Nucleic Acids Research, 2005, Vol. 33, No. 12 ***3907–3918***
doi:10.1093/nar/gki699

A-tract clusters may facilitate DNA packaging in bacterial nucleoid

Michael Y. Tolstorukov[1,3], Konstantin M. Virnik[2], Sankar Adhya[2] and Victor B. Zhurkin[1,*]

[1]Laboratory of Experimental and Computational Biology and [2]Laboratory of Molecular Biology, National Cancer Institute, Bethesda, MD 20892, USA and [3]Department of Biological and Medical Physics, V. Karazin Kharkov National University, Kharkov, 61077, Ukraine

Received January 25, 2005; Revised and Accepted June 22, 2005

ABSTRACT

Molecular mechanisms of bacterial chromosome packaging are still unclear, as bacteria lack nucleosomes or other apparent basic elements of DNA compaction. Among the factors facilitating DNA condensation may be a propensity of the DNA molecule for folding due to its intrinsic curvature. As suggested previously, the sequence correlations in genome reflect such a propensity [Trifonov and Sussman (1980) *Proc. Natl Acad. Sci. USA*, 77, 3816–3820]. To further elaborate this concept, we analyzed positioning of A-tracts (the sequence motifs introducing the most pronounced DNA curvature) in the *Escherichia coli* genome. First, we observed that the A-tracts are over-represented and distributed 'quasi-regularly' throughout the genome, including both the coding and intergenic sequences. Second, there is a 10–12 bp periodicity in the A-tract positioning indicating that the A-tracts are phased with respect to the DNA helical repeat. Third, the phased A-tracts are organized in ~100 bp long clusters. The latter feature was revealed with the help of a novel approach based on the Fourier series expansion of the A-tract distance autocorrelation function. Since the A-tracts introduce local bends of the DNA duplex and these bends accumulate when properly phased, the observed clusters would facilitate DNA looping. Also, such clusters may serve as binding sites for the nucleoid-associated proteins that have affinities for curved DNA (such as HU, H-NS, Hfq and CbpA). Therefore, we suggest that the ~100 bp long clusters of the phased A-tracts constitute the 'structural code' for DNA compaction by providing the long-range intrinsic curvature and increasing stability of the DNA complexes with architectural proteins.

INTRODUCTION

Multi-level DNA packaging in a bacterial nucleoid involves concerted interactions between genomic DNA, architectural proteins and RNA (1–4). It has been suggested that at higher levels of organization, topologically independent segments of genomic DNA (domains of supercoiling) are packaged into the rosette-like structure (1,5,6). However, this packaging mode cannot account for the total DNA compaction ratio (10^3–10^4) in bacteria (7,8). Hence, some lower levels of nucleoid organization should play a significant role in DNA condensation. The spatial organization of bacterial DNA at these levels remains unknown as bacteria lack apparent packaging subunits, such as nucleosomes in eukaryotes.

Among the factors facilitating DNA folding may be a propensity of DNA to form small coils and loops due to its sequence-dependent intrinsic curvature, in other words—a 'structural code' encrypted in the DNA sequence (9,10). Indeed, if the intrinsically curved DNA fragments were positioned periodically in phase with the DNA helical repeat of 10–11 bp, the local DNA bends would accumulate, causing formation of nearly planar arcs or open loops (11). Presence of the DNA loops at specific positions in genome would facilitate DNA packaging in a 'programmed' manner. Note that the DNA loops, typically containing ~10^2 bp, play important role in gene regulation in bacteria [see (12,13) for review].

The mono- and dinucleotide 10–11 bp periodicities were observed in genomic DNA in a number of studies, starting with a pioneering work by Trifonov and Sussman (14–17). Initially, these periodicities were associated with the DNA folding in eukaryotic nucleosomes (14). Later, it was shown that the coding sequences from both pro- and eukaryotic genomes are similar in this regard and an alternative interpretation

*To whom correspondence should be addressed. Tel: +1 301 496 8913; Fax: +1 301 402 4724; Email: zhurkin@nih.gov
Correspondence may also be addressed to Michael Y. Tolstorukov. Tel: +38 057 707 5576; Fax: +1 831 308 7657; Email: tolstorukov@gmail.com

was suggested (15), connecting such periodicities with the 3.6 amino acid periodicity observed in the protein α-helices (as 3.6 amino acid periodicity 'translates' into 10.8 bp periodicity in the coding DNA sequences).

However, it remains unclear how frequent are these 'periodically organized' DNA fragments in various genomes. Can these periodicities play a major role in DNA packaging? Do such fragments have a certain characteristic length (which would define the size of the DNA loops mentioned above), or are their lengths distributed in a statistically independent manner? So far, the latter question has been addressed only qualitatively, based on the 'visual' inspection of the genomic sequence periodicities (18). One of the reasons for this limitation is the absence of a rigorous quantitative procedure to estimate the 'length' of a sequence periodicity. We suggest such a procedure here and apply it to the analysis of the sequence periodicities in *Escherichia coli* and other bacterial genomes.

