doi:10.1006/jmbi.2000.3787 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2000) **299**, 907–930

JMB

A DNA Structural Atlas for *Escherichia coli*

Anders Gorm Pedersen†, Lars Juhl Jensen†, Søren Brunak
Hans-Henrik Stærfeldt and David W. Ussery*

Center for Biological Sequence Analysis, Department of Biotechnology, The Technical University of Denmark
Building 208, DK-2800
Lyngby, Denmark

*Corresponding author

We have performed a computational analysis of DNA structural features in 18 fully sequenced prokaryotic genomes using models for DNA curvature, DNA flexibility, and DNA stability. The structural values that are computed for the *Escherichia coli* chromosome are significantly different from (and generally more extreme than) that expected from the nucleotide composition. To aid this analysis, we have constructed tools that plot structural measures for all positions in a long DNA sequence (e.g. an entire chromosome) in the form of color-coded wheels (http://www.cbs.dtu.dk/services/GenomeAtlas/). We find that these ''structural atlases'' are useful for the discovery of interesting features that may then be investigated in more depth using statistical methods. From investigation of the *E. coli* structural atlas, we discovered a genome-wide trend, where an extended region encompassing the terminus displays a high of level curvature, a low level of flexibility, and a low degree of helix stability. The same situation is found in the distantly related Gram-positive bacterium *Bacillus subtilis*, suggesting that the phenomenon is biologically relevant. Based on a search for long DNA segments where all the independent structural measures agree, we have found a set of 20 regions with identical and very extreme structural properties. Due to their strong inherent curvature, we suggest that these may function as topological domain boundaries by efficiently organizing plectonemically supercoiled DNA. Interestingly, we find that in practically all the investigated eubacterial and archaeal genomes, there is a trend for promoter DNA being more curved, less flexible, and less stable than DNA in coding regions and in intergenic DNA without promoters. This trend is present regardless of the absolute levels of the structural parameters, and we suggest that this may be related to the requirement for helix unwinding during initiation of transcription, or perhaps to the previously observed location of promoters at the apex of plectonemically supercoiled DNA. We have also analyzed the structural similarities between groups of genes by clustering all RNA and protein-encoding genes in *E. coli*, based on the average structural parameters. We find that most ribosomal genes (protein-encoding as well as rRNA genes) cluster together, and we suggest that DNA structure may play a role in the transcription of these highly expressed genes.

Keywords: plectonemically supercoiled DNA; sequence-dependent DNA structure; promoter structural profile; whole-genome analysis

Introduction

Although *B*-form DNA is typically depicted as a uniformly straight and rigid double helix, it has in fact been found to possess inherent structural properties that play a role in many different biological processes (Horwitz & Loeb, 1990; Perez-Martin *et al.*, 1994; Perez-Martin & de Lorenzo, 1997; Sinden *et al.*, 1998). For instance, the exact positioning of nucleosomes in eukaryotic chromatin has in some cases been demonstrated to rely on local differences in DNA flexibility and curvature (Simpson, 1991; Iyer & Struhl, 1995; Wolffe & Drew, 1995; Ioshikhes *et al.*, 1996; Zhu & Thiele,

†These authors contributed equally to this work.
E-mail address of the corresponding author: dave@cbs.dtu.dk.

1996; Liu & Stein, 1997). Furthermore, target site recognition of a number of DNA-bending proteins (including the TATA box binding protein, TBP), has been found to depend on inherent DNA structure (Parvin *et al*., 1995; Starr *et al*., 1995; Grove *et al*., 1996). As a final example, there are several known cases in prokaryotes where inherent or induced DNA bending influences the interaction between RNA polymerase and regulatory proteins (Bracco *et al*., 1989; Hoover, *et al*., 1990; Lobel & Schleif, 1991; Serrano *et al*., 1991; Richet & Søgaard-Andersen, 1994; Valentin-Hansen *et al*., 1996).

It has been shown that such local DNA structural properties depend on the exact nucleotide sequence (Brukner *et al*., 1990; Bolshoy *et al*., 1991; Hassan & Calladine, 1996; Olson *et al*., 1998; Sinden *et al*., 1998). This phenomenon is, to some degree, caused by stacking interactions between adjacent base-pairs (Hunter, 1993, 1996), although sequence-dependent binding of cations may also be involved (Hud *et al*., 1999; Sines & Williams, 1999).

Several different experimental approaches have been taken to quantify the connection between DNA sequence and structure, resulting in a set of di- or trinucleotide-based models. We have used such models to study the DNA structural features of nucleosome positioning patterns, eukaryotic promoters, and triplet repeats involved in human hereditary disorders (Baldi *et al*., 1996, 1999; Pedersen *et al*., 1998). We have also used dinucleotide parameters to investigate the periodicity of structural features within complete genomes by means of autocorrelation functions (Worning *et al*., 2000). We found a period of about 11 bp for most eubacterial organisms, and of around 10 bp for some Archaea, although the periodicity spectra differs slightly with each organism.

Given the above discussion of the relationship between sequence and structure, it is interesting that a comparison of completely sequenced genomes has revealed that each organism has its own dinucleotide signature (Karlin, 1998; Campbell *et al*., 1999). However, the dinucleotide distribution is by no means homogeneously distributed throughout the genome. Thus, there are base composition differences between different codon positions (Majumdar *et al*., 1999) and different classes of genes can have different dinucleotide distributions (Karlin *et al*., 1998). Furthermore, many bacterial genomes display differences in nucleotide composition between the leading and lagging strands of DNA replication (Mrazek & Karlin, 1998).

Here, we perform a computational analysis of DNA structure in genomes of 18 different prokaryotes, with a special emphasis on *Escherichia coli*. We do this using five different models that capture different characteristics of DNA structure (curvature, flexibility, helix stability). The resulting ''structural atlases'' show predicted structural measures for all positions in a long DNA sequence in the form of color-coded wheels. We find that these atlases are a useful tool for discovery of structural features, which can then be examined in greater depth using statistical methods.

Results and Discussion

Strategy: structural measures

For the purpose of predicting structural features we selected five different models. These predict one of three different types of structural characteristics: DNA flexibility, DNA curvature, and DNA stacking energy. Flexibility and curvature are both important for interactions between DNA and protein, while stacking energy can be interpreted as a measure of how easily the two strands of a DNA helix are separated (DNA ''meltability''). Specifically, the models we use here are described below.

(1) A DNaseI sensitivity-based model of DNA flexibility (Brukner *et al*., 1990, 1995a, henceforth referred to as the DNaseI sensitivity model). Values are in arbitrary units with higher values corresponding to greater flexibility.

(2) An X-ray crystallography-based model of protein-induced DNA flexibility derived from structural characteristics of crystallized DNA-protein complexes (Olson *et al*., 1998, ''protein-induced deformability''). Higher values correspond to great levels of flexibility.

(3) A model of DNA flexibility derived from the preference shown by individual trinucleotides to be positioned in specific orientation in nucleosomal DNA (Satchwell *et al*., 1986; Pedersen *et al*., 1998, ''position preference''). Values indicate the fractional preference of triplets for being specifically positioned in nucleosomal DNA. Lower values correspond to great flexibility.

(4) A model of DNA stacking energy derived from quantum mechanical calculations (Ornstein *et al*., 1978 ''stacking energy''). Values are in units of kcal/mol, and more negative values correspond to great stability.

(5) A model of DNA curvature derived from relative gel mobility of DNA (Bolshoy *et al*., 1991; Shpigelman *et al*., 1993, ''curvature''). Higher values correspond to higher degrees of curvature, with a value of 1 corresponding to the degree of curvature seen in nucleosomal DNA.

Briefly, all five models consist of tables giving structural parameters for di- or trinucleotides. In the case of models 1-4, structural values are assigned to every nucleotide in a DNA sequence simply by looking up the values for corresponding di- or trinucleotides. For the fifth model (curvature) dinucleotide parameters are first used to predict 3D coordinates, and the path of the predicted structure is then used to calculate a measure of local curvature at each nucleotide (see Methods).

Statistical analysis of these models demonstrated a relatively strong correlation between the stacking energy scale and the protein-induced deformability scale. Furthermore, these two models also displayed correlation to AT content. The remaining models showed no, or only very little, correlation to each other or to AT content. All models are described in more detail in Methods, along with a thorough investigation of correlation between models, and further discussion of our experimental rationale. Except for stacking energy, all these models are based on experimental investigations of sequence-structure relationships.

Structural atlas of the entire *E. coli* genome

Using the structural models mentioned above, and based on the *E. coli* K-12 MG1655 genomic sequence (Blattner *et al*., 1997, version M54, GenBank accession number U00096), we calculated the values of each of the five measures at each position in the entire genome (i.e. five times 4,639,221 real numbers). From these values we were now able to construct plots (''structural atlases'') showing structural features in any region of the genome. Figure 1 gives an overview of structural features in the entire *E. coli* circular chromosome. In this atlas, the value of each measure along the DNA sequence is shown using colored concentric wheels (one for each measure) representing the circular chromosome. The color scales are constructed so that average values are light gray, while values that are at least one standard deviation from the genomic average are brightly colored. Regions where the measure is more than three standard deviations from the genomic average are colored black (see Methods for details on the color scheme). ''Up'' on the wheels corresponds to zero minutes on the *E. coli* linkage map, with increasing minute positions in the clockwise direction. The numbers on the inside of the innermost wheel are the positions relative to zero minutes measured in millions of base-pairs (Mbp). The resolution of this whole-chromosome atlas is 928 bp (i.e. the thinnest visible line in the innermost circle, corresponds to a DNA region of length 928 bp), and the structural values have been smoothed using a running average ten times this size (i.e. 9280 bp). On the atlas we have also indicated the locations of the origin of replication (oriC, upper left part of circle, around 4 Mbp) and the terminus region (delimited by TerE and TerG, lower right part of circle, centered around 1.5 Mbp). Finally, we have indicated 36 regions possessing extreme structure on the edge of the outermost circle. The labeled regions have been chosen using two different criteria: a stringent approach based on choosing 1000 bp regions where all five measures are significantly more extreme than expected from nucleotide composition (see below), and an aesthetic approach where we labeled regions with extreme structure identifiable by visual inspection (this approach therefore depends on the resolution of the whole-genome atlas). For each extreme region we have plotted the name of a central gene (see Table 1 for a list of the stringently selected regions). In addition to *E. coli*, we have also constructed whole-genome atlases for 17 other publicly available, completely sequenced prokaryotic genomes. These (and additional) atlases are available from our regularly updated web site: http://www.cbs.dtu.dk/services/GenomeAtlas/

One feature which is immediately visible on the atlas is the existence of several positions in the chromosome where all or most of the measures agree that the local region has extreme structural properties. For instance, the region labeled *ygeG* (around 3 Mbp), can be seen to be extreme in all five measures, and many similar regions are visible as dark or colored ''spokes'' that radiate across the concentric wheels. We take this as an indication that the structural features we predict do have biological relevance. As a further substantiation of this, we found that the correlation between the structural measures is significantly higher when measured on the *E. coli* genome, than when measured on random DNA (see Methods for details on the correlation between measures). In addition, a shuffled version of the entire chromosome displayed much less extreme properties than the real DNA (see below).

Another (unexpected) feature that can be observed from the atlas is a general tendency for high levels of curvature in an extended region encompassing the terminus (outermost wheel, the region extends from approximately 1 to 2 Mbp), and a somewhat less distinct region of low-level curvature near (but not exactly at) the origin of replication. This is also illustrated in Figure 2, upper panel, which shows a smoothed profile of curvature values along the chromosome. Further analysis showed that there is a corresponding peak in AT content and stacking energy, while the flexibility measures display correspondingly low values in the same region (data not shown). Thus, there appears to be a genome-wide trend for greater AT content, higher degrees of curvature, lower levels of flexibility, and less negative stacking energy (i.e. less stable DNA) near the terminus. The fact that the peak of the curvature profile exactly coincides with the position of TerC, which is believed to be the most commonly used terminator site in *E. coli* (Hill, 1996; Bussiere & Bastia, 1999), suggests that the connection between the terminus and the structural features we predict, may be biologically relevant. In agreement with this hypothesis we find a very similar situation in the chromosome of *B. subtilis*: thus, there is a (somewhat narrower) high-curvature region encompassing the terminus, and a region of low curvature near the origin (Figure 2, lower panel).

At least two observations may be relevant in connection with this finding. First, the nucleoid-associated protein HU has been found to play a role in proper chromosome partitioning (Jaffe *et al*., 1997). HU is known to bind preferentially unusual DNA structures such as kinked or cruciform DNA (Pontiggia *et al*., 1993; Bonnefoy *et al*., 1994), and also has some affinity for smoothly curved DNA of

