigued researchers (for reviews, see (1-4)). The hypotheses explaining the possible role of curvature in promoter regions fall into four groups: (i) curved DNA could form large loops around RNA polymerase, thereby enhancing the affinity of the complex (4,5); (ii) it has been demonstrated that even a small intrinsic curvature may greatly enhance the affinity of protein-DNA contacts, which might lead one to speculate that local curvature serves to fine-tune the interaction of promoters with regulatory factors (6); (iii) it has been suggested that DNA curvature or looping brings together components of the transcriptional complex that are distant along the DNA sequence (5); (iv) curvature and/or the superhelical structure of DNA results in torsional stress, which affects the energy of DNA melting and double helix unwinding, thus assisting, or even functioning in place of, the corresponding initiation proteins (7).

The role of curvature in promoter regions has been studied experimentally for several genes. Nickerson and Achberger (8) examined the role of curved DNA in the binding of _E. coli_ RNA polymerase to certain promoters. Carmona et al. (9,10) analyzed the effect of DNA curvature on transcription initiation in _E. coli_ for _α_54-dependent promoters and concluded that the initiation of transcription requires either intrinsic or induced DNA bending. The histone-like regulatory protein, H-NS of _E. coli_ was demonstrated to bind preferentially to curved DNA in the promoter region (11,12). Recently, Pedersen et al. (13) used hidden Markov models to identify sequence patterns in upstream and downstream regions of human promoters. Based on the periodicity of certain trinucleotides, they suggested that the distribution of bendability in downstream regions is similar to typical nucleosome binding sites. Many other examples of the role DNA bending plays in gene expression can be found in reviews by Hagerman (1) and
Harrington (2). It has also been speculated that intrinsic DNA curvature is a common structural element of eukaryotic promoters (14,15). New experimental data support this hypothesis. Nair reported the presence of an intrinsic DNA bend with a broad locus of curvature for the human cdc2 promoter and concluded that this intrinsic bend might influence the protein-induced bending of a promoter region and therefore the subsequent interaction of regulatory factors with the transcription machinery (16). Leech et al. reported variations in HLA-DQB1 promoter elements related to curvature, with distinct allelic differences in the predicted DNA curvature of these sites (17).

However, the mechanisms whereby curved DNA affects biological processes remain to be determined. Existing data imply that they are quite complex and that they vary depending on the presence or absence of certain regulatory proteins, other ligands, or ions. Questions remain whether intrinsic DNA curvature is a necessary component of all promoter architectures; whether DNA curvature affects promoter activity in general or only in specific cases; and what are the biological classes of genes that have either promoters with significant DNA curvature or promoters where DNA curvature is negligible. Such experimental investigations might be facilitated by a computational approach. Another more practical question is whether curvature determinations could be used to identify promoters in genomes, or at least assist in this process. Efficient discrimination requires that curvature is present predominantly in promoter regions, rather than in the rest of the genome.

Our goal was to compare the average curvature of promoters in a prokaryotic (E. coli) and a eukaryotic (human) genome with randomly shuffled fragments as well as with randomly selected sequences from the same genomes. There are different approaches to predicting DNA curvature, not all are based on the computation of the three-dimensional shape of DNA
molecules. To find curved segments in genomic sequences, we used three known algorithms (see Materials and Methods). Comparing data obtained with three distinct algorithms of curvature prediction allowed us to avoid biases of approach, scale, and parameters, whereby the consistency of the results increased the reliability. Our most interesting finding was the difference between human and *E. coli* promoters. While *E. coli* promoters were found on average to be more curved than all control sets, human promoters were less curved than intron sequences and practically indistinguishable from coding fragments from its genome. By our calculations, differences between average curvature in human sequences are not statistically significant.
MATERIALS AND METHODS

Curvature calculations:

We have used three different methods for the estimation of intrinsic DNA curvature. Two of them employ the static model, calculating the curvature based on the dinucleotide step parameters (see below). Only the CURVATURE program calculates the three-dimensional structure of DNA molecule. Another method (the program PERIOD) does not directly calculate the curvature, but instead estimates the periodicity of bendable DNA fragments, which leads to the DNA curvature. While the bendability itself does not correlate with the curvature, the periodicity of bendability does. The reason for this is the anisotropy of bendability (a preference for DNA bending towards the major groove). If highly bendable fragments are arranged according to the period of B-DNA (i.e. close to 10.2 bp), the resulting DNA segment would bend preferentially in one direction (towards the major groove) only, as distinct from the case where bendable fragments are distributed randomly. The preferential bending, therefore, directs the curvature of DNA segment.