Another ambiguity existing in the literature is related to the distribution of the intrinsically curved DNA fragments between various functional regions in genome. Using various DNA bending models, it was predicted that in the intergenic regions, especially in promoters, the DNA curvature is stronger on average than in the coding sequences (19–22). [The role of curved DNA elements in gene regulation has been extensively reviewed in (23–25)]. However, the 10–11 bp periodicities, which may contribute to the long-range 'systematic' DNA bending and folding, are present (and were first discovered) in the coding regions (14,15). As the coding sequences constitute a bulk of the bacterial genome, the curved fragments have to be omnipresent in the coding regions in order to play a role in the genome packaging. The question if this is the case remains open.

The best-known motifs causing intrinsic DNA bending (or curvature) are the A-tracts, defined as the sequences A_nT_m (26–31). The largest DNA bends are produced by periodically repeating 4–8 bp long A-tracts (30,31) separated by 'quasi-random' (predominantly GC-rich) DNA fragments, such that the overall sequence period is close to the DNA helical repeat, 10.5 bp (or its multiple).

There are several alternative models describing the A-tract-induced curvature of DNA, reviewed recently in (25,32). Two of them are used most frequently: the A-tract (30,31,33) and the non-A-tract (34–36) models. These models differ in stereochemical details (such as the sequence-dependent inclination of base pairs in solution), which are still a subject of heated discussion. However, one important fact is established unambiguously: when an A-tract is surrounded by 'quasi-random' DNA fragments, the directionality of bending is consistent with the 'bending vector' being directed into the minor groove approximately in the center of A-tract (30,31) (see Figure 1A). Our analysis of the A-tracts phasing in genomic sequences is based on this observation.

Another sequence motif, G-tracts (G_nC_m), is associated with DNA bending into the major groove, especially in the presence of divalent cations (37,38). To focus on the strongest possible 'structural signals' in the DNA sequence, we restricted our analysis to the distribution of the A- and G-tracts.

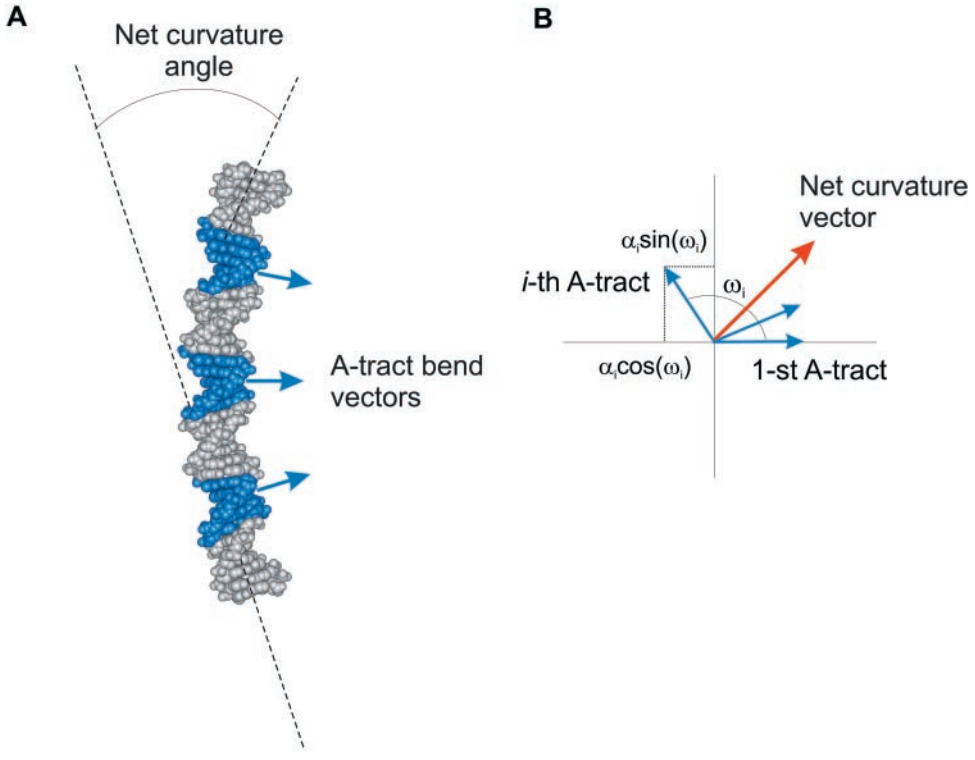

Figure 1. Schematic representation of the procedure used to calculate the 'A-tract curvature' for a DNA segment. (**A**) The A-tracts (shown in blue) interspersed with 'random' sequences (shown in gray) cause local DNA bending. Each bend is represented by a vector directed into the minor groove in the center of an A-tract, with the length proportional to the bending angle. (**B**) View along the DNA axis: the resulting 'curvature vector' (in red) is the vector sum of the A-tract bending vectors. The A-tract curvature was determined as the length of this vector.