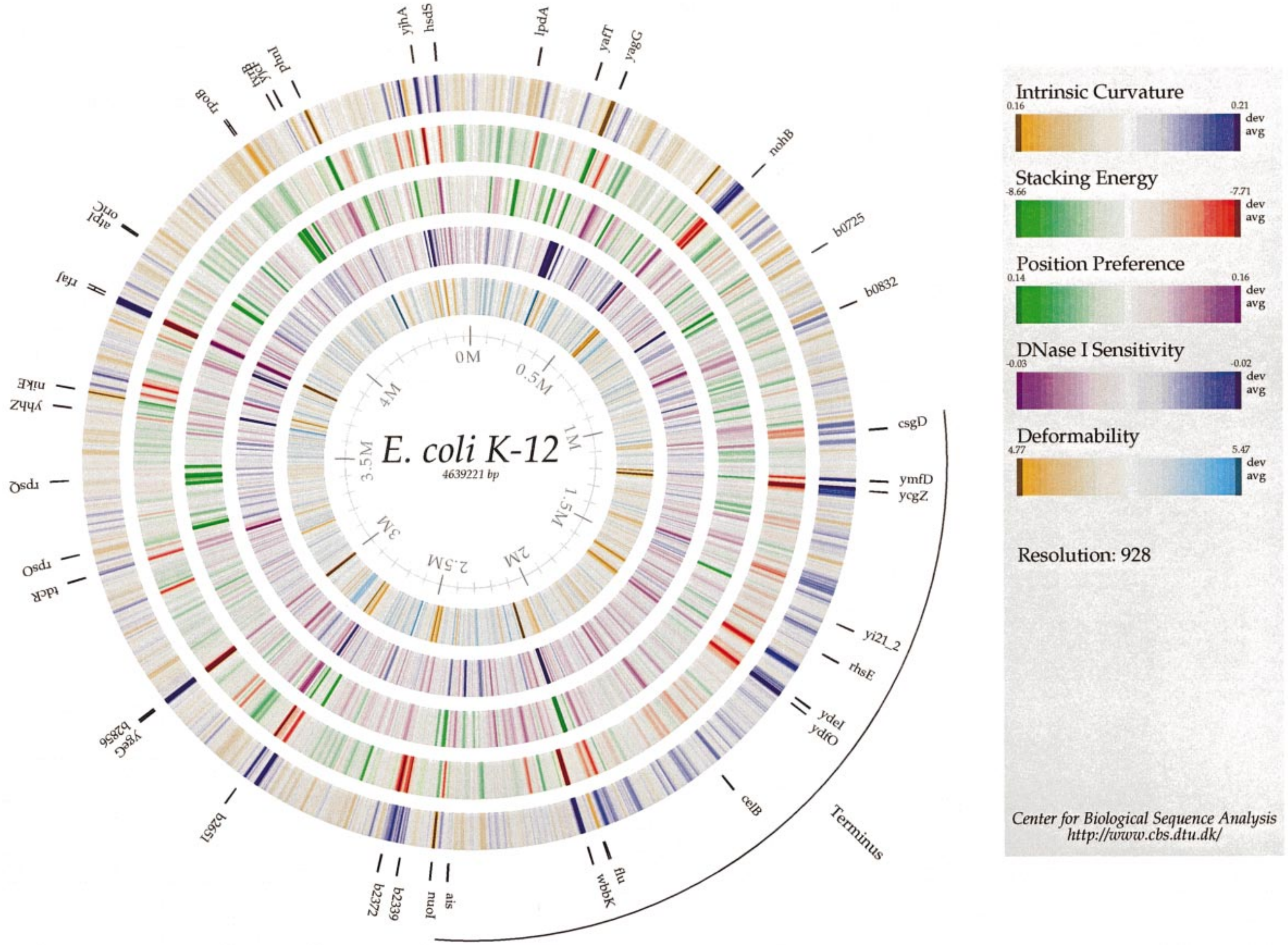


Figure 1 (*legend opposite*)

Table 1. Regions of extreme structure in the *E. coli* genome

Pos. (kbp)	Nearby genes				Class	Prom.
238-239	yafT	(+)	yafU	(−)	Ig	−
872-873	b0832	(+)	b0833	(+)	Ext[a]	+
1102-1103	csgD	(−)	csgB	(+)	Ig	+
1196-1197	ymfD	(−)	ymfE	(−)	CDS[b]	+
1528-1529	rhsE	(+)	ydcD	(+)	CDS[c]	+
1622-1623	ydeI	(−)	ydeJ	(+)	Ig	+
1819-1820	celB	(−)	celA	(−)	Ig	−
2102-2103	wbbK	(−)	wbbj	(−)	Ext[d]	−
2363-2364	ais	(−)	b2253	(+)	Ig[e]	+
2453-2454	b2339	(−)	b2340	(−)	Ig	+
2993-2994	b2856	(−)	b2857	(−)	Ext[f]	+
3265-3266	tdcR	(+)	yhaB	(+)	Ext[g]	+
3580-3581	yhhZ	(+)	yrhA	(+)	Ext[h]	−
3797-3798	rfaZ	(−)	rfaY	(−)	Ext[i]	−
3798-3799	rfaY	(−)	rfaJ	(−)	Ext	−
3802-3803	rfaS	(−)	rfaP	(−)	Ext	−
3920-3921	atpI	(−)	gidB	(−)	Ig	+
4266-4267	tyrB	(+)	aphA	(+)	Ig	+
4280-4281	yjcF	(−)	yjcG	(−)	CDS[j]	+
4578-4579	hsdS	(−)			Ext[k]	−

Pos., the position in the *E. coli* genome of the 1000 bp window (numbers are kbp measured from zero minutes). Nearby genes, the two genes nearest the extreme region (most upstream gene is mentioned first). The sign indicates whether the gene is transcribed in the clockwise (+) or counterclockwise (−) direction. Class, the type of region covered by the extreme window; can be either Ig (mainly intergenic), CDS (mainly coding), or Ext (extended). Prom., indicates whether the extreme region is known to contain a promoter.

[a] Most extreme in intergenic region, but structure high in 872-877 kbp.

[b] *ymfD* and *ymfE* are similar to phage genes. There are three predicted sigma-54 promoters overlapping *ymfE*.

[c] Structure mainly in ydcD. The *rhsE* gene has ''opposite'' structural characteristics (Figure 4, upper panel).

[d] Most extreme where genes overlap, but structure high in 2101-2108 kbp.

[e] Figure 4, lower panel.

[f] High structure in 2985-2994 kbp.

[g] High structure in 3264-3267 kbp, but curvature is most extreme within the *tdcR* gene.

[h] High structure mostly in CDSs between 3579 and 3581 kbp. Another high region between 3582 and 3583 kbp. In the low-structure interval there are two IS-related genes.

[i] High structure in 3795-3806 kbp, mostly within CDSs, and covering most of the two oppositely oriented *rfa*-operons (except for *rfaD*, *rfaF*, and *rfaC*). There is a shorter extreme region further upstream covering *htrL*. Interestingly, *E. coli* cells lacking the HU protein, known to interact with extremely structured DNA, have a phenotype that resembles that of *rfa* mutants (Painbeni *et al*., 1997).

[j] Most structure within *yjcF*, but extending into flanking intergenic regions.

[k] High structure between 4574 and 4579 kbp. Extreme structure within *hsdS*.

the type we predict in this study (Bracco *et al*., 1989; Tanaka *et al*., 1993; Shimizu *et al*., 1995). It is therefore tempting to suggest that the high-curvature (and generally extreme) region could be directly involved in segregation of the newly replicated nucleoids. Second, in *B. subtilis* the terminus region has been found to be attached to the cell membrane in a high-salt resistant manner (Sueoka, 1998) while it is unclear whether the same is true for *E. coli*. We speculate that the structurally extreme properties of the terminus region may potentially be connected to this phenomenon.

Comparison of shuffled and real DNA sequences

Since base composition clearly has an impact on the measures used here, we were interested in investigating to what degree the mono-nucleotide frequencies in *E. coli* explain the distribution of

Figure 1. Structural atlas for the entire *E. coli* chromosome. The value of each measure along the DNA sequence is shown using colored concentric wheels (one for each measure) representing the circular chromosome. The sequence of measures is indicated in the legend at the right (curvature values are plotted in the outermost circle, while stacking energy is in the innermost). Up on the wheels corresponds to zero minutes on the *E. coli* linkage map, with increasing minute positions in the clockwise direction. The numbers on the inside of the innermost wheel is the position relative to zero minutes measured in millions of base-pairs (Mbp). The resolution is 928 bp, meaning that the thinnest visible line in the innermost circle, corresponds to a DNA region of length 928 bp, and the structural values have been smoothed using a running average with a window size of 9280 bp. Also indicated are the locations of the origin of replication (oriC, upper left part of circle, around 4 Mbp) and the terminus region (TerE-TerG, lower right part of circle, around 1.5 Mbp). Finally, we have indicated central genes in regions possessing extreme structure on the edge of the outermost circle. See Table 1 for a more thorough description of some of these regions.

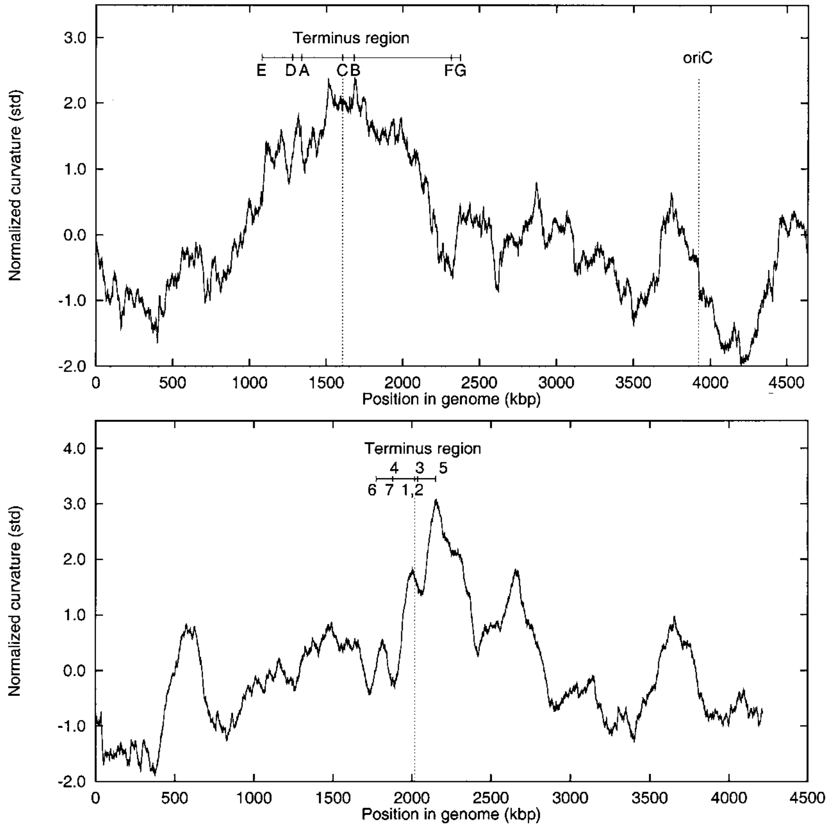


Figure 2. Curvature profiles. The profiles have been smoothed using a running average with a window size of 250 kilo-base-pairs (kbp), and values have furthermore been normalized and are in units of standard deviations from genomic average. Upper panel, curvature profile of the entire *E. coli* chromosome. The location of the origin of replication (oriC) and the terminus region (with terminator sites TerA, TerB, TerC, TerD, TerE, TerF, and TerG) is indicated. Note the broad curvature peak in the terminus region, centered on TerC (which is believed to be the most frequently used terminator site). Lower panel, curvature profile of the entire *B. subtilis* chromosome. The origin of replication is at position zero kbp. The location of the terminus region (with terminator sites TerI-TerVII) is indicated.

predicted structural values. For that purpose we constructed a shuffled version of the entire genome, and subsequently calculated the values of all five structural measures at all positions. Comparison of the distribution of values (using 1000 bp averaging windows) in the real and shuffled genomes, showed that in all cases the distributions were very significantly different (see Figure 3). In the case of curvature, protein-induced deformability, and stacking energy, the shuffled distributions were much narrower than the real distributions, meaning that the real biological DNA displayed significantly more extreme values in both directions than did the shuffled DNA. For DNaseI and nucleosome position preference values, however, the bulk of the shuffled distribution was shifted to one side, so that biological DNA, on average, is predicted to be less flexible (more rigid) than expected from the mononucleotide frequencies. In both these cases the most extremely flexible values of shuffled and real DNA are nevertheless of comparable size, see Figure 3. Thus, the real biological DNA is predicted to have significantly different (and generally more extreme) structural properties than would be expected from base composition alone. This observation is consistent with the hypothesis that DNA structure is one of the driving forces behind the evolution of chromosome sequence in *E. coli.*

In order to investigate to what degree the predicted structural characteristics are determined by sequence constraints present in coding DNA, we investigated the effect of shuffling within protein-encoding genes (CDSs). As for the entire genomic sequence, we observe significantly different distributions for the real DNA and the mono-nucleotide shuffled CDSs, in agreement with previous work (Jauregui *et al.*, 1998). However, we also observe significant differences when comparing the biological DNA, with DNA in which we have shuffled entire codons, thus maintaining the codon usage (data not shown). Specifically, we find that the median curvature of codon-shuffled DNA lies approximately mid-way between the median curvatures of mononucleotide shuffled and of biological DNA (data not shown). The same general relationship is also true for the remaining four structural parameters. We therefore conclude that although codon usage clearly has an effect on the predicted structural characteristics in coding DNA, it only accounts for about half of the observed bias (compared to what is expected from base composition), while the remainder is a consequence of sequence patterns that lie across codon-codon boundaries. This observation is also consistent with the idea that DNA structure is a driving force behind the evolution of the nucleotide sequence in the *E. coli* chromosome.

Structurally extreme regions

In order to identify regions with extreme structural characteristics, we used five differently shuffled versions of the *E. coli* chromosome to esti-

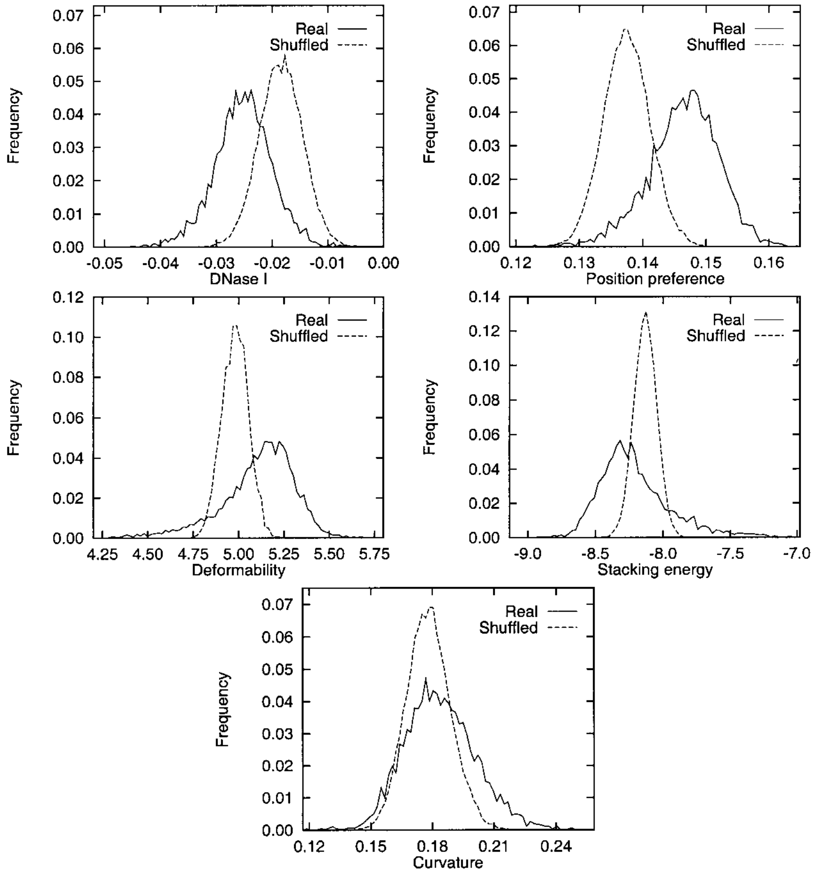


Figure 3. Distribution of structural parameters in real and shuffled DNA. The distribution of values was calculated for all five measures in 1000 bp non-overlapping windows in the entire *E. coli* genome and in a shuffled version of the same. Note that in the case of curvature, protein-induced deformability, and stacking energy, the shuffled distributions are much narrower than the real distributions, meaning that the real biological DNA displays significantly more extreme values in both directions compared to the shuffled DNA. For DNaseI and nucleosome position preference values, however, the bulk of the shuffled distribution is shifted to one side, so that biological DNA on average is predicted to be less flexible (more rigid) than expected from the mononucleotide frequencies (i.e. the range of values in the biological DNA is wider than that of shuffled DNA).

mate the most extreme structural values likely to be observed by chance (see Methods). These values can be thought of as (very strict) thresholds for when an observed value is significantly more extreme than expected from the nucleotide composition alone. We then selected the 1000 bp regions in the real genome where all five measures simultaneously exceed the thresholds defined on the basis of the shuffled data. This resulted in a list of 20 regions predicted to be extreme by all five structural measures (Table 1). In all these, the structural measures are extreme in the same direction (this is mainly caused by the fact that in the biological DNA, one end of the DNaseI and stacking energy distributions extend to approximately the same values as in the shuffled DNA, see Figure 3). Specifically, the structural features observed in all 20 regions are: high curvature, high stacking energy, high position preference (corresponding to rigid DNA), low DNaseI sensitivity (rigid DNA), and low deformability (rigid DNA). Thus, the measures agree that the DNA in these extreme regions is significantly more curved, less stable, and less flexible than the genomic average. We speculate that these regions may function in delimiting topological domains in the *E. coli* chromosome (see below).