It is important to understand that none of the methods below (or other existing method) can predict the curvature of DNA complexed with a protein. It would only be possible to predict the DNA structure in a DNA-protein complex if the three-dimensional structure of similar complexes is known and if all DNA sequences important for binding have been determined. We do not consider such analyses in this work.

(1) Calculation of the helical asymmetry of bendability coefficient of DNA sequences using the PERIOD program (18).

It has been shown that experimentally mapped curved DNA sites frequently possess a periodic distribution of bendable and rigid fragments (19). The numerical coefficient "helical
asymmetry" (analogous to that used for protein \(\alpha\)-helices) uses Fourier analysis to estimate the periodicity of bendability based on the bending propensities of overlapping trinucleotides. While correlating with curvature, helical asymmetry of bendability is not a geometric parameter. It can be considered as a measure of "compatibility" of DNA sequence with a curved conformation. It was demonstrated that helical asymmetry of bendability, but not bendability itself, correlates well with experimental gel mobility data (i.e. curvature) for synthetic oligonucleotides (19). Sequence-dependent bendability data were taken from (18). The "consensus" scale was obtained by averaging two bendability scales of Satchwell et al. (20) and Brukner et al. (21,22), derived from statistics of nucleosome positioning and DNaseI cutting frequencies. Previously, it had been shown that this scale could recognize both AA/TT and CC/GG types of curvature more efficiently than did its predecessors (18). The program for calculating helical asymmetry is available at the Web site http://www.icgeb.trieste.it/dna.

(2) Calculation of the bending and curvature of DNA sequences using the BEND program (23).

This program utilizes a simplified approach for calculating the amplitude of curvature, without prior calculation of the three-dimensional structure of DNA. With this algorithm, we used two sets of roll, tilt, and twist angles; those of DeSantis et al. (24) and Bolshoy et al. (25). We slightly modified the published version of the BEND program by excluding the arbitrary pre-smoothing of angles before calculation of curvature. It is worth mentioning that despite the implication of the standard dinucleotide parameters (i.e. roll, tilt, and twist angles), this is a simplified approach, which is not mathematically equivalent to the calculation of curvature on the basis of pre-computed or known three-dimensional structures.

(3) Calculation of the distribution of curvature of DNA sequences using the CURVATURE program (26).
This program calculates the three-dimensional path of DNA molecules and estimates the segment curvature by computing the radius of the arc approximating to the path of the axis of the DNA fragment (26). The dinucleotide wedge angles of Bolshoy et al. (25) and the twist angles of Kabsch et al. (27) were used for all calculations.

The CURVATURE algorithm is based on the stepwise calculation of geometric transformations according to the set of roll, tilt and twist angles. The BEND algorithm is a simplified calculation of the amplitude of curvature, that is satisfactory only for very small roll and tilt angles. The BEND algorithm, being a simplification of approaches based on actual structure calculations, is however, not based on a different concept like the helical asymmetry of bendability (PERIOD program). Our results showed that for the purposes of the statistical characterization of large amounts of genomic information, both BEND and CURVATURE algorithms produce similar results. The subsequent analyses of the detailed structure of promoters is possible with the CURVATURE program only.

Despite these different approaches, we conditionally use the term "curvature" for all predictions made by the above three methods, since they effectively identify regions of intrinsic DNA curvature.