We show that the distribution of the two types of tracts is strikingly different—the A-tracts are abundant and omnipresent, while the G-tracts are under-represented in the bacterial genomes. In addition, the distribution of the A-tracts is highly non-random and 'quasi-periodic', and thus is likely to contribute to the DNA packaging.

METHODS

The A-tracts are defined as the sequences A_nT_m, where $4 \leqslant (n + m)$, i.e. A-tracts comprise the dimeric steps AA:TT and AT, but not TA. Similarly, G-tracts are the sequences G_nC_m, where $4 \leqslant (n + m)$. Only those A- and G-tracts were taken into account that cannot be extended further. In particular, the sequence A_nT_m was counted as an A-tract if it occurred in the context BA_nT_mV (both m, $n \neq 0$), or BA_nB ($n \neq 0$, $m = 0$) or VT_mV ($m \neq 0$, $n = 0$), where B $\neq$ A and V $\neq$ T.

To describe the distribution of A- and G-tracts, the distance autocorrelation function (DAF) was used, similar to that introduced earlier (14,15). This function, $F(x)$, represents the number of the pairs of A-tracts with distance x between their centers (Figure 2). The $F(x)$ values were calculated as follows. If the distance x is integral (e.g. $x_3 = 20$ in Figure 2), then $F(x)$ is increased by 1. Otherwise, the function values for the two nearest integers, $x + 0.5$ and $x - 0.5$, are increased by 0.5. In the example given in Figure 2, two distances are non-integral, $x_1 = 9.5$ and $x_2 = 10.5$. Accordingly, the values $F(9)$ and $F(11)$ are increased by 0.5, while $F(10)$ is increased by $0.5 + 0.5 = 1$. By analogy, the same procedure was applied to the G-tracts.

The resulting function $F(x)$, defined for the integral arguments x, was expanded into the Fourier series (39) of N frequency harmonics with coefficients A_j, $0 \leqslant j \leqslant N - 1$. Here, the magnitude of a zero frequency component, A_0, is the average value for the function $F(x)$, and the rest of the coefficients A_j ($j = 1, 2, 3 \ldots$) are the magnitudes of the higher-frequency components. Next, we calculated the 'intensity' of each non-zero frequency component, $\bar{A}_j = |A_j| / \sqrt{\sum_{k=1}^{N-1} |A_k|^2}$, which characterizes a contribution of this component (with a period N/j) to the net oscillation of the autocorrelation function.

To introduce a measure of the A-tracts phasing within a given DNA segment, the following procedure was applied. The bends caused by A-tracts were represented by vectors going through the centers of the A-tracts and pointing toward the minor groove (Figure 1A); see Introduction and refs (30,31). Then, the '2D vector sum' of the bending vectors was calculated and the net A-tract curvature, κ, was obtained as the length of the resulting vector (Figure 1B):

$$\kappa = \left[\left(\sum_i \alpha_i \cos(\omega_i) \right)^2 + \left(\sum_i \alpha_i \sin(\omega_i) \right)^2 \right]^{1/2}, \qquad \mathbf{1}$$

where α_i is the bending angle for the i-th A-tract, ω_i is the net DNA twisting between the i-th and the first bending vectors (Figure 1B). This simplified procedure is similar to those used previously for the analyses of the physical properties of both DNA (40) and protein (41) sequences. Note that the '2D vector sum' defined above does not represent actual equilibrium curvature in a DNA segment, but rather is a measure of the A-tracts phasing, and as such it serves the purposes of the present study. Furthermore, the 'direct' summation of the A-tract bends in the 3D space—rather than using '2D vector sum'—would require introducing at least 10 more parameters into the model (wedge angles at each dinucleotide step). Such an approach would make the A-tract curvature estimation less reliable (25).

Among various A-tracts, the bend angle is known to be the largest for the A_6-tract, estimated as 17–21° from cyclization experiments (42). The bend angle magnitude depends on the ionic conditions, specifically on the concentration of divalent cations Mg^{2+} (43). Topological measurements of supercoiled DNA under the ionic conditions comparable with physiological conditions (in the presence of Mg^{2+}) found the A_6-tract bend angle to be 22° at room temperature (44). Thus, we consider a value 20 ± 2° to be the best current estimate of the DNA bending per A_6-tract. Based on the quadratic relationship between the gel retardation and the DNA bending angle per helical pitch (30), we used the following bend angles: 20° for the 5–6 bp A-tracts, and 15° for the 4 and 7 bp long A-tracts. An additional assumption here is that the DNA bending angle is the same for all A_nT_m-tracts with a given length ($m + n$), i.e. A_6 and A_3T_3 are assumed to produce the same DNA bending (25,28). Approximate magnitudes of the A-tract bends are suitable for estimating the degree of A-tract phasing.