The regions found can roughly be divided into three different classes depending on the location and extent of the extreme structure: (1) structure is mainly in intergenic region (referred to as Ig in Table 1); (2) structure is mainly in coding region(s) (CDS); and (3) structure covers an extended region including both coding and intergenic DNA (Ext). In the Table it is also indicated whether the extreme region includes a promoter. Of the 20 selected regions, eight belong to the mainly intergenic class, three belong to the mainly CDS class, and nine belong to the extended class. Furthermore, 12 of the 20 regions (60 %) contain promoters (for comparison, 47 % of all 1000 bp windows contain promoter sequence). Several features specific to individual regions are commented upon in the footnotes of Table 1.

Figure 4, upper panel, shows a close-up of the extended extreme region adjacent to the *rhsE* gene (position of window: 1528-1529 kbp.) It should be emphasized that this plot represents a short, non-circular region of the chromosome, as indicated by the gap at the top of the circles. Also shown in this plot are the positions of annotated genes (third tier of circles; CDSs are shown as colored boxes; the direction of translation is indicated by the shade of the box). Note the characteristic pattern where the core of the *rhsE* element displays levels of low curvature, low stacking, and high flexibility, while the adjoining region has exactly opposite characteristics. Closer inspection of the genome atlas revealed the presence of several other similarly structured *rhs*-containing regions. The *rhs* elements

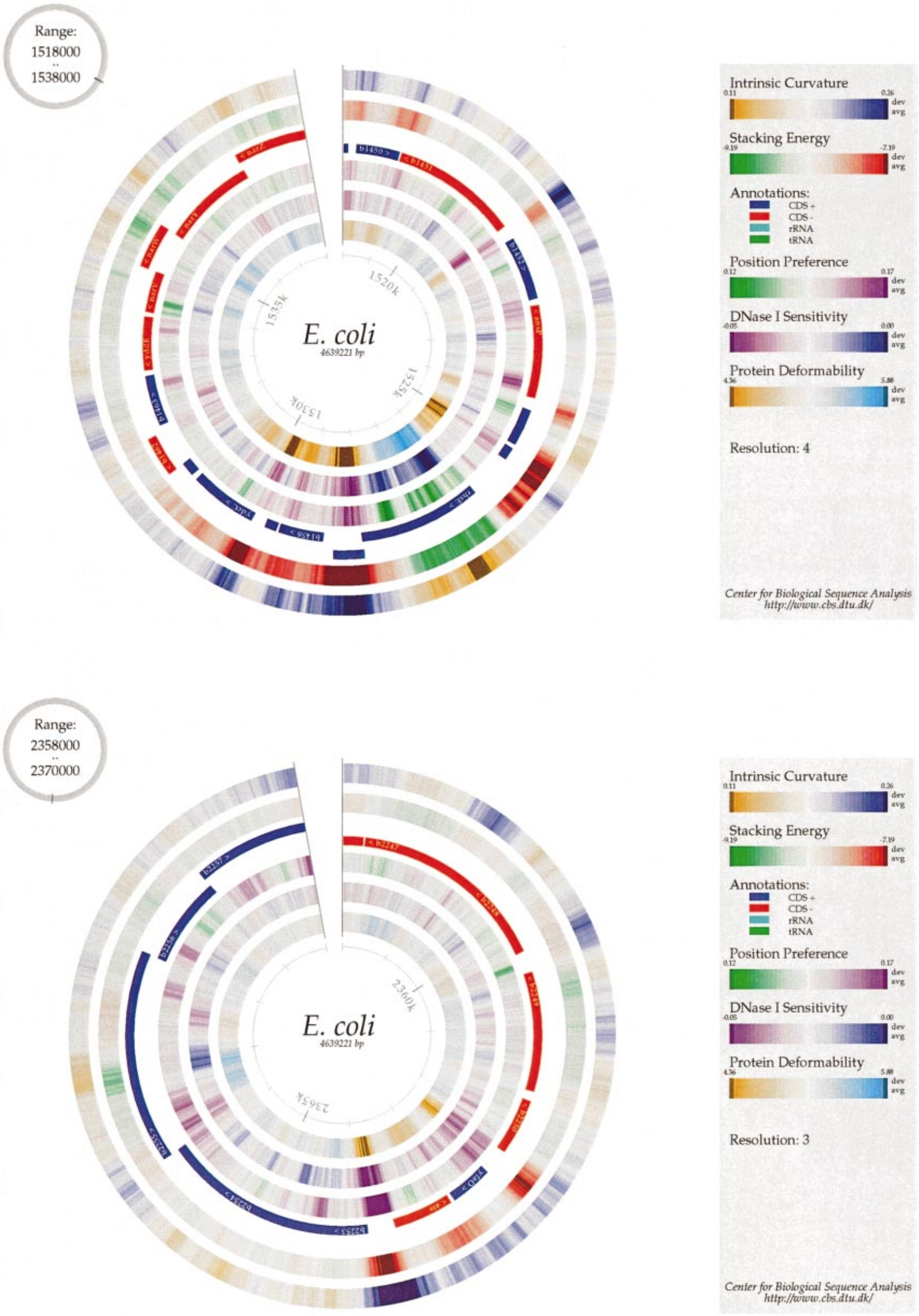


Figure 4. Structural atlas: this plot represents a short, non-circular region of the chromosome, as indicated by the gap at the top of the circles. Also shown are the positions of annotated genes (third tier of circles; CDSs are shown as colored boxes; the direction of translation is indicated by the shade of the box). Upper panel, close-up of the *rhsE* region. Lower panel, close-up of the region near 2363 kbp containing typical intergenic structure.

are large repeats with unusual base composition (Hill, 1999), which makes them stand out in the DNA structural atlases. Although the *rhs* family was first described 15 years ago (Lin *et al.*, 1984), the function of these elements is still unknown, and neither the putative 300 kDa protein or the

7000 nt mRNA have been detected in *E. coli* (Hill *et al.*, 1994). The number of *rhs* genes varies within different strains of *E. coli* and, although *E. coli* strain K12 contains five *rhs* elements, some strains contain no *rhs* element, and exhibit no evident phenotypic change under normal growth conditions (Sadosky *et al.*, 1991; Wang *et al.*, 1998).

Topological domains

DNA in the *E. coli* chromosome is negatively supercoiled *in vivo*, and is segregated into separate domains of supercoiling (Worcel & Burgi, 1972; Pettijohn, 1996). Based on an analysis of the number of nicks that are needed to relax supercoiling fully in the chromosome, it has been estimated that there are 43(±10) such topological domains in growing *E. coli* cells (Sinden & Pettijohn, 1981; Sinden & Ussery, 1992). Microscopy of the isolated *E. coli* nucleoid has shown that it is present in the form of a rosette with about 20-50 long loops emanating from a dense node of DNA (Pettijohn, 1996; Hinnebusch & Bendich, 1997). This is consistent with the idea that each loop corresponds to a single topological domain. It is not known what constitutes the domain borders *in vivo*, but investigations based on a transposon-derived, site-specific recombination system have indicated that the supercoil barriers are stationary and stochastic, i.e. the number and/or position of barriers vary from cell to cell, and possibly also over time (Higgins, 1996; Staczek & Higgins, 1998). For some plasmids it has been found that simultaneous transcription, translation, and secretion of membrane proteins can act to anchor the DNA polymerase to the membrane, thereby restricting the rotation of the DNA template and potentially resulting in formation of temporary supercoil barriers (Lynch & Wang, 1993). However, since it has been reported that rifampicin treatment of cells has no effect on domain structure in the *E. coli* chromosome (Sinden & Pettijohn, 1981), it seems unlikely that a similar mechanism is responsible for creating the nucleoid domain boundaries.

Investigations of the level of supercoiling at defined locations in the chromosomes of *E. coli* and *Salmonella typhimurium*, have indicated that under normal circumstances, all domains have very similar levels of supercoiling (Miller & Simmons, 1993; Pavitt & Higgins, 1993). This suggests that the domain structure does not have an important role in differential regulation of gene expression. However, supercoil barriers may nevertheless be important for biological processes, by limiting the potentially adverse effects of topology-changing processes (e.g. recombination, repair, and replication) to the domain in which they take place.

Prokaryotic supercoils seem to exist largely in the form of long rods consisting of interwound double helices (so-called plectonemic supercoils) (Bliska & Cozzarelli, 1987). It has been found that inherently curved DNA is able uniquely to orient supercoils of this type in a way such that the apex of the rod is positioned in the curved DNA region (Laundon & Griffith, 1988). Based on these observations, it seems likely that the regions of extreme structure listed above (Table 1) will be strong organizers of plectonemically supercoiled DNA. It is probably reasonable to assume that the energetic benefit of positioning curved DNA at an apex, will be proportional to the degree of curvature. This, in turn would mean that the more curved a piece of DNA is, the more the equilibrium will be shifted towards apical positioning of the curve, which is equivalent to saying that the curved DNA will spend more time in this position. As long as the apex of a plectonemic supercoil is fixed, slithering of the interwound DNA is also restrained. We therefore suggest that the highly curved (and generally extreme) regions listed above may function as domain boundaries: since the listed regions are highly curved, they are likely to be positioned at apices a large fraction of the time, thereby restraining the diffusion of supercoils. Furthermore, the observed stochastic properties of domain boundaries fits nicely with the dynamic nature of the equilibrium outlined above. Consistent with this hypothesis, the number of very extreme regions found in this study (20, of which 19 are not immediate neighbors) is in the same order of magnitude as the number of domain boundaries (20-50). It should be noted that according to this hypothesis the 20 regions mentioned above are by no means the only domain boundaries in the *E. coli* chromosome. In fact, any curved piece of DNA will function as a domain boundary some of the time. However, the 20 regions will probably be in a conformation where they restrict supercoil diffusion for a relatively large fraction of the time, perhaps making it possible to detect the effect experimentally.

Analysis of structural features in promoters, intergenic regions, and coding DNA

From browsing through close-up views of the genome atlas, it quickly became apparent that intergenic regions generally appear to have distinct structural features. A typical example is shown in Figure 4, lower panel, where intergenic regions with non-average structural properties are clearly visible. In order to investigate the extent of this phenomenon, we performed a statistical analysis of structural properties in several different classes of DNA. Specifically, the classes were: protein-encoding DNA (CDSs), entire intergenic regions with promoters, entire intergenic regions that do not contain promoters, sigma-70 promoters, and sigma-54 promoters.

Statistical analysis of differences between promoters, intergenic regions, and protein-encoding DNA

The first step in analyzing whether there are structural differences between the different sets of sequences was to determine the distribution of all five measures for each of the five sequence classes. Specifically, the distributions were calculated for all non-overlapping 30 bp windows in each sequence class (see Methods). The five sets of sequences were selected using annotation from the GenBank file. Figure 5 shows a typical example of the resulting histograms. In this plot the (clearly different) distributions of stacking energy in CDSs, intergenic regions, and intergenic regions with promoters, can be seen.

We then compared the distributions of all five structural measures in the five classes, using the Kolmogorov-Smirnov two-sample test (Young, 1977). Table 2 shows a summary of the results for stacking energy distributions. As it can be seen, the three distributions of stacking energies depicted in Figure 5 are all significantly different. The same is true for most other combinations, except that the distribution of stacking energy in sigma-54 promoters is not significantly different from that observed within coding DNA. The sigma-54 promoters are, however, very significantly different from sigma-70 promoters, in accordance with the fact that sigma-54 is quite different from the sigma-70 family, both structurally and functionally (Gross *et al.*, 1992; Merrick, 1993). It should be noted that practically all of the sigma-54 sequences annotated in the GenBank file are predicted, and the credibility of this result therefore depends strongly on the quality of those predictions. As can be seen from Figure 5, the differences are such that coding DNA (and hence also sigma-54 promoters) have the most negative stacking energy values (corresponding to more stable DNA), while intergenic regions are less stable, and intergenic regions with promoters, less stable still. The latter is in agreement with the need for opening up the double helix in promoters prior to initiation of transcription.

Comparison of the five classes for the remaining measures, showed that in all cases there is a significant difference between CDSs, intergenic regions with promoters, and intergenic regions without promoters (data not shown). This difference is much less pronounced in the case of DNaseI sensitivity and position preference, but in all cases the differences are significant with $p < 0.01$, and the three flexibility measures always agree on the direction of the difference. The general picture that emerged from this analysis is that intergenic regions with promoters are generally more curved, less flexible and less stable than coding DNA. Intergenic regions without promoters occupy a position between these other two classes (data not shown, but see Figure 6). The two promoter classes were always found to be very significantly different for all measures. As was the case for stacking energy, sigma-54 promoters were not significantly different from coding DNA in all other measures except position preference, according to which they were predicted to be slightly less flexible than coding DNA.