Comparison of the results

Since we did not know a priori whether overall or local curvature is more important, we decided to produce results using a range of sliding windows. For all sequences, we calculated the curvature of sequence segments with windows of 21, 30, 60, and 90 bases. Thus, for a sequence of length L, we obtained \((L - W + 1)\) values of curvature, where \(W\) is the window length. The resulting curvature of each sequence was calculated as the average of curvature for overlapping windows. Another measure we tested was the maximal curvature for the sequence, which
produced results very similar to those obtained from the mean curvature. For comparison, all results were normalized either to the average genome curvature for *E. coli* sequences (i.e. predominantly coding sequence) or to the average curvature of exons of human sequences (i.e. likewise, predominantly coding sequence), calculated with the same window length. Thus, the normalization procedures are similar. The comparison of two datasets was based on the difference between curvature values averaged over the whole dataset. More detailed comparisons were based on histograms of distributions for average curvature values, which allowed us to see where datasets differed most.

**Databases**

Sequences of *E. coli* and human promoters were taken from the databases of Lisser and Margalit (28) and from Release 50 of Eukaryotic Promoter Database (EPD) of Bucher and Trifonov (29). The database of *E. coli* promoter sequences consists of 290 fragments, each 101 nucleotides in length. To facilitate comparisons with the *E. coli* database, we analyzed human promoters of the same length (101 nucleotides), starting at position -75 relative to the transcription start site. All calculations were done for this "core" region, except when we compared curvature upstream of core promoter regions (see Results and Discussion) (see Figures 2 and 4). Control sets of random DNA fragments were obtained by shuffling the corresponding promoter sequences. Shuffled promoters have the same nucleotide composition as real ones, but they lack the specific sequence patterns. Other control sets were prepared by random fragment selection from appropriate sequences. The entire *E. coli* genome is available and the boundaries of coding regions are known (30). Three *E. coli* control (non-promoter) datasets were compiled by random selection of 290 segments with the same length of 101 nucleotides from coding and non-coding regions, and by shuffling promoter sequences. Sequences of human exons and
introns were taken from the Non-Redundant Functionally Equivalent Sequences Database (NRFES) of Konopka (31). Fragments of length 101 nucleotides were randomly selected from the sets of human exon and intron sequences.
RESULTS

Curvature of *E. coli* sequences

The mean calculated curvature for promoters and the three controls (non-promoter datasets) is presented in Table 1. It is striking that for all windows and scales the main conclusion is the same: the average curvature is highest for promoter sequences. All methods, except for the helical asymmetry coefficient method calculated with a window size of 90 nucleotides, ranked coding sequences as the least curved. In all cases, shuffling decreased the curvature of the corresponding fragment, suggesting that the presence and order of existing sequence patterns in promoters favor the formation of curvature. The non-coding set had smaller average curvature than promoters, but exceeded that of other control datasets (coding and shuffled promoters). It is evident that, on average, *E. coli* promoters are significantly more curved than control sequences, a difference that is particularly noticeable when curvature is measured in the window of 90 nucleotides.

More detailed information was obtained by comparing the curvature distributions of promoters and control sets. In Figure 1, which is a histogram of the curvature distributions for promoters and control sequences, we show such a comparison. The distributions for promoters and other sequences differ, although sometimes they overlap considerably. Therefore, as a set, the *E. coli* promoters exhibit a different range of curvature from those of the control datasets. The difference is particularly evident when these promoters are compared with coding sequences; in fact, the difference is of sufficient magnitude that it could be used to distinguish different classes of sequences. Based on these histograms, one could ascertain a curvature threshold; since only a few coding sequences would exceed this threshold, these threshold values could complement the identification of promoter sequences in large-scale genomic sequencing.
projects. Our calculations therefore lead to the conclusion that *E. coli* promoters are more frequently curved than coding sequences, and that this difference is at least partially due to sequence patterns characteristic of promoters.

In the *E. coli* promoter database (28), each promoter sequence is represented by the region close to the start of transcription (from 75 bases upstream to 25 bases downstream from the start; 101 nucleotides in total). In order to ascertain the importance of the localization of curved DNA, we determined the distribution of curved segments in "extended" promoter regions (i.e. whether curvature is localized in the "core" promoter regions or is located further upstream). For this purpose, we selected a subset of the Lisser-Margalit compilation which is a set of 136 unique promoters, whose positions in the *E. coli* genome were determined. Each sequence in this dataset was extended to 500 nucleotides; 400 nucleotides upstream to 100 nucleotides downstream from the start of transcription. In this promoter subset, the distribution of the average curvature in the interval from -75 to +25 is indistinguishable from that of the whole set of 290 promoters (data not shown). The histograms depicted in Figure 2 show that the "core" regions, between -100 and the transcription start site (i.e. position 0), are indeed most curved for the majority of promoters.