The genome of *E.coli* K-12 MG1655 (and other bacterial genomes used in the present study) was retrieved from the GenBank of the National Center for Biotechnology Information (ftp://ftp.ncbi.nih.gov/genbank/genomes). The coding and intergenic regions were determined according to the GenBank annotations [coding DNA sequence (CDS) coordinates]. As the bacterial genomes have very few introns, they were omitted from our analysis. For comparison, the 'random' sequences were generated, with the same nucleotide composition as in the corresponding genome. The results for genomic sequences were compared with the corresponding means over 10 implementations of the independently generated random sequences; the corresponding variances were used to evaluate statistical significance of the differences.

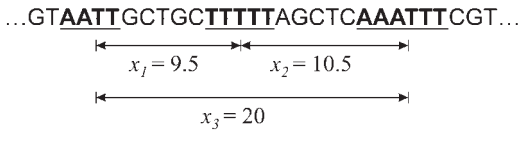

Figure 2. A scheme illustrating definition of the DAF for A-tracts (underlined). The distances between the tract centers, x_i, are in base pairs. If the length of an A-tract is odd (e.g. A_5), its center coincides with the central base pair. If the A-tract length is even (e.g. AATT), its center is placed between the base pairs of its central dimeric step.

RESULTS

Occurrences of the A- and G-tracts

First, we calculated the numbers of occurrences of the A- and G-tracts of different lengths in the *E.coli* genome (Table 1) and compared them with the corresponding values for random

Table 1. Occurrences of the A- and G-tracts of different length in the *E.coli* genome and in the coding sequences, CDS[a]

Tract length (bp)	A-tracts (A_nT_m)		G-tracts (G_nC_m)	
	Genome	CDS	Genome	CDS
4	51603	41360	37904	33814
5	23384	18832	8446	7403
6	8603	6306	1624	1380
7	2794	1858	315	241
8	999	661	41	23
9	227	138	3	2
10	28	17	2	2

[a]The length of A- and G-tracts is defined as $(n + m)$.

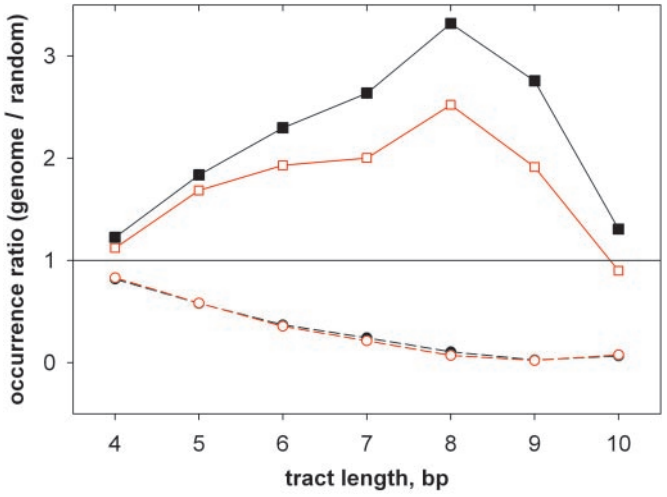

Figure 3. Relative occurrences of the A- and G-tracts in the *E.coli* genome. Given are the ratios of occurrences in genome to the average occurrences in 10 random sequences of the same base composition. Over-representation of the genomic A-tracts (A_nT_m) is statistically significant ($P < 0.001$, *t*-test) for the lengths $(n + m) = 5$–9 bp. Black lines with filled symbols represent the data for the entire genome and the red lines with open symbols represent the data for the coding sequences only (CDS). The A-tracts (solid lines with squares) are over-represented in the genome, while the G-tracts (dashed lines with circles) are under-represented (see Table 1 for the absolute numbers).

sequences (Figure 3). Notice that the G-tracts are under-represented (Figure 3), the ratio 'genome/random' being as low as 0.1 for the tract length of 10 bp. In contrast, the short A-tracts are over-represented, with the 'genome/random' ratio reaching its maximum for 8 bp long tracts (both for the entire genome and for the CDSs). In terms of the absolute numbers, the situation is different (Table 1)—the main contribution to the total number of A-tracts comes from the shorter A-tracts. Therefore, when analyzing the distances between the tracts, we restricted ourselves to the 4–7 bp long A- and G-tracts.

Distance autocorrelations of the A- and G-tracts

In Figure 4, the distance autocorrelations of the A- and G-tracts in the *E.coli* genome are compared with those in random sequences. There are several features distinguishing the distributions of these tracts in the genome. First, the 'distance occurrences' for the genomic A-tracts (Figure 4A, solid lines) are 2–3 times higher than for random sequences (dotted lines), while the situation is reverse for G-tracts (Figure 4B).