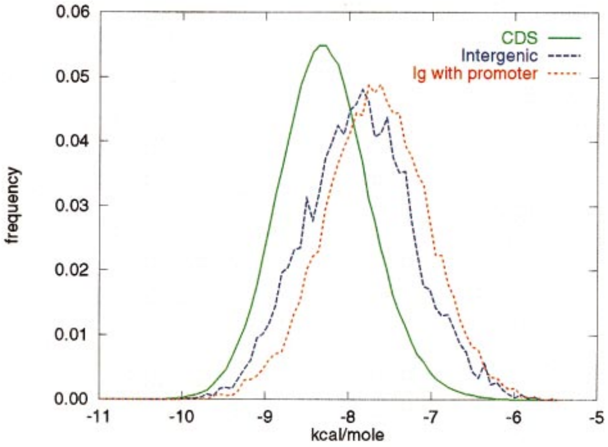


Figure 5. Distribution of stacking energies in CDSs, intergenic regions without promoters, and intergenic regions with promoters.

Promoter structural profiles

To analyze further structural features in promoters, we constructed average structural profiles for the four major sigma classes (sigma-32, sigma-38, sigma-54, and sigma-70). Briefly, these were constructed by aligning promoters of a given class at the transcriptional start point, and subsequently calculating the average of all structural parameters at every position in the alignment. Figure 6 shows the resulting plots which have been normalized based on the genomic average (corresponding to zero on the y-axis in the plot). Note that in these profiles we have inverted the sign of the position preference measure, so that high values mean flexible DNA (similar to the other two flexibility measures).

The sigma-70 profile shows a large region of non-average structure centered a little upstream of the transcriptional start point, and extending into the transcribed DNA. The structured region is much wider than the region where sequence is conserved and, in agreement with the general trends found above, is predicted to be more curved, less stable, and less flexible than the genomic average. A profile constructed from a set of well-documented promoters with mapped transcriptional start points (Lisser & Margalit, 1993), was almost identical with this one (data not shown). According to the two trinucleotide-based flexibility models (DNaseI and position preference), it is mainly the region upstream of the transcriptional start point that is rigid, while the downstream region is predicted to have about average flexibility. This is reminiscent of the situation in eukaryotic promoters, where we have found that the region downstream of the transcription start is more flexible than the region upstream (Pedersen *et al.*, 1998). However, the ‘‘deformability’’ measure does not agree on this feature, but instead predicts that the entire structured region is

Table 2. Summary of Kolmogorov-Smirnov tests for differences between distributions of stacking energy in different classes of DNA

	CDS	sigma-70	sigma-54	IG with prom.	IG no prom.
CDS		0.33	0.06	0.42	0.29
sigma-70	+		0.30	0.09	0.05
sigma-54	−	+		0.38	0.25
IG with prom.	+	+	+		0.13
IG no prom.	+	+	+	+	

CDS, protein-encoding DNA sequences; sigma-70, sigma-70, promoters (predicted and documented), sigma-54, sigma-54 promoters (predicted and documented); IG with prom., intergenic regions that contain predicted or documented promoters; IG no prom, intergenic region that does not contain predicted or documented promoters. In the upper triangle the Kolmogorov-Smirnov statistic is given. This is essentially a measure of the distance between the distributions. In the lower triangle is an indication of whether the difference is significant with $p < 0.001$ (+). Note that the sigma-54 distribution was found not to be significantly different from the CDS distribution ($p > 0.05$, −).

quite rigid. In this context, it should be noted that a structural profile shows the general, average picture. Hence, if every member in a set of promoters contains a single highly flexible region that is located at different positions in every individual promoter, then this will not be visible in the profile.

Our results are consistent with several experimental studies demononstrating the presence of curved DNA upstream of genes in *E. coli* (for reviews, see Perez-Martin *et al.*, 1994; Perez-Martin & de Lorenzo, 1997). They are also in agreement with investigations involving random cloning of curved DNA from *E. coli*, which demonstrated that most of the curved fragments are located immediately upstream of genes and furthermore contain promoters (Mizuno, 1987; Tanaka *et al.*, 1991). The center of curvature found in this study (around −40) and the presence of upstream curvature also agrees quite well with previous theoretical work (Plaskon & Wartell, 1987; Wye *et al.*, 1991; Gabrielian & Bolshoy, 1999). Our prediction that, on average, promoter DNA is less stable than DNA in most other regions, is consistent with the fact that the DNA helix in promoters is melted during initiation of transcription, and is furthermore in accordance with computational results (Lisser & Margalit, 1994). However, in the same study it was found that promoters are generally more flexible than random DNA, which is in contradiction to our observations; we find that promoters are more rigid than the genomic average, which is more rigid than random DNA. One important reason for this apparent discrepancy is the fact that the flexibility model used by Lisser and Margalit (Sarai *et al.*, 1989) deals with flexibility along the twist coordinate (corresponding to rotation of the double helix around its center), while the type of flexibility discussed here is more closely connected with base-pair roll (corresponding to bending of the helix backbone; see Sinden *et al.* (1998) for definitions of DNA helical characteristics). Examination of the underlying models further shows that the discrepancy is essentially caused by the flexibility value assigned to the dinucleotide AT (data not shown). Specifically, the twist-flexibility model assigns the highest flexibility to the dinucleotide step AT, while this is the most rigid step according to the dinucleotide-based roll-flexibility models used here (Hassan & Calladine, 1996; Olson *et al.*, 1998). AT is also predicted to be quite rigid by dinucleotide versions of the triplet-based measures used in this study (Satchwell *et al.*, 1986; Brukner *et al.*, 1995b). Since all four experimentally based models used here agree on this feature, we believe that our conclusion is correct as far as flexibility in the roll-direction (corresponding to backbone bending) is concerned.

Ozoline *et al.* (1999) demonstrated that *E. coli* promoters contain the highly deformable dinucleotide TA positioned with a weak periodicity of approximately 5.6 base-pairs. It was suggested that this could indicate macroscopic flexibility of the promoters, since 5.6 is approximately half of the helical repeat of *B*-DNA. In this context it would be interesting to know how strong the periodic positioning of TA is outside of promoters relative to within.

There is mounting evidence that during transcriptional initiation in *E. coli*, DNA flanking the transcriptional start point is wrapped nearly one full turn around the RNA polymerase (Amouyal & Buc, 1987; Schickor *et al.*, 1990; Craig *et al.*, 1995; Nickerson & Achberger, 1995; Polyakov *et al.*, 1995; Rivetti *et al.*, 1999). The highly curved region that we predict could easily be imagined to be involved in this process. Furthermore, in agreement with our prediction, which indicates that the structured region extends into the transcribed DNA, it has been shown that about one-third of the 90 bp of DNA involved in wrapping around RNAP are downstream of the transcriptional start point (Craig *et al.*, 1995; Rivetti *et al.*, 1999).

Taking into account that the sigma-32 and sigma-38 profiles are based on much fewer sequences, and are therefore more noisy, they look remarkably similar to the profile of sigma-70 (Figure 6). Thus, there is an extended region of non-average structure centered upstream of the transcriptional start point, and again the overall trend is towards the promoters being more curved, less stable and more rigid than the genomic average. From these plots, it is not clear whether the

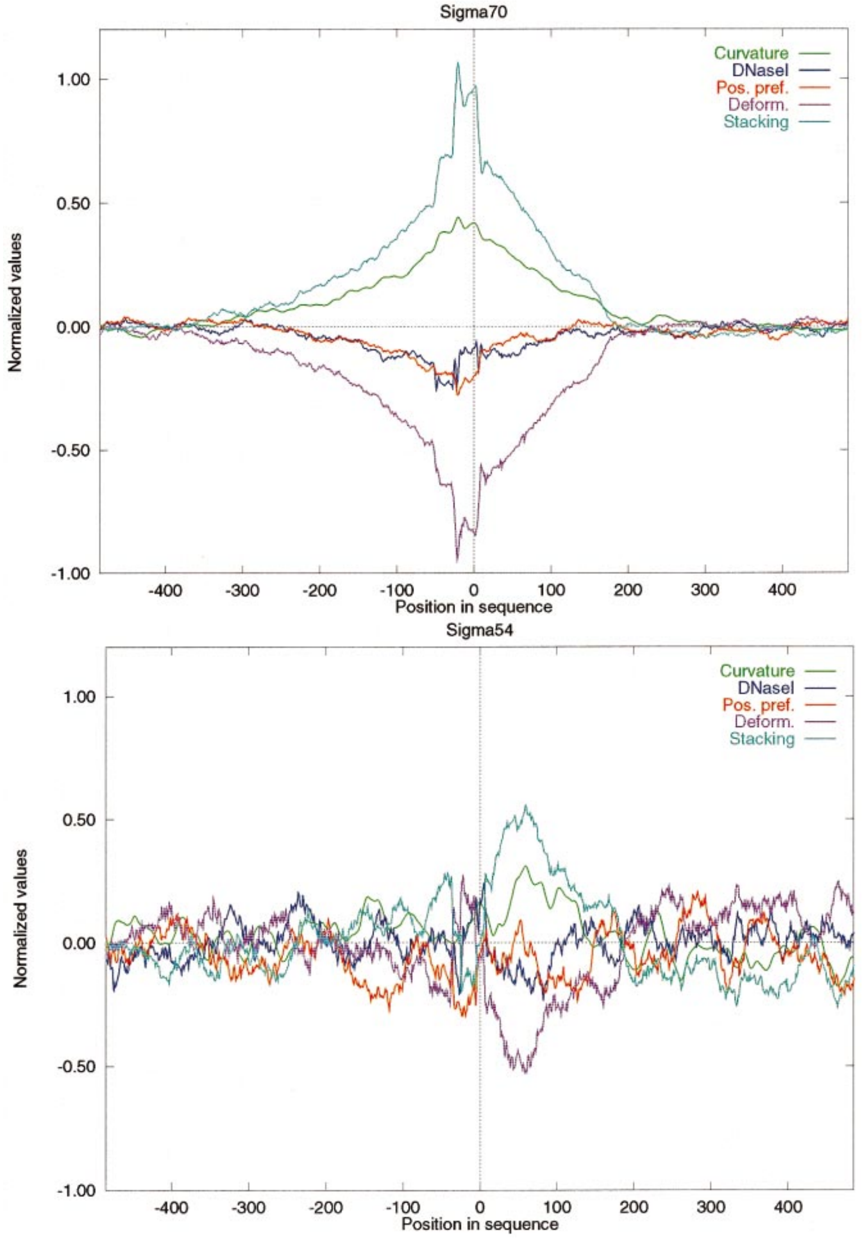


Figure 6 (*legend opposite*)

sigma-38 (sigma-S) promoters are significantly more curved than the other classes, as it has been reported (Espinosa-Urgel & Tormo, 1993; Espinosa-Urgel *et al*., 1996). It is, however, possible that there is a qualitative difference in the exact location of local curvature maxima between the classes.

The sigma-54 profile, on the other hand, is very different from the other profiles, in agreement with the statistical analysis above. Specifically, the entire region upstream of transcription start has an approximately average structure, while the region immediately downstream displays characteristics that are more typical of the other promoter profiles (curved, unstable, and rigid DNA). Again, it should be emphasized that since the profile is based on mostly predicted sigma-54 promoters, the credibility of this result depends strongly on the quality of those predictions. Consistent with this finding, and as mentioned above, sigma-54 is known to be structurally and functionally unrelated to sigma-70. Sigma-54 promoters generally require upstream activators for full activity (Gross *et al*., 1992; Merrick, 1993), and in many cases the activator and RNA polymerase are brought into close contact assisted by the protein IHF binding and bending the DNA between them (Hoover *et al*., 1990; Carmona *et al*., 1997). In addition to helping

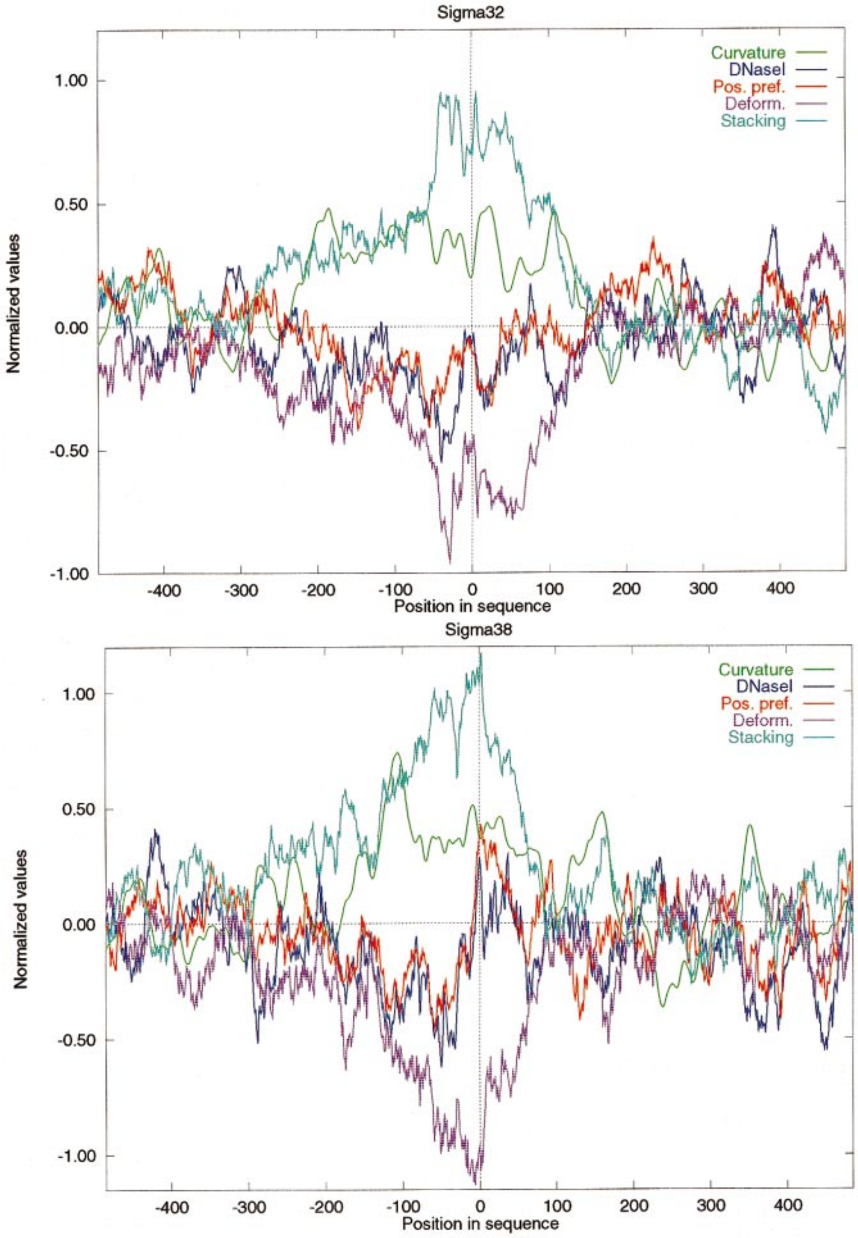


Figure 6. Structural profiles of the four major sigma-classes. The sign of position preference has been reversed, so that it is oriented in the same direction as the other flexibility measures. The profiles were calculated from promoter sequences aligned at the transcription start. Furthermore, the plots have been smoothed using a running average with window size 31 bp and normalized based on the genomic average and standard deviation.

establish the correct promoter architecture with activator and RNA polymerase in direct contact, this also seems to have the effect of disfavoring interactions between the polymerase and heterologous activators bound to other sites on the DNA or from solution (Perez-Martin & de Lorenzo, 1995, 1997). We speculate that the putative lack of curvature in sigma-54 promoters may be related to an evolutionary pressure for maintaining this regulatory system, which could potentially be short-circuited by a large, generally curved, region.