Recently, Collado-Vides and co-workers published a complete set of predictions for promoters and regulatory sites in *E. coli* genome ((32), (33) and http://www.cifn.unam.mx/Computational_Biology/E.coli-predictions). These include the location of 3398 core promoters (from the -35 box to the transcription start) as well as 699 sequences containing binding sites for known regulatory proteins. We have calculated the average curvature for all these sequences and found that the binding sequences were, on average, the most curved sequences in the *E. coli* genome (data not shown). The curvature of these core promoters is close to the average curvature
values that were obtained from the smaller Lisser-Margalit database (see Table 1). We did not find any correlation between the presence of either the -35 or -10 box motifs and promoter curvature. Previously, Collado-Vides et al. collated the information about the positions of activator and repressor binding sites in promoter regions (34). The majority of these sites are positioned between the transcription start and -70 nucleotides upstream. Thus, it is tempting to conclude that DNA curvature in the promoter regions of E. coli genes is mostly a consequence of the regulatory factor binding sites. Predicted regulatory sites were found in 16% of regions upstream of operons in their promoter regions and 10% of intra-operon, non-coding regions (32).

If curvature is predominantly associated with DNA binding sites, it would be difficult to identify promoters based solely on calculations of DNA curvature. Conversely, many of the DNA binding sites for regulatory proteins currently remains unknown, therefore it is impossible to verify whether all curved sites correspond to DNA binding sites in a genome.

**Curvature of human promoters**

The calculated mean curvature of promoter sequences and control sets are presented in Table 2. The results for human promoter sequences were quite distinct from those obtained for E. coli promoter sequences. The average curvature of the human promoter sequences do not differ significantly from curvature of shuffled promoter sequences, lying between those of introns (the most curved group in humans) and those of exons (the least curved group). Curvature histograms of human datasets (Figure 3) clearly demonstrate these tendencies, with only the distribution for introns slightly shifted. On the other hand, promoters, shuffled promoters, and exons exhibit very similar curvature distributions. Similarly, graphs analogous to Figure 3 were obtained for all other scales and window parameters. Comparing absolute curvature values (e.g. Figure 3 vs. Figure 1), we found that, on average, curvature of human promoter sequences for all scales and
windows are smaller than the corresponding curvature values of the *E. coli* promoter sequences. Even introns, the most curved sequences, are on average as curved as *E. coli* coding sequences, which are far less curved than *E. coli* promoter sequences.

We also tested whether neighboring regions contain more curved sites than do the cores of promoters. For this more detailed examination, we calculated the distributions of curvature in the 500 nucleotide long upstream regions of human promoters, divided into 100 nucleotide non-overlapping intervals (Figure 4). All datasets exhibit curvature distributions similar to those for "core" promoters. Thus, as distinct from *E. coli*, there seems to be no predominant curvature property in DNA sequences upstream of human genes.

**Testing statistical significance**

To estimate the statistical significance of the difference between the average curvature for promoters and coding sequences, we performed a Student's t-test to estimate the probability of obtaining such differences by chance. Probabilities were calculated for the results of the BEND program and De Santis matrices of roll, tilt and twist angles, with window size of 90 nucleotides; analogous calculations were done for other scales, approaches and windows with similar results. The differences between the average curvature of *E. coli* promoter and coding sequences as well as between promoters and randomized promoters were found highly significant (p < 0.001 and p < 0.01). We did not find a statistically significant difference between the average curvature of *E.coli* promoter and non-coding sequences. It would be expected that most of the *E. coli* intergenic, non-coding sequences encompass significant amounts of promoter sequences and binding sites for regulatory proteins. For human sequences, the average curvature for all sequence sets was too similar to be considered statistically different. When we performed the standard Student's t-test, even the most curved human sequences (introns) were not significantly
different from the least curved human sequences (exons) (p < 0.2; data for BEND program calculations, De Santis matrices, window of 90 nucleotides). Thus, the differences that we estimate when comparing the average curvature of human promoters, exons and introns, are subtle. It is particularly evident when comparing the histograms of curvature distribution.