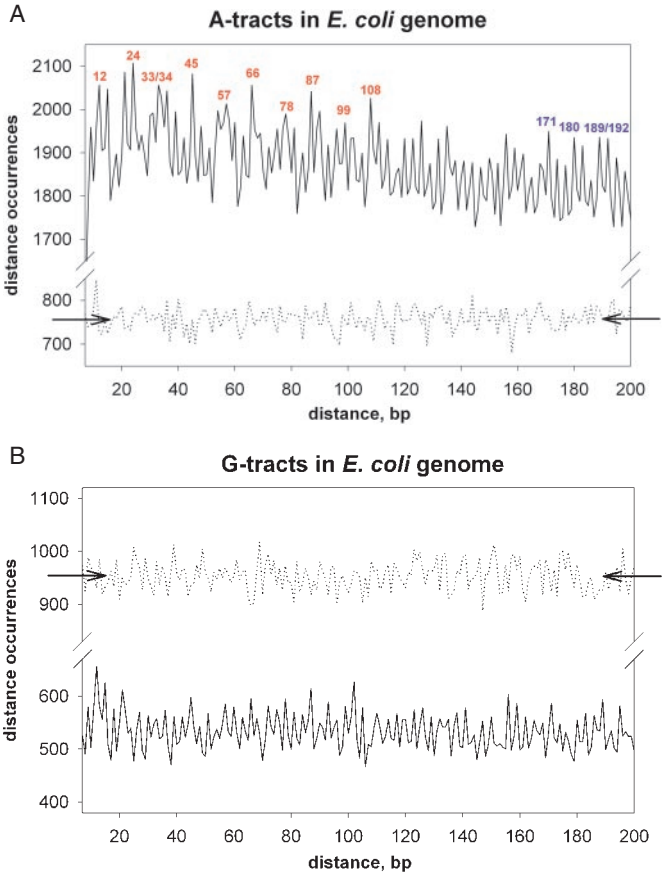

Figure 4. Distance autocorrelations of the (**A**) A-tracts and (**B**) G-tracts in the *E.coli* genome. The tract lengths vary from 4 to 7 bp; therefore, the distances between the tracts are considered to be 7 bp and larger. The data for the genomic DNA are represented by solid lines, and those for random sequences by dotted lines. Both genomic and random sequences have [G + C] = 51%. Note that in random sequences, the average occurrence of the A-tracts is somewhat less than that of the G-tracts [see the arrows in (A) and (B), respectively]; this is consistent with the GC-content exceeding 50%. The genomic DNA reveals an opposite trend: the absolute numbers for the A-tracts (A) are three times higher than those for the G-tracts (B). Also, periodicity of 10–12 bp is seen in the positioning of the peaks for the A-tracts [numbers in (A)], while no apparent regularity is seen for the G-tracts. The peaks in (A) marked with blue numbers are out of phase with the peaks marked with red numbers (see the main text).

These trends are consistent with the over-representation of A-tracts and under-representation of G-tracts in the genome (Figure 3). Second, the magnitudes of the oscillations are larger for the A-tracts than for the G-tracts. Third, there is an apparent 10–12 bp periodicity in the peak positioning for the A-tracts (cf. red numbers in Figure 4A), while the G-tract peaks do not reveal any kind of regularity (Figure 4B). Finally, the absolute values of occurrences decrease with the distance quite noticeably for the A-tracts, but not for the G-tracts. This indicates that the A-tracts are not evenly positioned throughout the genome but are grouped together, so that these A-tracts are more frequently separated by short than by long distances (will be discussed in detail below). If the A-tracts were distributed randomly, there would be no decrease in their occurrences with the distance, since the probabilities of finding two A-tracts separated by 20 or 200 bp are the same. Indeed, no such decrease was observed for random sequences (Figure 4A, dotted line).

Overall, the distance autocorrelations of the A-tracts in the genome differ principally from those in random sequences. At the same time, the G-tract autocorrelations calculated for the genome and for random sequences are quite similar, indicating that the genomic G-tracts are distributed quasi-randomly. Therefore, further analysis is concentrated solely on the A-tract distribution.

Fourier analysis of the distance autocorrelations

To quantify periodic patterns in the A-tract autocorrelation function, DAF, the Fourier transform technique was used. We calculated the intensities of Fourier harmonics (periodicities) as a function of the oscillation period (see Methods for details). In short, the intensity of a certain periodicity represents the relative contribution of this periodicity into the net oscillating component of the initial function. As mentioned above, the DAF values vary significantly with distance (cf. magnitudes of the peaks in Figure 4A for the distances shorter than 100 bp and for those longer than 100 bp). Therefore, we analyzed the A-tract autocorrelations in two data intervals (sampling windows) separately: (i) the 'first 100 bp' window, with the distances between the A-tracts from 7 to 106 bp and (ii) the 'second 100 bp' window, with the distances from 107 to 206 bp. (The minimal distance analyzed, 7 bp, corresponds to the shortest distance between the centers of two non-overlapping 7 bp A-tracts.)