Structural features in other prokaryotic genomes

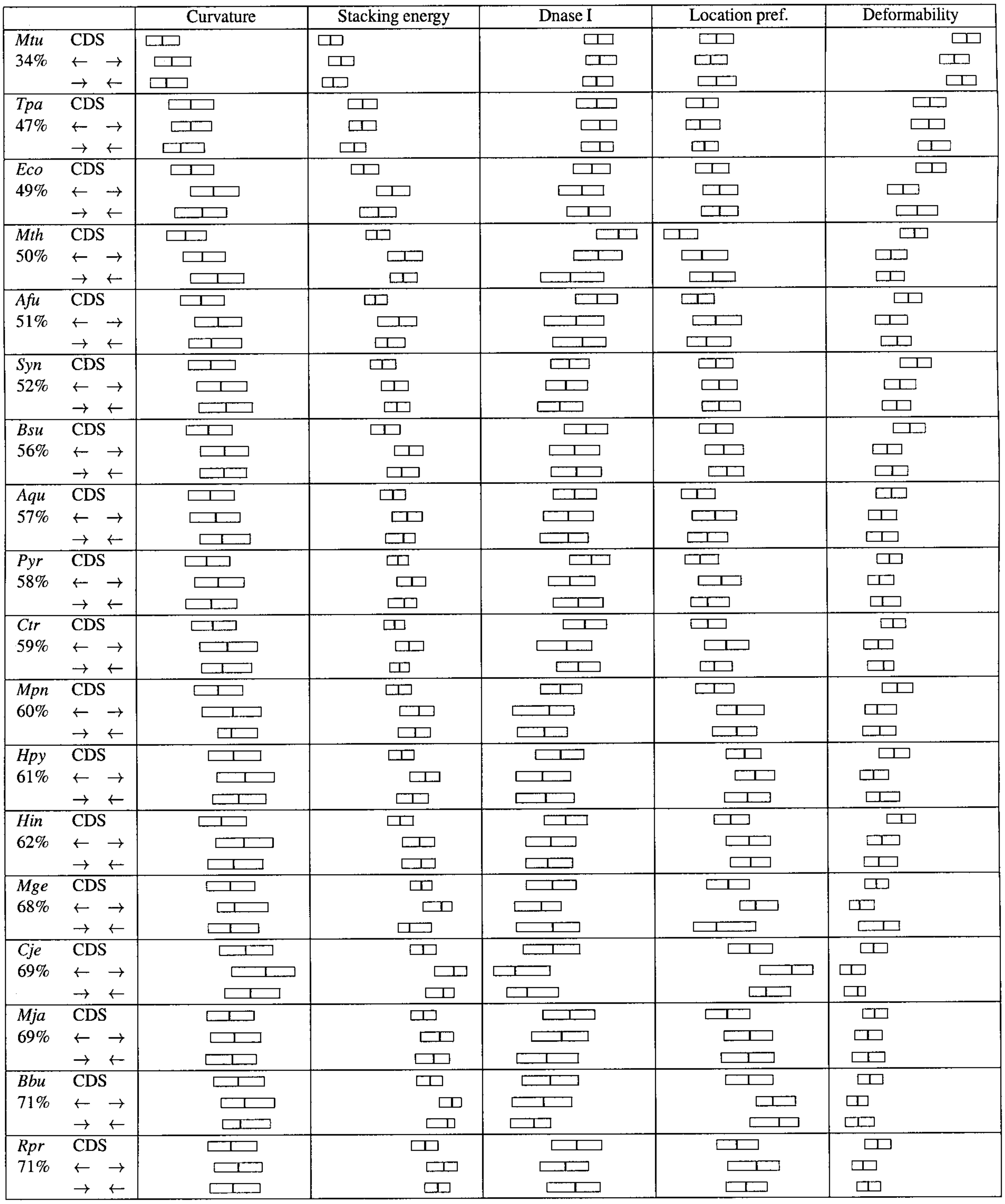
In order to investigate the general applicability of our findings, we were interested in evaluating the structural features in other genomes besides that of *E. coli*. For this purpose we collected a set of 18 prokaryotic chromosomes, including both eubacterial and archaeal examples. However, although there is a wealth of fully sequenced genomes, many of these are not annotated to the same degree of detail as is the case for *E. coli*. In order to avoid this problem we came up with an indirect approach that we believe can be used to find promoter-containing regions, and that at the same time removes the problem of unequivocally identifying regions which do not contain promoter activity. Specifically, we use the direction of the two genes flanking an intergenic region as an indicator of whether there is a promoter present or not. Prokaryotes generally have very compact genomes with correspondingly short 5′ untranslated regions (5′ UTRs). It should therefore be possible to select a set of intergenic regions that very likely contain promoters by selecting regions where the two flanking genes both point away from the intergenic DNA (henceforth denoted ← →). Using a similar line of reasoning, intergenic regions with two genes pointing towards the region (→ ←), probably do not contain a promoter. In accordance with this idea none of the 306 annotated experimentally verified promoters in the *E. coli* geliome (GenBank file, version M54) is present in regions of type → ←. Furthermore, based on the annotation in the GenBank file, 96 % of all regions of type - ← → do contain promoters, while 99.6 % of regions of type → ← do not. It should be noted that this approach assumes that genes are correctly annotated. Furthermore, both protein-encoding genes and RNA-genes have to be included for the method to work.

Using this approach on our set of 18 prokaryotic genomes, we selected three different types of region in each chromosome: (1) protein encoding genes (CDSs); (2) intergenic regions that probably contain a promoter (← →); and (3) intergenic regions that probably do not contain a promoter (→ ←). We then performed an analysis of the distribution of structural values in 30 bp windows, similar to that performed for *E. coli* described above. The results of this analysis are summarized in Table 3 in the form of box plots. Specifically, we have indicated for each of the three classes and for each measure the location of the median of the distribution (central, thick, vertical bar in boxes), as well as the location of the 15th and 85th percentiles (narrower, vertical lines at the ends of boxes). The scale of the *x*-axis in each column is indicated in the legend. The percentiles were chosen so that they correspond approximately to plus/minus one standard deviation for normal distributions. The reason why we report medians and percentiles is that some measures (especially of curvature) follow skewed distributions, and thus the average is not a good indicator. As it can be seen, the absolute values of the different measures vary widely between the different organisms (see for instance, stacking energy in *M. tuberculosis*, *Mtu* and *B. burgdorferi*, *Bbu*). Nevertheless, the general trends that we noted for *E. coli* hold true for practically all other organisms (both archaeal and eubacterial) and for all measures. Thus, in both of the eubacteria mentioned above, it can be seen that promoter-containing regions (← →) are more curved, less stable, and less flexible than CDSs, while intergenic regions without promoters (→ ←) have intermediate values (Table 3). Furthermore, this is true even though the most curved region in some organisms (e.g. *M. tuberculosis*) is less curved than the least curved region in others (e.g. *B. burgdorferi*). The clearest exceptions to this general picture are *Methanobacterium thermoautotrophicum* (*Mte*, archaea), *Synechosystis* sp. (*Syn*, eubacterium), and *Treponema pallidum* (*Tpa*, eubacterium). The relative curvature levels in *E. coli*, *Baccillus subtilis*, *H. influenzae*, and *M. genitalium* are consistent with a previous analysis (Gabrielian & Bolshoy, 1999).

General features of promoter DNA in prokaryotes

The fact that promoter-containing DNA is predicted to be more curved and less stable than CDSs and intergenic DNA in practically all the investigated organisms (including both members of archaea and eubacteria), while the absolute levels vary widely between different organisms, suggests that it is the relative level of structure that is important. One biological phenomenon that is consistent with the lower level of stability of promoter DNA is the need for opening up the double helix during initiation of transcription. It is therefore possible that the observed structural characteristics have evolved to ensure the preferential melting of DNA in promoter regions. Another phenomenon that presumably relies more on relative structural characteristics than on absolute values is the organization of plectonemically supercoiled DNA mentioned above. Thus, it seems likely that regions with stronger curvature will be located at the apices in plectonemically supercoiled DNA more often than less curved DNA, regardless of the absolute level of curvature. We therefore suggest that promoters may generally be positioned at terminal loops of interwound supercoiled DNA molecules. Consistent with this hypothesis, it was found in one study that around 95 % of bound RNA polymerases were in fact located at the apices of plectonemically supercoiled DNA (ten Heggeler-Bordier *et al.*, 1992). During transcription, the loop is shifted along the DNA so that the apical position of the polymerase is maintained, a phenomenon which has the advantageous consequence that the nascent RNA chain does not become wrapped around the DNA template (ten Heggeler-Bordier *et al.*, 1992). It is tempting to suggest that such apical positioning of promoters may have an influence on promoter-recognition: since an apical loop in supercoiled DNA always points towards the outside of the circular DNA molecule, it is presumably more accessible for polymerase binding (ten Heggeler-Bordier *et al.*, 1992).

Table 3. Box-plot of structural features for CDSs, intergenic regions with promoters (← →) and intergenic regions without promoters (→ ←), for 18 prokaryotic genomes



For each class and for each measure, the distribution of values is summarized by indicating the median (central bar in box), as well as the 15th and 85th percentiles (narrower bars at end of box). The Table has been sorted according to genomic AT-content, which is also indicated in the first column (%). Scale of columns is as follows. Curvature, 0.114 to 0.329; stacking energy, −9.218 to −6.136 kcal/mol; DNaseI, −0.085 to 0.003; position preference, 0.789 to 0.886; deformability: 3.661 to 5.954. Abbreviations: *Afu, Archaeoglobus fulgidus; Aqu, Aquifex aeolicus; Bbu, Borrelia burgdorferi; Bsu, Bacillus subtilis; Cje, Campylobacter jejuni; Ctr, Chlamydia trachomatis; Eco, Escherichia coli; Hin, Haemophilus influenzae; Hpy, Helicobacter pylori; Mge, Mycoplasma genitalium; Mja, Methanococcus jannaschii; Mpn, Mycoplasma pneumoniae; Mth, Methanobacterium thermoautotrophicum; Mtu, Mycobacterium tuberculosis; pyr, Pyrococcus horikoshii; Rpr, Rickettsia prowazekii; Syn, Synechosystis* sp.; *Tpa, Treponema pallidum*. See Methods for references to original sequence publications.

Structural cluster analysis

In order to investigate whether there is a connection between the predicted structural characteristics of a gene on one hand, and its function on the other hand, we performed a cluster analysis based on the five structural parameters. Briefly, we calculated the average of each measure for all 5000 bp windows centered on RNA or protein-encoding genes, normalized the values (based on the genomic average and standard deviation), and then used the resulting five values to cluster the windows. We thus treat each gene-centered window as a point in a five-dimensional "structure space", and use the Euclidean distance between them as a simple measure of structural similarity (see Methods for more details on clustering). Figure 7 is a distance tree that summarize the overall topology of this space. In this plot, all genes have been divided into 11 clusters and the tree shows the relative position of the mid-points of these (the five coordinates of each centroid, in standard deviation units, is given in the legend to Figure 7). At the base of each branch, the number of genes in that cluster is indicated.

To analyze whether there is a connection between the tree structure and gene functionality, we used word-analysis software that we have developed for investigation of yeast promoters (Jensen & Knudsen, 2000). Briefly, the approach was as follows. From the GenBank file we first collected all the functional annotations for each gene. For each cluster we then divided the annotation into two groups: one containing the annotation for the investigated cluster (the positive set) and another group containing the annotation for all the remaining genes (the negative set). By counting word frequencies in the two sets and using hypergeometric statistics, it was then possible to find annotation keywords that are significantly over-represented in the positive set (Jensen & Knudsen, 2000). On Figure 7 we have indicated the most significant words found in this way.

As it can be seen, several groups did display significant over-representation of keywords. In some cases a cluster contains only a few genes, most of which belong to one or more operons of related function. For example, among the 20 genes in cluster 11, ten belong to the *phn* operon, giving the entire cluster an over-representation of the keywords "phosphonate metabolism". Another example is cluster 4 which contains a number of genes from two different groups of genes that each are involved in lipopolysaccharide synthesis (the *rfa* and *wbb* genes). Cluster 4 is characterized by extreme values for all five measures, and displays a very high level of curvature, very low flexibility and extremely high stacking energy (corresponding to unstable DNA). Cluster 5 is structurally similar to cluster 4 and contains three additional *rfa* genes. Interestingly, *E. coli* strains lacking the HU protein, display a phenotype that resembles the deep-rough phenotype seen in *rfa* mutants (Painbeni *et al.*, 1997). Since HU is known to interact with curved or kinked DNA, and since most of the *rfa* operon displays extreme structural properties (including very high curvature), it is tempting to suggest that the interaction between HU and this chromosomal region plays a role in expression of the genes, although it has been reported that apparently the HU-deficient mutants do not have a truncated lipopolysaccharide (Painbeni *et al.*, 1997).