DISCUSSION

Our analysis does not attempt to answer all questions about the role of intrinsic DNA curvature in promoters. Nevertheless, the extensive comparison of curvature in promoter sequences and in coding regions as well as neighboring non-coding regions, seems a necessary step in understanding the significance of DNA curvature in transcription regulation.

Intrinsic DNA curvature predictions can be computed at different levels of granularity. In this study, we chose to characterize the curvature of a whole DNA fragment by a single, averaged value. There are several striking differences between the DNA curvature distributions for human and *E. coli* promoter sequences. In *E. coli* the average curvature of promoter sequences exceeds that of coding sequences. Since curved DNA regions are often good targets for regulatory proteins, and the lowest average curvature for the *E. coli* genome was found among coding sequences, it would not seem unreasonable to assume that most of the curvature is located in the intergenic regions which contains these regulatory elements. Our finding that the average curvature of predicted regulatory factor binding sites exceeds that of core promoters, further supports this hypothesis. It remains to be seen what classes of regulatory proteins require the presence of the intrinsic curvature and whether variations of this intrinsic curvature correspond to the affinity of protein-DNA contacts.
Both for *E. coli* and human promoters, none of the predictive methods showed that curvature is particularly pronounced in regions further upstream of core promoters (Figure 2 and 4). For *E. coli*, core promoters (i.e. sequences between the transcription start site and position -100) are the most curved when compared to other *E. coli* sequences, whereas for similar human promoter sequence regions, there is hardly any difference when compared to either exons, introns, or shuffled sequences.

If the average level of curvature (or the frequency of occurrence of extremely curved sites) in a set of sequences with the same or similar function, significantly exceeds that of a comparison group, one might speculate that curvature plays a role in a common function. However, that does not mean that curvature may be used to distinguish such sequence sets. The question we have addressed is whether it is possible to predict the location of promoters in the *E. coli* genome based on curvature calculations from the sequence data. It is expected that the known promoter sequences would exhibit a distinct, elevated, average curvature distribution compared to the rest of the genome. However, this is not the case. If we specify that the threshold for "very curved" sequences is three standard deviations above the average curvature level for the genome, then 99.3% of the genome sequences would not be classified as very curved DNA (data obtained with helical asymmetry coefficients calculated with a consensus bendability scale). Among the 290 known promoter sequences, 55 sequences have DNA curvature values that exceed this threshold (data not shown). Similar data obtained with the BEND program and the DeSantis scale, where 0.8% of the genome sequences and 38 out of 290 promoter sequences display curvature levels that exceed the average genomic curvature by three standard deviations. Similar results were obtained in the study of VanWye et al. (35). Our results clearly show that curvature can help in identifying some *E. coli* promoters, but cannot be used as
the sole criterion to search the genome for candidate promoter sequences. A more effective approach would need to combine knowledge of specific sequence patterns with curvature calculations as well as experimental, biologically relevant data. While the presence of DNA curvature was previously demonstrated for several bacterial promoters (4), our calculations suggest that there are many non-promoter sites in the genome with comparable levels of curvature.

In contrast to the *E. coli* genome, our knowledge of the gene regulatory regions in the human genome is largely incomplete. Nevertheless, the body of existing data on the regulation in the human genes and the number of sequenced genes and their promoters are large enough to perform similar comparative studies on human promoter sequences. As stated above, the results of the analysis of average curvature of human exons, introns, and promoters were very similar for all algorithms and scales used (see Table 2). The overall curvature level is noticeably lower in all human sequences than in *E. coli* sequences: the most curved human sequences (introns) are comparable to the least curved *E. coli* sequences (coding). One of the possible explanation could be that the human genome is less AT-rich than the *E. coli* genome. DNA curvature was first discovered in AT-rich sequences (36,37)]. However, it was also shown that GGGCCC-containing sequences are curved (38). These two types of curvature differ in direction and their response to the presence of certain ions and other ligands (38,39)]. It has been noted that existing scales of dinucleotide angles are not efficient in predicting curvature in GC-rich sequences (18,23,40). Satchwell et al. (20) have posited that their bendability scale is sensitive to GGGCCC-curvature. To ascertain whether this was indeed an issue in this study, we calculated the average curvature for all sequence sets using the Satchwell bendability scale (20). The differences between the average curvature of promoters, exons, and introns sequences were the
same as for other scales (data not shown). Thus, the overall nucleotide composition of a genome does not seem to affect the distribution of curvature in that genome.