Consider the results of the Fourier series expansion of the A-tract DAF in the 'first 100 bp' window (Figure 5A, solid line). These results reveal a strong periodic pattern: there are two main peaks in the intensity plot at 3 and 11.1 bp, representing two major non-zero harmonics. Notice that the sum of these two harmonics reproduces the oscillatory behavior of the DAF remarkably well, accounting for every local maximum in its irregular profile (Figure 5B). Similar peaks were observed in the previous analyses of the mono- and dinucleotide autocorrelations in bacterial genomes [reviewed in (9)]. The 3 bp periodicity is related to the protein coding (14), and the 11.1 bp periodicity apparently reflects the phasing of the A-tracts along the DNA helix, because it suggests the frequent occurrence of the distances between the A-tracts, which are close to the DNA helical repeat (~10.5 bp) or its multiples.

Since the 'spectral resolution' is relatively low when only 100 data points are used in the Fourier transform, we have expanded the DAF over a larger interval of distances (up to 1200 bp) to achieve a higher 'resolution'. The results did not change principally; however, the 11.1 bp peak has decomposed into several peaks between 10 and 12 bp, with two main peaks at 10.6 and 11.2 bp (resolution ±0.1 bp; data not shown). Therefore, here and below, we refer to the 11.1 bp peak in the intensity plot as the 10–12 bp peak.

Next, we consider the results of the Fourier analysis of the A-tract DAF in the 'second 100 bp' window (Figure 5A, dotted line). The 10–12 bp peak disappears, while the 3 bp peak retains nearly the same amplitude. The change in the amplitude of the 10–12 bp periodicity is clearly visible in Figure 5C, where the sums of the zero frequency and 11.1 bp harmonics are plotted against the 'first' and the 'second 100 bp' data intervals. Based on these results, we conclude that the 10–12 bp periodicity, or the A-tract helical phasing, is strongly

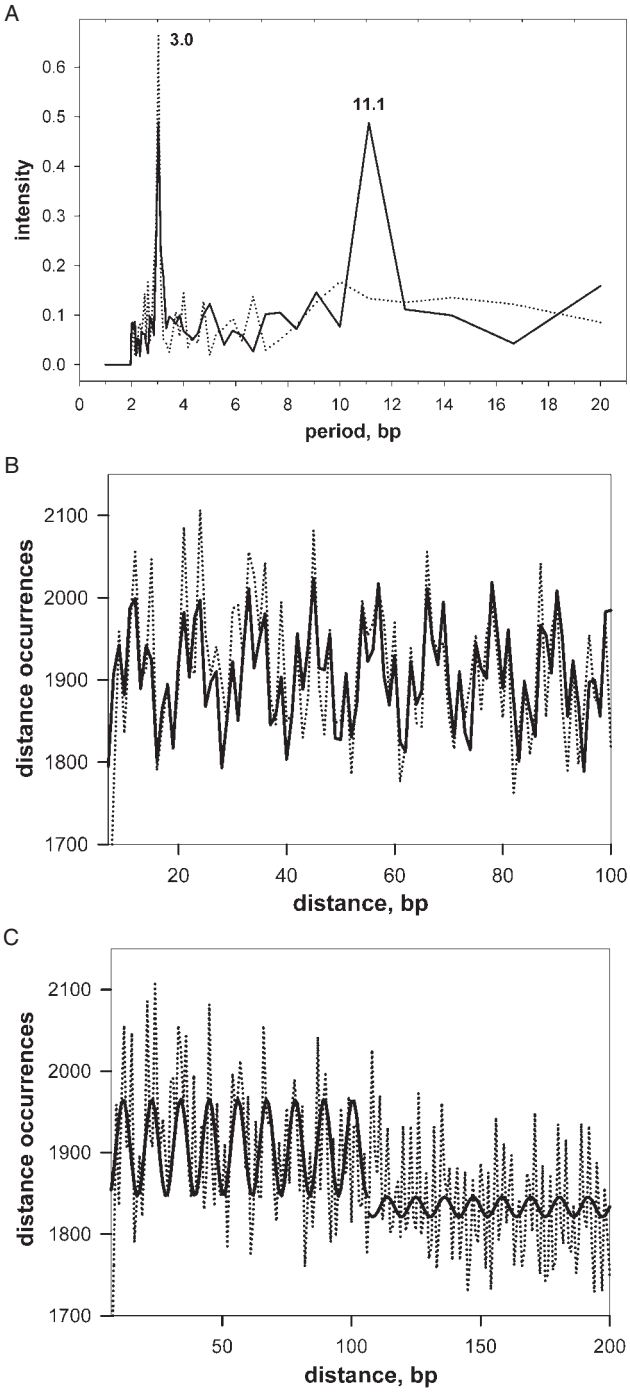

Figure 5. Relative contributions of the periodicities into the net oscillating component of the A-tract autocorrelation function, DAF. (**A**) Intensity of the periodicity as a function of its period. Results were obtained by expanding the DAF (Figure 4A, solid line) into the Fourier series (see Methods for details). Calculations were performed for two data sets (sampling windows): solid line—the 'first 100 bp' (the DAF values for the distances 7–106 bp) and dotted line—the 'second 100 bp' (the DAF values for the distances 107–206 bp). Note that the peak at 11.1 bp is present only for the first window, while the peak at 3.0 bp is present for both windows. (**B**) Superposition of the zero frequency, 3.0 and 11.1 bp harmonics (solid line) is plotted against the DAF (dotted line). Note that the two non-zero harmonics are sufficient to reproduces the 'spiky' behavior of the autocorrelation function. (**C**) Superposition of the zero frequency and the 11.1 bp harmonics with the intensities calculated for the 'first' and the 'second 100 bp' windows (solid line) are plotted against the corresponding intervals of the DAF (dotted line).

pronounced for the distances up to ~100 bp, but vanishes when the distance increases up to 200 bp.