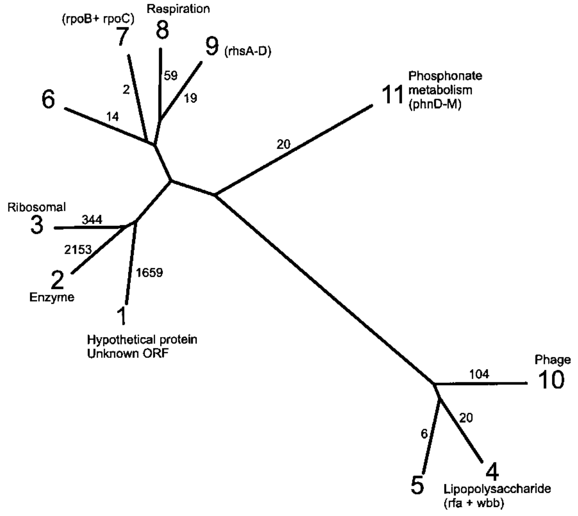


Figure 7. Structural cluster analysis. Distance tree showing the relative location of 11 gene clusters based on average structural measures. The number of genes in each cluster is indicated at the base of the branch. Significantly over-presented annotation keywords are indicated at the end of branches. Names of genes mentioned in the text are indicated in parentheses. The centroid coordinates of the 11 clusters are given below (values are normalized and are in units of standard deviations from genomic average. They are listed in the following order: curvature, DNaseI, position preference, deformability, and stacking energy). Cluster 1: 0.7, −0.6, 0.3, −0.6, 0.6. Cluster 2: −0.6, 0.3, 0.1, 0.6, −0.6. Cluster 3: −0.3, 0.9, −1.6, −0.3, 0.1. Cluster 4: 3.3, −2.4, 2.1, −4.4, 4.6. Cluster 5: 3.0, −3.3, 2.2, −2.8, 3.1. Cluster 6: −0.5, 2.6, −3.3, 0.7, 0.5. Cluster 7: −1.6, 1.1, −4.6, 0.1, −0.8. Cluster 8: −1.7, 2.0, −1.7, 0.8, −0.9. Cluster 9: −2.1, 3.9, −2.8, 0.7, −0.6. Cluster 10: 2.3, −1.3, 0.7, −2.8, 2.8. Cluster 11: −2.9, 2.2, 1.1, 2.7, −2.8.

Cluster 10 is adjacent to clusters 4 and 5, and displays similar (but less extreme) structural features of high curvature, low flexibility, and high stacking energy. Consistently, ten of the 20 extreme 1000 bp regions discussed above (and which displayed the same structural features; Table 1) are included in one of these three clusters. In agreement with the genome-wide trend for higher curvature near the terminus, the genes in cluster 10 show a weak tendency to be located in this region (data not shown). This fact is also consistent with the observation that the keyword "phage" is over-represented in cluster 10 (20 of the 104 genes have the word in their annotation), since the terminus region is known to contain many phage and transposon-related genes (Hill, 1996).

Most of the previously mentioned *rhs* elements are in cluster 9, consistent with the general features displayed by this group (low curvature, and very high flexibility). The one exception is *rhsE* which is in cluster 1. This is presumably related to the fact that in *E. coli* K-12, the *rhsE* gene is truncated and is about half the length of the other *rhs* elements. Other *E. coli* strains have full-length *rhsE* (Hill, 1999). The two genes that encode the beta and beta-prime subunits of RNA polymerase, together make up cluster 7.

Clusters 1, 2, and 3 together contain the bulk of the genes (about 94 %), and consequently have average structural values. Interestingly, the keywords ''hypothetical protein'' and ''unknown ORF'' are over-represented in cluster 1 (657 of the 1659 genes). In cluster 2, the keyword ''enzyme'' is over-represented, indicating that most of the house-keeping genes are in this structurally average cluster. Cluster 3 is interesting in that it only contains 344 genes, but among these are all the ribosomal RNAs, and the majority of the ribosomal proteins. Compared to clusters 1 and 2, the most important difference with cluster 3 is that it has higher DNaseI and lower position preference, both indicative of more flexible DNA. Since the ribosomal genes are among the most highly expressed in growing *E. coli*, cells, it seems possible that the common structural features may play a role in this.

Methods

Data

From the GenBank (Benson *et al.*, 1999) database we downloaded 18 sequenced eubacterial and archaeal genomes. Specifically, the following archaea were included: *Archaeoglobus fulgidus* (Klenk *et al.*, 1997), *Methanococcus jannaschii* (Bult *et al.*, 1996), *Methanobacterium thermoautotrophicum* (Smith *et al.*, 1997), and *Pyrococcus horikoshii* (Kawarabayasi *et al.*, 1998).

The eubacterial genomes were: *Aquifex aeolicus* (Deckert *et al.*, 1998), *Borrelia burgdorferi* (Fraser *et al.*, 1997), *Bacillus subtilis* (Kunst *et al.*, 1997), *Campylobacter jejuni* (Parkhill *et al.*, 2000), *Chlamydia trachomatis* (Stephens *et al.*, 1998), *Escherichia coli* (Blattner *et al.*, 1997), *Haemophilus influenzae* (Fleischmann *et al.*, 1995), *Helicobacter pylori* (Tomb *et al.*, 1997), *Mycoplasma genitalium* (Fraser *et al.*, 1995), *Mycoplasma pneumoniae* (Himmelreich *et al.*, 1996), *Mycobacterium tuberculosis* (Cole *et al.*, 1998), *Rickettsia prowazekii* (Andersson *et al.*, 1998), *Synechosystis* sp. (Kaneko *et al.*, 1996), and *Treponema pallidum* (Fraser *et al.*, 1998).

Structural parameters

We initially selected a set of six different structural models. Based on investigations of correlation between measures (see below) we then chose five of these for this study. The six original models are described below. All six models consist of tables giving structural values for each di- or trinucleotide. For the first five models, prediction of the structural features of any given DNA sequence is done simply by reading along the sequence, and for each position the value for the di- or trinucleotide that the present nucleotide is part of is looked up. For trinucleotides the value was assigned to the central nucleotide, while dinucleotide values were assigned to the second nucleotide. In the case of the sixth model (curvature) dinucleotide parameters are first used to predict 3D coordinates, and the path of the predicted structure is then used to calculate a measure of local curvature at each nucleotide.

DNaseI sensitivity

DNaseI is known to bind preferably and cut DNA that is bent, or bendable, towards the major groove (Lahm & Suck, 1994). Thus, DNaseI cutting frequencies on naked DNA can be interpreted as a quantitative measure of major groove compressibility or anisotropic bendability. Such data have been used to calculate bendability parameters for the 32 complementary trinucleotide pairs (Brukner *et al.*, 1995a).

Nucleosome position preference

From experimental investigations of the positioning of DNA in nucleosomes, it has been found that certain trinucleotides have strong preference for being positioned in phase with the helical repeat. Depending on the exact rotational position, such triplets will have minor grooves facing either towards or away from the nucleosome core (Satchwell *et al.*, 1986). Based on the premise that flexible sequences can occupy any rotational position on nucleosomal DNA, these preference values can be used as measures of DNA flexibility. Hence, in this model, all triplets with close-to-zero preference are assumed to be flexible, while triplets with preference for facing either in or out are taken to be more rigid (Pedersen *et al.*, 1998).

Propeller twist

Based on X-ray crystallography of DNA oligomers it has been found that there is a correlation between the propeller twist angle of a base-pair (i.e. the angle between the planes of the two aromatic bases in the base-pair) and the standard deviation on base-pair ''slide'' (essentially the displacement of a base-pair in the direction perpendicular to the helix axis), as estimated from averaging over dinucleotide parameters in a large set of crystals (Hassan & Calladine, 1996). Since the ability of a base-pair to adopt widely different sidewise positions (and thus to have a large standard deviation in slide) probably can be taken as an indication of local DNA flexibility, the same must therefore be true of the propeller twist angles. Generally, dinucleotides with a small propeller twist angle tend to be more flexible than dinucleotides with a high (more negative) propeller twist angle.

Protein-induced deformability

Protein-induced deformability is a dinucleotide model for how easily DNA is deformed by proteins. The values have been determined by investigating a set of crystal structures of DNA/protein complexes (Olson *et al.*, 1998). The protein-induced deformability value used here is a measure of the size of the conformational space covered by DNA dimers in protein complexes (specifically, we have used the V_{step} parameter).

Stacking energy

Stacking energy can be thought of as the strength with which the planar aromatic bases in adjacent base-pairs interact. All stacking energies are negative, since base stacking is an energetically favorable interaction that stabilizes the double helix. This means that regions with lower (i.e. more negative) stacking energies are strongly stabilized and are therefore less likely to de-stack or melt than regions with higher (less negative) stacking energies. Here we have used a set of dinucleotide values (in kcal/mol) determined by quantum mechanical calculations (Ornstein *et al.*, 1978).

Curvature

Intrinsic curvature is a property of DNA that is closely related to anomalous gel mobility, as DNA fragments with high intrinsic curvature migrate slower on polyacrylamide gels than non-curved fragments of the same length. Here, we have used the CURVATURE program (which is based on a dinucleotide model derived from gel mobility data) for prediction of intrinsic curvature (Bolshoy *et al.*, 1991; Shpigelman *et al.*, 1993). Briefly, the program uses a set of values for the twist, wedge, and direction angles of dinucleotides to calculate the three-dimensional path of the input sequence. The curvature at any given nucleotide in the DNA is then taken to be a value reciprocal to the radius of a 21 bp arc centered at the reference point. Other theoretical models for DNA curvature exist, and these give very similar predictions (Haran *et al.*, 1994; Gabrielian & Bolshoy, 1999).

Correlation of structural parameters

In order to understand whether the six measures mentioned above give independent information about structural features, we performed a thorough analysis of correlation between the scales. To do this, we first calculated the six structural parameters at all positions in a piece of ''random DNA' of the same length as the *E. coli* genome (approx. 4.6 Mbp). We then divided the DNA into non-overlapping fragments of length 31 bp and calculated the average values in each of the fragments. For each of the 15 possible pairs of parameters, we then used the sets of 31 bp averages to calculate Pearson linear correlation coefficients (Table 4). This approach is one way to overcome the problem of comparing di- and trinucleotide-based scales.

The results showed that three of the six structural properties and AT content displayed relatively high levels of correlation. Specifically, the propeller twist and protein-induced deformability scales are directly correlated (a fact which can also be seen from direct comparison of the dinucleotide values (Baldi *et al.*, 1999)), while the stacking energy scale is inversely correlated to propeller twist and protein-induced deformability. All three measures also show correlation to AT content. We observed only very little correlation between curvature and any of the other structural measures, and we also found virtually no correlation between the two trinucleotide-based flexibility measures (DNaseI sensitivity and position preference) in agreement with previous studies (Brukner *et al.*, 1995b; Baldi *et al.*, 1998). The fact that there is a relatively high correlation between the DNA flexibility measures derived from pure oligonucleotide crystals on one hand and from protein-DNA complexes on the other hand, probably indicates that the conformation adopted by DNA bound to protein to a large degree depends on the inherent structural features of the DNA. This is consistent with several investigations of the interaction between DNA-bending proteins and their binding sites (Parvin *et al.*, 1995; Starr *et al.*, 1995; Grove *et al.*, 1996). We generally found that the correlations measured on biological DNA were higher than those measured on random sequence (Table 4, values in parentheses are on biological DNA). Again, this is consistent with the notion that DNA structure is one of the driving forces behind the evolution of nucleotide sequence in the *E. coli* chromosome.

One goal with using multiple models was to obtain a set of independent flexibility predictions that could then be compared. This approach was motivated by the observation that the correlation between many of the existing flexibility models is quite low (Brukner *et al.*,

Table 4. Linear correlation coefficients between structural scales measured on 31 bp fragments in approximately 4.6 Mbp of random DNA

	DNaseI	Curvature	Prop. twist	Deformability	Stacking	Position pref.	AT %
DNaseI		−0.27 (−0.33)	0.37 (0.47)	0.11 (0.25)	−0.18 (−0.30)	−0.21 (−0.33)	−0.15 (−0.28)
Curvature	−0.27 (−0.33)		−0.33 (−0.42)	−0.24 (→0.34)	0.25 (0.36)	0.15 (0.21)	0.26 (0.36)
Prop. twist	0.37 (0.47)	−0.33 (−0.42)		0.80 (0.87)	−0.74 (−0.82)	−0.15 (−0.22)	−0.88 (−0.91)
Deformability	0.11 (0.25)	−0.24 (−0.34)	0.80 (0.87)		−0.80 (−0.86)	0.06 (0.00)	−0.78 (−0.85)
Stacking	−0.18 (−0.30)	0.25 (0.36)	−0.74 (−0.82)	−0.80 (−0.86)		−0.03 (0.04)	0.90 (0.94)
Position pref.	−0.21 (−0.33)	0.15 (0.21)	−0.15 (−0.22)	0.06 (0.00)	−0.03 (0.04)		0.02 (0.09)
AT %	−0.15 (−0.28)	0.26 (0.36)	−0.88 (−0.91)	−0.78 (−0.85)	0.90 (0.94)	0.02 (0.09)	

Values in parentheses are the coefficients measured on 31 bp fragments in the real *E. coli* genome. Note that generally the correlation observed in the biological DNA is higher than that seen in random DNA. Also indicated is the correlation between structural measures and AT-content.

1995b; Baldi *et al*., 1998; Pedersen *et al*., 1998) and was furthermore based on the rationale that if several independent models agree on a predicted feature, then that feature is likely to be true. In this context it should be noted that there are two different aspects of correlation between models. On one hand, if two measures are highly correlated then one would only gain a small amount of extra information by using both models. On the other hand, if two widely different approaches have been used to quantify the relation between DNA sequence and structure, and these approaches have resulted in two highly correlated scales, then this is in itself an indication that the scales are measuring something meaningful and that predictions based on them should therefore be trusted. As a compromise between these two opposing views, we chose to exclude one of the three correlated measures mentioned above. Specifically, we did not use the propeller twist-based model, while retaining the protein-induced deformability model, and the stacking energy scale in our analysis. We thus ended up using three different models of DNA flexibility: the DNaseI sensitivity model, the protein-induced deformability model and the nucleosome position preference model. In addition we used one model for stacking energy and one model for curvature, giving a total of five structural models which were included in the analysis.

We have noticed that the correlation between some measures is significantly higher when estimated from a longer stretch of DNA than when calculated from direct comparison of the scales themselves. For instance, direct comparison of the dinucleotide-based propeller twist and stacking energy scales gives a linear correlation coefficient of −0.29 when calculated from the 16 possible dinucleotides. When calculated from only the ten independent dinucleotides it is slightly higher: −0.30. In previous work we estimated this correlation on the ''trinucleotide level'', by taking the sum of all overlapping dinucleotides, and comparing those values (Baldi *et al*., 1998). This gave a linear correlation coefficient of −0.55 i.e. significantly higher. As can be seen from Table 4, the correlation coefficient calculated from 31 bp averages in 4.6 Mbp of random DNA is higher still: −0.74. We believe this phenomenon is connected to the fact that while the physical reality behind DNA structure probably involves influences from many neighboring base-pairs, then these parameters have been calculated by fitting experimental data to di- or trinucleotide models only. This presumably means that if there is an underlying correlation between different measures, then this will in effect be divided between overlapping sequence fragments. Furthermore, the choice of ''dinucleotide frame'' when fitting data is also arbitrary. For example, if the flexibility at the central A in the sequence CGATC has been estimated by some experimental procedure, then it is possible to tabulate this value as originating from either the dinucleotide GA or the dinucleotide AT. This will further lead to discrepancies when comparing different scales at the dinucleotide level, while significant correlation may be observed at trinucleotide or higher levels.