One important aspect of DNA curvature is that it is a highly cooperative and additive feature that cannot be attributed to short DNA fragments. It should not be confused with local bending. Local bends do not necessarily result in curvature which could be viewed as a product of several co-oriented bends. The question of positioning of curved fragments according to, for example, the start of transcription, is extremely difficult to address because any curved fragment ought to span at least several helical turns. Therefore, we will not discuss the orientation of DNA curvature because of a vague definition. On the other hand, the distance between the start of transcription and a local bend could be easily measured, but we have not studied the statistics of local DNA bending.

Although curvature may play a role in promoter function, it is hard to speculate on the mechanism because we do not know much about the functional aspects related to DNA curvature (e.g. influence of curvature on protein and small ligand binding, and on hydration, strand-strand interactions, and stacking energies). There are several hypotheses that are not mutually exclusive that could explain the differences between the curvature distributions in human and E. coli promoter sequences. Nucleosomes tend to be positioned on curved DNA (41,42). Furthermore, recent studies on the sequence patterns in nucleosomal DNA indicates that only certain DNA sequences bind to the most stable nucleosome subset and that there is a significant amount of these sequences in centromeres suggesting a functional role for these stable nucleosomes (43). Many of these stable nucleosomal sequences are predicted to be strongly curved (Bolshoy, Gabelian, and Landsman, unpublished data). Thus, it is not surprising that strongly curved DNA is absent from core promoters, because the presence of particularly stable nucleosomes in
promoter regions could interfere with transcription (44,45). In addition, the set of regulatory proteins in human and *E. coli* differs considerably and the more extensive set of human regulatory factors consists of a wider variety of structural types, with only a small subset requiring curved DNA for specific binding. Thus, the more diverse set of human regulatory proteins might exclude the need for using DNA curvature as an effector of DNA-protein interactions. Finally, since the intergenic regions in bacterial genomes are shorter than those in eukaryotic genomes, the presence of curvature might be more effective for the targeting of regulatory proteins.

Although currently there is no algorithm or scale that guarantees the accurate prediction of DNA conformation, the significant similarity of the results obtained by using several widely accepted approaches to calculate DNA curvature has substantiated the implications of our findings. Our analysis was not a comparison of DNA curvature models or different dinucleotide angles. This task would require a more comprehensive coverage of existing scales and a broader selection of known curved DNA fragments. Our goal is strictly to compare DNA curvature distributions between promoter and non-promoter sequences. The answer to this question would ascertain whether DNA curvature could be used for the prediction of promoters and/or a variety regulatory sites. Using the current computational approaches for DNA curvature calculations, we can not reliably predict the positions of promoters in all *E. coli* and human genes. However, our results do show that, on average, *E. coli* promoters are the most curved regions in this genome. This could suggest further analysis of prokaryotic promoters after an appropriate classification into subgroups or after development of novel algorithms for DNA structure prediction in solution.
Acknowledgments

A.B. was supported by the NCBI Scientific Visitors Program at the National Library of Medicine. Dr. Birgit An der Lan provided excellent editorial assistance.
References.