The A-tract periodicity length

To determine the characteristic length of the A-tract phasing, we used the 'sliding' window technique. Namely, we analyzed how the intensity of the 10–12 bp periodicity changes when the sampling window (in which DAF was expanded into Fourier series) gradually moved from the 'first 100 bp' (strong 10–12 bp peak in the intensity plot, Figure 5A) to the 'second 100 bp' (weak 10–12 bp peak). This procedure is illustrated in Figure 6A, where it is applied to a simple test function, $f(x)$, representing an ideal sine wave with period of 10 arbitrary units (au) in the interval $1 \leqslant x \leqslant 100$ and a constant for $x > 100$. In this case, the 10 au harmonic intensity changes from one to zero, diminishing 2-fold at such a position of the sampling widow, when the left half of the window contains a perfect sine wave, and its right half contains a constant function. Accordingly, the position of the window center corresponds to the periodicity length, which is 100 au for the test function.

To examine how the size of the sampling window affects the results, we used three window sizes: 80, 100 and 120 au. The half-drops in the intensity occur at 100 au—as expected—for the 80 au (dotted line) and 100 au (solid line) sampling windows. In the case of the 120 au window (dashed line), the intensity plot does not reach the plateau at the beginning, because the window size is larger than the periodicity length. As a result, the intensity half-drop point is slightly shifted to the higher lengths (Figure 6A). Thus, in order to estimate the periodicity length correctly, one has to consider several window sizes and select those for which the plateau at the beginning of the plot is observed.

Next, we applied the described procedure to the genomic A-tract DAF. The calculated dependence of the intensity of 10–12 bp periodicity on the position of the sampling window is shown in Figure 6B. Again, the windows of three sizes have been used: 80, 100 and 120 bp. Overall, the plots have the shapes similar to those calculated for the test function (Figure 6A). In the case of the 80 and 100 bp windows, the 'intensity versus window position' plots have plateaus at the beginning, validating usage of these windows for determining the periodicity length. Our procedure for the 80 and 100 bp windows gives an estimate of this length, which is ~100 bp. In the case of 120 bp window, wherein plateau at the beginning of the plot is absent, the half-drop of intensity is also close to 100 bp. Taking into account complexity of the genomic sequence and the consequent irregular oscillations of the A-tract DAF, we consider this value as a plausible assessment of the characteristic length of the A-tract phasing in the *E.coli* genome.

The presence of a broad intensity maximum in the range of 170–210 bp (Figure 6B) can be interpreted as follows. There are two groups of peaks in the A-tract DAF (marked with red and blue numbers in Figure 4A). The peak-to-peak distances within each group are multiples of 10–12 bp (the peaks are phased), but there is no correlation between positioning of the 'red' and 'blue' peaks. Also, the 'red' and 'blue' peaks are not correlated with the peaks in the range 110–170 bp. Accordingly, when the sampling window moves out of the region of