Construction of the visualization software

We have constructed a computer program (GeneWiz), for visualization of DNA structural features of long DNA sequences (e.g. entire chromosomes). GeneWiz displays the data using various filters and graphical techniques. The plots show structural data as well as annotations for any specified region. The raw data calculations are based on dinucleotide and trinucleotide models, while the annotations often come from GenBank records. The GeneWiz visualization program allows access to various parameters for manipulation of filtering and output features. Output is in postscript for easy cross platform viewing. The output is a wheel-shaped graphic, either of the whole genome or subsections. It allows both for large-scale and finer detailed analysis. Here, the data are color-coded depending on the value. The configuration allows for the display of data from several different data filters (such as box filters, or filters detecting contiguous regions of relatively high or low values in the data sets). We can display the annotations using a series of icons with user-defined colors. This allows for the identification of short or long annotated regions of interest. We designed the package for the specific purpose of examining structural parameters in genomic data, but it is also useful for visualizing other parameters, such as for instance DNA repeats, base composition, and GC-skew (Jensen *et al*., 1999). The program will be made publicly available.

Color scheme

The first step in color-coding a wheel plot is to determine the distribution of structural values in the analyzed piece of DNA. This is done using overlapping windows of the same size as the wheel plot resolution (which is calculated from the size of the analyzed DNA and the size of the plot). The color scheme used here was constructed so that average values are rendered a light gray, while more extreme values are progressively more brightly colored and can be more clearly distinguished. The progression in color from the average to three standard deviations above or below average is linear (in terms of RGB color codes) in three steps. Thus, the color intensity increases very slowly between the average and one standard deviation. Between one and two standard deviations, the rate of increase is five times higher than the above, while the rate is twofold higher still between two and three standard deviations from the genomic average. When a value is more than three standard deviations from the genomic average it is rendered black (or actually, a very dark version of the relevant color).

Structurally extreme regions

In order to identify regions with extreme structure, we first needed to define a set of stringent thresholds for when any given measure can be considered extreme. For this purpose we first constructed five different mononucleotide-shuffled versions of the *E. coli* chromosome. We then divided the shuffled chromosomes into non-overlapping 1000 bp regions, and for each of these regions calculated the averages of the five structural parameters. For each of the five measures and in each of the five shuffled chromosomes, we then noted the largest and smallest values. Thus, in the case of, for instance DNaseI, we ended up with five different largest random values, and five different smallest random values. Finally, we calculated the averages in each end and used these as thresholds for when an observed value is significantly more extreme than what should be expected from base composition alone.

Analysis of structural features in promoters, intergenic regions, and coding DNA

We first constructed five different sets of DNA sequences, based on annotation in the GenBank file. The five classes are: coding DNA (CDSs); entire intergenic regions with promoters; entire intergenic regions that do not contain promoters; sigma-70 promoters; and sigma-54 promoters. In the three sets containing only promoter sequences we selected 60 bp windows that were placed with the downstream end of the window aligned with the downstream end of the −10 box.

For all classes we then divided the sequences into 30 bp long non-overlapping fragments and calculated the average structural values in each of these. The purpose of using a fixed-length window to determine all distributions was to avoid problems with comparing classes that are of different lengths. The variance of any distribution is inversely proportional to the length of the fragments, meaning that shorter fragments, e.g. promoters, will adopt more extreme values by chance. If a region was less than half the window size long (i.e. less than 15 bp) it was not used for the analysis.

For each measure, all pairwise comparisons of the distributions (for all the five classes) were performed using Kolmogorov-Smirnov two-sample statistics (Young, 1977). The results for stacking energy are shown in Table 2. Specifically, this Table lists the Kolmogorov-Smirnov statistic (essentially a measure of the distance between the distributions) along with an indication of whether this difference is significant with $p < 0.001$. For a visualization of the actual distributions see Figure 5.

Promoter profiles

Average promoter profiles were constructed essentially as described (Pedersen *et al*., 1998), except that the profiles were normalized so that all measures would fit on one plot. Briefly, promoter sequences were first aligned at the transcriptional start point, and for each sequence the corresponding structural profiles were then calculated. Subsequently, the average of all these profiles was constructed and smoothed with a 31 bp window. Finally, the resulting five average structural profiles (one for each measure) were normalized so they could be shown on the same plot. This was done by calculating the genomic average and standard deviation of all measures using 31 bp windows, and subsequently normalizing the actual value by subtracting the average and dividing with the standard deviation. Sequences were selected based on annotation in the GenBank file, and, in the case of sigma-38, also based on original literature (Espinosa-Urgel *et al*., 1996; Loewen *et al*., 1998; Hengge-Aronis, 1996; Schellhorn *et al*., 1998).

Cluster analysis

In order to analyze the connection between gene structure and function, we performed a cluster analysis of all RNA and protein-encoding genes in *E. coli*. The structural features of each gene were summarized by calculating the average of the five structural measures in a 5000 bp window centered on the gene. Again, fixed-size windows were used in order to avoid problems with length-dependence of the average values. The five averages were then normalized by, for each measure, subtracting the genomic average (of all 5000 bp windows) and dividing with the genomic standard deviation (using the same window size). This resulted in a set of five normalized coordinates for each window. From these values we calculated all pairwise Euclidean distances, and then performed hierarchical clustering using the UPGMA algorithm included in the PHYLIP package (Felsenstein, 1989). The 11 cluster level was chosen after inspection of several different cluster-levels (using software that we developed specifically for the purpose of interpreting trees as hierarchical clusters).

Acknowledgments

The authors acknowledge helpful discussions with Alex Bolshoy and other people at the Center for Biological Sequence Analysis. The work of all authors was supported by a grant from the Danish National Research Foundation.

References

Amouyal, M. & Buc, H. (1987). Topological unwinding of strong and weak promoters by RNA polymerase. A comparison between the *lac* wild-type and the UV5 sites of *Escherichia coli*. *J. Mol. Biol.* **195**, 795-808.

Andersson, S., Zomorodipour, A., Andersson, J., Sicheritz-Ponten, T., Alsmark, U., Podowski, R., Naslund, A., Eriksson, A., Winkler, H. & Kurland, C. G. (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature,* **396**, 133-140.

Baldi, P., Brunak, S., Chauvin, Y. & Krogh, A. (1996). Naturally occurring nucleosome positioning signals in human exons and introns. *J. Mol. Biol.* **263**, 503-510.

Baldi, P., Chauvin, Y., Brunak, S., Gorodkin, J. & Pedersen, A. G. (1998). Computational applications of DNA structural scales. In *Proceedings of the Sixth Conference on Intelligent Systems for Molecular Biology (ISMB98)*, pp. 35-42, The AAAI Press, Menlo Park, CA.

Baldi, P., Brunak, S., Chauvin, Y. & Pedersen, A. G. (1999). Structural basis for triplet repeat disorders: a computational analysis. *Bioinformatics,* **15**, 918-929.

Benson, D., Boguski, M. S., Lipman, D. J., Ostell, J., BF, O., BA, R. & DL, W. (1999). Genbank. *Nucl. Acids Res.* **27**, 12-17.

Blattner, F., Plunkett, G., 3rd, Bloch, C., Perna, N., Burland, V., Riley, M., Collado-Vides, J., Glasner, J., Rode, C., Mayhew, G., Gregor, J., Davis, N., Kirkpatrick, H., Goeden, M., Rose, D., Mau, B. & Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science,* **277**, 1453-1474.

Bliska, J. B. & Cozzarelli, N. R. (1987). Use of site-specific recombination as a probe of DNA structure and metabolism *in vivo*. *J. Mol. Biol.* **194**, 205-218.

Bolshoy, A., McNamara, P., Harrington, R. E. & Trifonov, E. N. (1991). Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc. Natl Acad. Sci. USA,* **88**, 2312-2316.

Bonnefoy, E., Takahashi, M. & Rouviere-Yaniv, J. (1994). DNA-binding parameters of the HU protein of *Escherichia coli* to cruciform DNA. *J. Mol. Biol.* **242**, 116-129.

Bracco, L., Kotlarz, D., Kolb, A., Diekman, S. & Buc, H. (1989). Synthetic curved DNA sequences can act as transcriptional activators in *Escherichia coli*. *EMBO J.* **8**, 4289-4296.

Brukner, I., Jurukovski, V. & Savic, A. (1990). Sequence-dependent structural variations of DNA revealed by DNase I. *Nucl. Acids Res.* **18**, 891-894.

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

Brukner, I., Sanchez, R., Suck, D. & Pongor, S. (1995b). Trinucleotide models for DNA bending propensity: comparison of models based on DNase I digestion and nucleosome positioning data. *J. Biomol. Struct. Dynam.* **13**, 309-317.

Bult, J., White, O., Olsen, O., Zhou, L., Fleischmann, R., Sutton, G., Blake, J., FitzGerald, L., Clayton, R., Gocayne, J., Kerlavage, A., Dougherty, B., Tomb, J., Adams, M., *et al*. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science,* **273**, 1058-1073.

Bussiere, D. E. & Bastia, D. (1999). Termination of DNA replication of bacterial and plasmid chromosomes. *Mol. Microbiol.* **31**, 1611-1618.

Campbell, A., Mrazek, J. & Karlin, S. (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial dna. *Proc. Natl Acad. Sci. USA,* **96**, 9184-9189.

Carmona, M., Claverie-Martin, F. & Magasanik, B. (1997). DNA bending and the initiation of transcription at σ^{54}-dependent bacterial promoters. *Proc. Natl Acad. Sci. USA,* **94**, 9568-9572.

Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S., Eiglmeier, K., Gas, S., 3rd, C, B., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., *et al*. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature,* **393**, 537-544.

Craig, M. L., Suh, W. C. & Record, M. T. (1995). HO and DNaseI probing of $E\sigma^{70}$ RNA polymerase-λ_{PR} promoter open complexes: Mg^{2+} binding and its structural consequences at the transcription start site. *Biochemistry,* **34**, 15634-15632.

Deckert, G., Warren, P., Gaasterland, T., Young, W., Lenox, A., Graham, D., Overbeek, R., Snead, M., Keller, M., Aujay, M., Huber, R., Feldman, R., Short, J., Olsen, G. & Swanson, R. (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature,* **392**, 353-358.

Espinosa-Urgel, M., Chamizo, C. & Tormo, A. (1996). A consensus structure for σ^{s}-dependent promoters. *Mol. Microbiol.* **21**, 657-659.

Espinosa-Urgel, M. & Tormo, A. (1993). σ^{s}-dependent promoters in *Escherichia coli* are located in DNA regions with intrinsic curvature. *Nucl. Acids Res.* **21**, 3667-3670.

Felsenstein, J. (1989). Phylip - phylogeny inference package (version 3.2). *Cladistics,* **5**, 164-166.

Fleischmann, R. D., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J., *et al*. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* **269**, 496-512.

Fraser, C., Gocayne, J., White, O., Adams, M., Clayton, R., Fleishmann, R., Bult, C., Kerlavage, A., Sutton, G., Kelley, J., Fritchman, J., Weidman, J., Small, K., Sandusky, M., *et al*. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science,* **270**, 397-403.

Fraser, C., Casjens, S., Huang, W., Sutton, G., Clayton, R., Lathigra, R., White, O., Ketchum, K., Dodson, R., Hickey, E., Gwinn, M., Dougherty, B., Tomb, J., Fleischmann, R., *et al*. (1997). Genomic sequence of a lyme disease spirochaete, *Borrelia burgdorferi*. *Nature,* **390**, 580-586.

Fraser, C., Norris, S., Weinstock, G., White, O., Sutton, G., Dodson, R., Gwinn, M., Hickey, E., Clayton, R., Ketchum, K., Sodergren, E., Hardham, J., McLeod, M., Salzberg, S., *et al*. (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science,* **281**, 375-388.

Gabrielian, A. & Bolshoy, A. (1999). Sequence complexity and DNA curvature. *Comput. Chem.* **23**, 263-274.

Gross, C. A., Lonetto, M. & Losick, R. (1992). Bacterial sigma factors. In *Transcriptional Regulation* (McKnight, S. & Yamamoto, Y., eds), pp. 129-176, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Grove, A., Galeone, A., Mayol, L. & Geiduschek, E. P. (1996). Localized DNA flexibility contributes to target site selection by DNA-bending proteins. *J. Mol. Biol.* **260**, 120-125.

Haran, T., Kahn, J. & Crothers, D. (1994). Sequence elements responsible for DNA curvature. *J. Mol. Biol.* **225**, 729-738.

Hassan, M. A. E. & Calladine, C. R. (1996). Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J. Mol. Biol.* **259**, 95-103.

Hengge-Aronis, R. (1996). Regulation of gene expression during entry into stationary phase. In Escherichia coli *and* Salmonella - *Cellular and Molecular Biology* (Neidhardt, F., ed.), 2nd edit., pp. 1497-1512, ASM Press, Washington, DC.

Higgins, N. P. (1996). Surveying a supercoil domain by using the γδ resolution system in *Salmonella typhimurium*. *J. Bacteriol.* **178**, 2825-2835.

Hill, C., Sandt, C. & Vlanzy, D. (1994). Rhs elements of *Escherichia coli*: a family of genetic composites each encoding a large mosaic protein. *Mol. Microbiol.* **12**, 865-871.

Hill, T. M. (1996). Features of the chromosomal terminus region. In Escherichia coli *and* Salmonella typhimurium*: Cellular and Molecular Biology* (Neidhardt, F. C., ed.), 2nd edit., pp. 1602-1614, American Society for Microbiology, Washington, DC.

Hill, C. W. (1999). Large genomic sequence repetitions in bacteria: lessons from rRNA operons and *Rhs* elements. *Res. Microbiol.* **150**, 665-674.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. & Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucl. Acids Res.* **24**, 4420-4449.

Hinnebusch, B. J. & Bendich, A. J. (1997). The bacterial nucleoid visualized by fluorescence microscopy of cells lysed within agarose: comparison of *Escherichia coli* and spirochetes of the genus *Borrelia*. *J. Bacteriol.* **179**, 2228-2237.

Hoover, T. R., Santero, E., Porter, S. & Kustu, S. (1990). Integration host factor stimulates interaction of RNA polymerase with NifA, the transcriptional activator for nitrogen fixation operons. *Cell,* **63**, 11-22.