Table 1
The ratios of average curvature in *E. coli* promoter sequences to the average of curvature in *E. coli* genomic sequences

A. Curvature calculated by the BEND program (23) using the De Santis et al. dinucleotide angles (24).

<table>
<thead>
<tr>
<th>sequence set</th>
<th>window size (nucleotides)</th>
<th>21</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>promoters</td>
<td></td>
<td>1.17</td>
<td>1.21</td>
<td>1.23</td>
<td>1.25</td>
</tr>
<tr>
<td>shuffled promoters</td>
<td></td>
<td>1.00</td>
<td>1.04</td>
<td>1.03</td>
<td>1.06</td>
</tr>
<tr>
<td>coding</td>
<td></td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td>non-coding</td>
<td></td>
<td>1.12</td>
<td>1.15</td>
<td>1.17</td>
<td>1.21</td>
</tr>
</tbody>
</table>

B. Curvature calculated by the BEND program (23) using the Bolshoy et al. dinucleotide angles (25).

<table>
<thead>
<tr>
<th>sequence set</th>
<th>window size (nucleotides)</th>
<th>21</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>promoters</td>
<td></td>
<td>1.19</td>
<td>1.19</td>
<td>1.17</td>
<td>1.18</td>
</tr>
<tr>
<td>shuffled promoters</td>
<td></td>
<td>1.08</td>
<td>1.06</td>
<td>1.07</td>
<td>1.05</td>
</tr>
<tr>
<td>coding</td>
<td></td>
<td>1.00</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>non-coding</td>
<td></td>
<td>1.14</td>
<td>1.14</td>
<td>1.15</td>
<td>1.17</td>
</tr>
</tbody>
</table>

C. Curvature calculated by the consensus bendability scale, helical asymmetry coefficient (18).

<table>
<thead>
<tr>
<th>sequence set</th>
<th>window size (nucleotides)</th>
<th>21</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>promoters</td>
<td></td>
<td>1.16</td>
<td>1.17</td>
<td>1.21</td>
<td>1.17</td>
</tr>
<tr>
<td>shuffled promoters</td>
<td></td>
<td>0.96</td>
<td>0.98</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>coding</td>
<td></td>
<td>0.98</td>
<td>1.00</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>non-coding</td>
<td></td>
<td>1.13</td>
<td>1.16</td>
<td>1.20</td>
<td>1.21</td>
</tr>
</tbody>
</table>

D. Curvature calculated by the CURVATURE program (26) using the Bolshoy et al. dinucleotide angles (25).

<table>
<thead>
<tr>
<th>sequence set</th>
<th>window size (nucleotides)</th>
<th>21</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>promoters</td>
<td></td>
<td>1.22</td>
<td>1.20</td>
<td>1.18</td>
<td>1.20</td>
</tr>
<tr>
<td>shuffled promoters</td>
<td></td>
<td>1.11</td>
<td>1.13</td>
<td>1.09</td>
<td>1.06</td>
</tr>
<tr>
<td>coding</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>non-coding</td>
<td></td>
<td>1.17</td>
<td>1.18</td>
<td>1.16</td>
<td>1.20</td>
</tr>
</tbody>
</table>
Table 2
The ratios of average DNA curvature of human promoter sequences to average curvature of exon sequences

A. Curvature calculated by the BEND program (23) using the De Santis et al. dinucleotide angles (24).

<table>
<thead>
<tr>
<th>sequence set</th>
<th>window size (nucleotides)</th>
<th>21</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>promoters</td>
<td></td>
<td>1.03</td>
<td>1.03</td>
<td>1.10</td>
<td>1.05</td>
</tr>
<tr>
<td>shuffled promoters</td>
<td></td>
<td>1.03</td>
<td>1.03</td>
<td>1.04</td>
<td>1.00</td>
</tr>
<tr>
<td>exons</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>introns</td>
<td></td>
<td>1.12</td>
<td>1.12</td>
<td>1.12</td>
<td>1.08</td>
</tr>
</tbody>
</table>

B. Curvature calculated by the BEND program (23) using the Bolshoy et al. dinucleotide angles (25).

<table>
<thead>
<tr>
<th>sequence set</th>
<th>window size (nucleotides)</th>
<th>21</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>promoters</td>
<td></td>
<td>1.03</td>
<td>1.04</td>
<td>1.10</td>
<td>1.13</td>
</tr>
<tr>
<td>shuffled promoters</td>
<td></td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>exons</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>introns</td>
<td></td>
<td>1.12</td>
<td>1.11</td>
<td>1.13</td>
<td>1.15</td>
</tr>
</tbody>
</table>