Horwitz, M. S. Z. & Loeb, L. A. (1990). Structure-function relationships in *Escherichia coli* promoter DNA. *Prog. Nucl. Acid. Res. Mol. Biol.* **38**, 137-164.

Hud, N. V., Sklenar, V. & Feigon, J. (1999). Localization of ammonium ions in the minor groove of DNA duplexes in solution and the origin of DNA A-tract bending. *J. Mol. Biol.* **286**, 651-660.

Hunter, C. A. (1993). Sequence-dependent DNA structure: the role of base stacking interactions. *J. Mol. Biol.* **230**, 1025-1054.

Hunter, C. A. (1996). Sequence-dependent DNA structure. *Bioessays,* **18**, 157-162.

Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M. & Trifonov, E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.* **262**, 129-139.

Iyer, V. & Struhl, K. (1995). Poly (dA:dT), a ubiquitous promoter element that stimulates transcription *via* its intrinsic DNA structure. *EMBO J.* **14**, 2570-2579.

Jaffe, A., Vinella, D. & D'Ari, R. (1997). The *Escherichia coli* histone-like protein HU affects DNA initiation, chromosome partitioning *via* MukB, and cell division *via* MinCDE. *J. Bacteriol.* **179**, 3494-3499.

Jauregui, R., O'Reilly, F., Bolivar, F. & Merino, E. (1998). Relationship between codon usage and sequence-dependent curvature of genomes. *Microb. Comp. Genom.* **3**, 243-253.

Jensen, L. & Knudsen, S. (2000). Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics,* In the press.

Jensen, L. J., Friis, C. & Ussery, D. W. (1999). Three views of microbial genomes. *Res. Microbiol.* **150**, 773-777.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., *et al*. (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**, 109-136.

Karlin, S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.* **1**, 598-610.

Karlin, S., Miazek, J. & Campbell, A. (1998). Codon usages in different gene classes of the Escherichia coli genome. *Mol. Microbiol.* **29**, 12341-1355.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., *et al*. (1998). Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5**, 147-155.

Klenk, H., Clayton, R., Tomb, J., White, O., Nelson, K., Ketchum, K., Dodson, R., Gwinn, M., Hickey, E., Peterson, J., Richardson, D., Kerlavage, A., Graham, D., Kyrpides, N., Fleischmann, R., *et al*. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature,* **390**, 364-370.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A., Alloni, G., Azevedo, V., Bertero, M., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S., *et al*. (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature,* **390**, 249-256.

Lahm, A. & Suck, D. (1991). DNase I-induced DNA conformation: 2 Å structure of a DNase I-octamer complex. *J. Mol. Biol.* **222**, 645-667.

Laundon, C. H. & Griffith, J. D. (1988). Curved helix segments can uniquely orient the topology of supertwisted DNA. *Cell,* **52**, 545-549.

Lin, R., Capage, M. & Hill, C. (1984). A repetitive DNA sequence, rhs, responsible for duplications within the *Escherichia coli* K-12 chromosome. *J. Mol. Biol.* **177**, 1-18.

Lisser, S. & Margalit, H. (1993). Compilation of *E. coli* mRNA promoter sequences. *Nucl. Acids Res.* **21**, 1507-1516.

Lisser, S. & Margalit, H. (1994). Determination of common structural features in *Escherichia coli* by computer analysis. *Eur. J. Biochem.* **223**, 823-830.

Liu, K. & Stein, A. (1997). DNA sequence encodes information for nucleosome array formation. *J. Mol. Biol.* **270**, 559-573.

Lobell, R. B. & Schleif, R. (1991). AraC-DNA looping: orientation and distance dependent loop breaking by cyclic AMP receptor protein. *J. Mol. Biol.* **218**, 45-54.

Loewen, P., Hu, B., Strutinsky, J. & Sparling, R. (1998). Regulation in the rpoS regulon of *Escherichia coli*. *Can. J. Microbiol.* **44**, 707-717.

Lynch, A. S. & Wang, J. C. (1993). Anchoring of DNA to the bacterial cytoplasmic membrane through cotranscriptional synthesis of polypeptides encoding membrane proteins or proteins for export: a mechanism of plasmid hypernegative supercoiling in mutants deficient in DNA topoisomerase I. *J. Bacteriol.* **175**, 1645-1655.

Majumdar, S., Gupta, S., Sundararajan, V. & Ghosh, T. (1999). Compositional correlation studies among the three different codon positions in 12 bacterial genomes. *Biochem. Biophys. Res. Commun.* **266**, 66-71.

Merrick, M. J. (1993). In a class of its own - the RNA polymerase sigma factor σ^{54} (σ^{N}). *Mol. Microbiol.* **10**, 903-909.

Miller, W. G. & Simmons, R. W. (1993). Chromosomal supercoiling in *Escherichia coli*. *Mol. Microbiol.* **10**, 675-684.

Mizuno, T. (1987). Random cloning of bent DNA segments from *Escherichia coli* chromosome and primary characterization of their structures. *Nucl Acids Res.* **15**, 6827-6841.

Mrazek, J. & Karlin, S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl Acad. Sci. USA,* **95**, 3720-3725.

Nickerson, C. A. & Achberger, E. C. (1995). Role of curved DNA in binding of *Escherichia coli* RNA polymerase to promoters. *J. Bacteriol.* **177**, 5756-5761.

Olson, W., Gorin, A., Lu, X., Hock, L. & Zhurkin, V. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA,* **95**, 11163-11168.

Ornstein, R., Rein, R., Breen, D. & MacElroy, R. (1978). An optimized potential function for the calculation of nucleic acid interaction energies. I. Base stacking. *Biopolymers,* **17**, 2341-2360.

Ozoline, O., Deev, A. & Trifonov, E. (1999). DNA bendability - a novel feature in *E. coli* promoter recognition. *J. Biomol. Struct Dynam.* **16**, 825-831.

Painbeni, E., Carnoff, M. & Rouviere-Yaniv, J. (1997). Alterations of the outer membrane composition in *Escherichia coli* lacking the histone-like protein HU. *Proc. Natl Acad. Sci. USA,* **94**, 6712-6717.

Parkhill, J., Wren, B., Mungall, K., Ketley, J., Churcher, C., Basham, D., Chillingworth, T., Davies, R., Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, A., Moule, S., Pallen, M., Penn, C. W., *et al*. (2000). The genome sequence of the foodborne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**, 665-668.

Parvin, J. D., McCormick, R. J., Sharp, P. A. & Fisher, D. E. (1995). Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. *Nature*, **373**, 724-727.

Pavitt, G. & Higgins, C. (1993). Chromosomal domains of supercoiling in *Salmonella typhimurium*. *Mol. Microbiol.* **10**, 685-696.

Pedersen, A. G., Baldi, P., Brunak, S. & Chauvin, Y. (1998). DNA structure in human RNA polymerase II promoters. *J. Mol. Biol.* **281**, 663-673.

Perez-Martin, J. & de Lorenzo, V. (1995). Integration host factor supresses promiscuous activation of the σ^{54}-dependent promoter Pu of *Pseudomonas putida*. *Proc. Natl Acad. Sci. USA*, **92**, 7277-7281.

Perez-Martin, J. & de Lorenzo, V. (1997). Clues and consequences of DNA bending in transcription. *Annu. Rev. Microbiol.* **51**, 593-628.

Perez-Martin, J., Rojo, F. & de Lorenzo, V. (1994). Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. *Microbiol. Rev.* **58**, 268-290.

Pettijohn, D. E. (1996). The nucleoid. In Escherichia coli *and* Salmonella typhimurium*: Cellular and Molecular Biology*, 2nd edit., pp. 158-166, American Society for Microbiology, Washington, DC.

Plaskon, R. R. & Wartell, R. M. (1987). Sequence distributions associated with DNA curvature are found upstream of strong *E. coli* promoters. *Nucl. Acids Res.* **15**, 785-796.

Polyakov, A., Severinova, E. & Darst, S. A. (1995). Three-dimensional structure of *E. coli* RNA-polymerase: promoter binding and elongation conformation of the enzyme. *Cell*, **83**, 365-373.

Pontiggia, A., Negri, A., Beltrame, M. & Bianchi, M. E. (1993). Protein HU binds specifically to kinked DNA. *Mol. Microbiol.* **7**, 343-350.

Richet, E. & Søgaard-Andersen, L. (1994). CRP induces the repositioning of MalT at the *Escherichia coli* malKp promoter primarily through DNA bending. *EMBO J.* **13**, 4558-4567.

Rivetti, C., Guthold, M. & Bustamante, C. (1999). Wrapping of DNA around the *E. coli* RNA polymerase open promoter complex. *EMBO J.* **18**, 4464-4475.

Sadosky, A., Gray, J. & Hill, C. (1991). The RhsD-E subfamily of *Escherichia coli* K-12. *Nucl. Acids Res.* **19**, 7177-7183.

Sarai, A., Mazur, J., Nussinov, R. & Jernigan, R. L. (1989). Sequence dependence of DNA conformational flexibility. *Biochemistry*, **28**, 7842-7849.

Satchwell, S. C., Drew, H. R. & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659-675.

Schellhorn, H., Audia, J., Wei, L. & Chang, L. (1998). Identification of conserved, rpos-dependent stationary-phase genes of *Escherichia coli*. *J. Bacteriol.* **180**, 6823-6831.

Schickor, P., Metzger, W., Werel, W., Lederer, H. & Heumann, H. (1990). Topography of intermediates in transcription initiation of *E. coli*. *EMBO J.* **9**, 2215-2220.

Serrano, M., Barthelemy, I. & Salas, M. (1991). Transcription activation at a distance by phage ϕ29 protein p4: effect of bent and non-bent intervening DNA sequences. *J. Mol. Biol.* **219**, 403-414.

Shimizu, M., Miyake, M., Kanke, E., Matsumoto, U. & Shindo, H. (1995). Characterization of the binding of HU and IHF, homologous histone-like proteins of *Escherichia coli*, to curved and uncurved DNA. *Biochim. Biophys. Acta*, **1264**, 330-336.

Shpigelman, E. S., Trifonov, E. N. & Boishoy, A. (1993). CURVATURE: software for the analysis of curved DNA. *CABIOS*, **9**, 435-440.

Simpson, R. T. (1991). Nucleosome positioning: occurrence, mechanisms, and functional consequences. *Prog. Nucl. Acids Res. Mol. Biol.* **40**, 143-184.

Sinden, R. R. & Pettijohn, D. E. (1981). Chromosomes of living *Escherichia coli* cells are segregated into domains of supercoiling. *Proc. Natl Acad. Sci. USA*, **78**, 224-228.

Sinden, R. R. & Ussery, D. W. (1992). Analysis of DNA structure *in vivo* using psoralen photobinding: measurement of supercoiling, topological domains, and DNA-protein interactions. *Methods Enzymol.* **212**, 319-335.

Sinden, R. R., Pearson, C. E., Potaman, V. N. & Ussery, D. W. (1998). DNA: structure and function. *Advan. Gen. Biol.* **5A**, 1-141.

Sines, L. M.-I. C. C. & Williams, L. D. (1999). DNA structure: cations in charge? *Curr. Opin. Struct. Biol.* **9**, 298-304.

Smith, D., Doucette-Stamm, L., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., *et al*. (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135-7155.

Staczek, P. & Higgins, N. P. (1998). Gyrase and topo IV modulate chromosome domain size *in vivo*. *Mol. Microbiol.* **29**, 1435-1448.

Starr, D. B., Hoopes, B. C. & Hawley, D. K. (1995). DNA bending is an important component of site-specific recognition by the TATA binding protein. *J. Mol. Biol.* **250**, 434-446.

Stephens, R., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R., Zhao, Q., Koonin, E. & Davis, R. (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science*, **282**, 754-759.

Suck, D. (1994). DNA recognition by DNase I. *J. Mol. Recogn.* **7**, 65-70.

Sueoka, N. (1998). Cell membrane and chromosome replication in *Bacillus subtilis*. *Prog. Nucl. Acid Res. Mol. Biol.* **59**, 35-53.

Tanaka, K., Muramatsu, S., Yamada, H. & Mizuno, T. (1991). Systematic characterization of curved DNA segments randomly cloned from *Escherichia coli* and their functional significance. *Mol. Gen. Genet.* **226**, 367-376.

Tanaka, H., Goshima, N., Kohno, K., Kano, Y. & Imamoto, F. (1993). Properties of DNA-binding of HU heterotypic and homotypic dimers from *Escherichia coli*. *J. Biochem.* **113**, 568-572.

ten Heggeler-Bordier, B., Wahli, W., Adrian, M., Stasiak, A. & Dubochet, J. (1992). The apical localization of transcribing RNA polymerases on supercoiled DNA prevents their rotation around the template. *EMBO J.* **11**, 667-672.

Tomb, J., White, O., Kerlavage, A., Clayton, R., Sutton, G., Fleischmann, R., Ketchum, K., Gill, H. K. S., Dougherty, B., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E., Peterson, S.,; Loftus, B., *et al*. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, **388**, 539-547.

Valentin-Hansen, P., Gaard Andersen, L. S. & Pedersen, H. (1996). A flexible partnership: the CytR anti-activator and the cAMP-CRP activator protein, comrades in transcription control. *Mol. Microbiol.* **20**, 461-466.

Wang, Y., Zhao, S. & Hill, C. (1998). Rhs elements comprise three subfamilies which diverged prior to acquisition by *Escherichia coli*. *J. Bacteriol.* **180**, 4102-4110.

Wolffe, A. P. & Drew, H. R. (1995). DNA structure: implications for chromatin structure and function. In *Chromatin Structure and Gene Expression* (Elgin, S. C. R., ed.), pp. 27-48, IRL Press, Oxford.

Worcel, A. & Burgi, E. (1972). On the structure of the folded chromosome of *Escherichia coli*. *J. Mol. Biol.* **71**, 127-147.

Worning, P., Jensen, L., Nelson, K., Brunak, S. & Ussery, D. (2000). Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*. *Nucl. Acids Res.* **28**, 706-709.

Wye, D., Bronson, E. C. & Anderson, J. N. (1991). Species-specific patterns of DNA bending and sequence. *Nucl Acids Res.* **19**, 5253-5261.

Young, I. T. (1977). Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources. *J. Histochem. Cytochem.* **25**, 935-941.

Zhu, Z. & Thiele, D. J. (1996). A specialized nucleosome modulates transcription factor access to a *C. glabrata* metal responsive promoter. *Cell*, **87**, 459-470.

Edited by T. Richmond

(Received 11 August 1999; received in revised form 10 April 2000; accepted 11 April 2000)