C. Curvature calculated by the consensus bendability scale, helical asymmetry coefficient (18).

<table>
<thead>
<tr>
<th>sequence set</th>
<th>window size (nucleotides)</th>
<th>21</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>promoters</td>
<td></td>
<td>1.05</td>
<td>1.03</td>
<td>1.05</td>
<td>1.01</td>
</tr>
<tr>
<td>shuffled promoters</td>
<td></td>
<td>1.02</td>
<td>1.03</td>
<td>1.05</td>
<td>1.24</td>
</tr>
<tr>
<td>exons</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>introns</td>
<td></td>
<td>1.11</td>
<td>1.11</td>
<td>1.15</td>
<td>1.12</td>
</tr>
</tbody>
</table>

D. Curvature calculated by the CURVATURE program (26) using the Bolshoy et al. dinucleotide angles (25).

<table>
<thead>
<tr>
<th>sequence set</th>
<th>window size (nucleotides)</th>
<th>21</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>promoters</td>
<td></td>
<td>1.03</td>
<td>1.08</td>
<td>1.17</td>
<td>1.26</td>
</tr>
<tr>
<td>shuffled promoters</td>
<td></td>
<td>1.01</td>
<td>1.05</td>
<td>1.06</td>
<td>1.11</td>
</tr>
<tr>
<td>exons</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>introns</td>
<td></td>
<td>1.13</td>
<td>1.15</td>
<td>1.22</td>
<td>1.29</td>
</tr>
</tbody>
</table>
Figure legends.

**Figure 1. Curvature distribution of *E. coli* sequences.**
DNA curvature was calculated using the CURVATURE program (26) with the Bolshoy *et al.*
dinucleotide angles (25), a window size of 21 bp, and averaging over the whole sequence. The red line shows the average curvature of 290 promoters from the Lisser and Margalit compilation (28); the blue and green lines, the average curvature of randomly selected fragments from coding and non-coding regions, respectively; the cyan line, the curvature of sequences obtained by random shuffling of the promoters; and the vertical dashed magenta line is the mean curvature of the *E. coli* genome.

**Figure 2. Curvature distribution in *E. coli* upstream promoter regions.**
DNA curvature calculations were performed as described above in Figure 1. The distribution of the average curvature of 136 promoters selected from (28). The orange line is a histogram of curvature of downstream promoter regions in the range of -1 to +100. Red, blue, green and cyan lines are histograms of curvature of upstream promoter regions in ranges of windows [+1 to -100], [-100 to -200], [-200 to -300] and [-300 to -400], respectively. The vertical dashed magenta line is the mean curvature of the *E. coli* genome.

**Figure 3. Curvature distribution in human DNA sequences.**
DNA curvature calculations were performed as described above in Figure 1. All histogram curves are smoothed by a running average of three. The red line shows the average curvature of 210 human promoters from Eukaryotic Promoter Database of Bucher & Trifonov (29) of length 101 bp, starting from position -75 relative to the start of transcription. The green and blue lines correspond to histograms of the average curvature of fragments of the same length of 101 bp randomly selected from human exons and introns, respectively, from the NRFES database (31). The cyan line is the average curvature of 210 shuffled human promoter sequences.

**Figure 4. Curvature distribution in the upstream regions of human promoters.**
DNA curvature was calculated using the CURVATURE program (26) with the Bolshoy *et al.*
dinucleotide angles (25), with a window size of 90 bp. All histogram curves are smoothed by a running average of three. In total, six sets of sequence fragments were studied; five upstream sets and one downstream set of fragments. All sequences have the length of 101 bp; the location of fragments referred to the start of transcription (distance in bp.) and corresponding colors are
indicated on the graph. The beige line is a histogram of the curvature of downstream promoter regions in the range of $-1$ to $+100$. Red, blue, green, yellow, and magenta lines are histograms of curvature of upstream promoter regions in the ranges of windows $[+1 \text{ to } -100]$, $[-100 \text{ to } -200]$, $[-200 \text{ to } -300]$, $[-300 \text{ to } -400]$, and $[-400 \text{ to } -500]$, respectively.