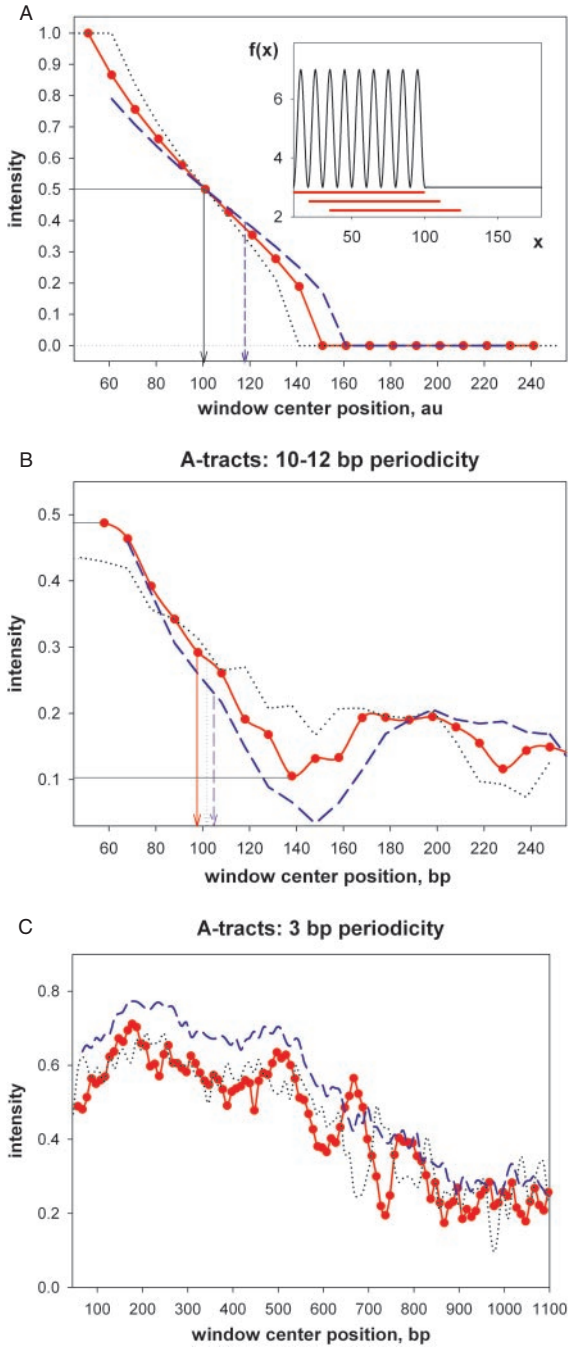

Figure 6. Intensity of an oscillating component as a function of the sampling window position. The window positioning step was 10 bp. Three sizes of the sampling window were used: 80 bp (dotted black lines), 100 bp (solid red lines with circles) and 120 bp (dashed blue lines). (**A**) Intensity of the 10 au periodicity for the test function: $f(x) = 5 - 2\cos(2\pi x/10)$ for $1 \leqslant x \leqslant 100$ and $f(x) = 3$ for $x > 100$ ('au' stands for arbitrary units; the test function is shown in the inset). Using all three window sizes results in a similar drop in the intensity when the window slides out of the periodicity range (from 1 to 100 au for the test function). When the window size is smaller than the periodicity length, a plateau precedes the drop, otherwise there is no such a plateau. The length of the periodicity can be evaluated as a half-height of the intensity drop (shown with an arrow). (**B**) Intensity of the 10–12 bp periodicity for the A-tract DAF. The periodicity lengths (shown with the arrows) amounted to ~100 bp for all three sizes of the sampling window. Possible origin of the second maximum in the range of 170–210 bp is discussed in the text. (**C**) Intensity of the 3 bp periodicity for the A-tract DAF. Note that there is no drop in the intensity within the first 500 bp.

the 'red' peaks, there is a drop in the intensity of 10–12 bp periodicity, and when the 'blue' peaks get into the sampling window the second intensity maximum appears. This finding implies that the groups of independently phased A-tracts (or the A-tract clusters) are often located close to each other, with the distance between the cluster centers being ~200 bp.

As a control, we applied the same procedure to the intensity peak at 3 bp, to find out the length of periodicity in this case (Figure 6C). Since the 3 bp periodicity is believed to be associated with the protein coding (14) one would anticipate that its characteristic length is close to an average gene length (~1000 bp in the case of *E.coli*). As expected, the 3 bp periodicity expands much further than does the 10–12 bp periodicity, and a noticeable drop in the intensity is observed in the range of 800–1000 bp.

Summarizing, our results indicate that the A-tracts are not distributed randomly in the genome, but rather are grouped into clusters containing several A-tracts phased with respect to each other. The characteristic length of such clusters is ~100 bp. Next, we tackle the distribution of these clusters in the *E.coli* genome, focusing on the comparative analysis of the coding and regulatory regions.

The A-tract clusters in the coding and intergenic regions

To analyze the distribution of the phased A-tracts in the genome, we calculated the 'A-tract curvature' in the sliding 100 bp windows (Figure 7). As described in Methods, this parameter characterizes the number of A-tracts and the degree of their phasing in a window. The 100 bp window corresponds to the characteristic length of the clusters of phased A-tracts as estimated above.

Two important inferences follow from the data presented in Figure 7A. First, the phased A-tract clusters are abundant in the genome. In particular, there are ~9000 non-overlapping clusters with the curvature >40°, compared with ~3000 such clusters in random sequences of the same base composition (data for random sequences are not shown). Notice that the A-tract curvature of 40° requires the presence of at least two phased 5–6 bp A-tracts in a window. The average peak-to-peak distance is ~500 bp for peaks >40° in the genome, compared with ~1500 bp in a random sequence (calculated from the data shown in Figure 7A and the data for random sequences).

Second, many DNA fragments that are noticeably curved due to the presence of the phased A-tracts are located inside CDS regions (Figure 7A, red bars). About 70% of the A-tract curvature peaks >40° lie in the CDS; the net number of such curved DNA fragments in CDS exceeds 2.5 times the corresponding value for a random sequence. This is an unexpected observation since significant DNA curvature was reported previously only in promoter regions of the bacterial genomes (21,22), but not in the CDS. Finally, the large peaks in A-tract curvature (>100°) were observed both in the intergenic regions and in the CDS (Figure 7B and C, respectively).

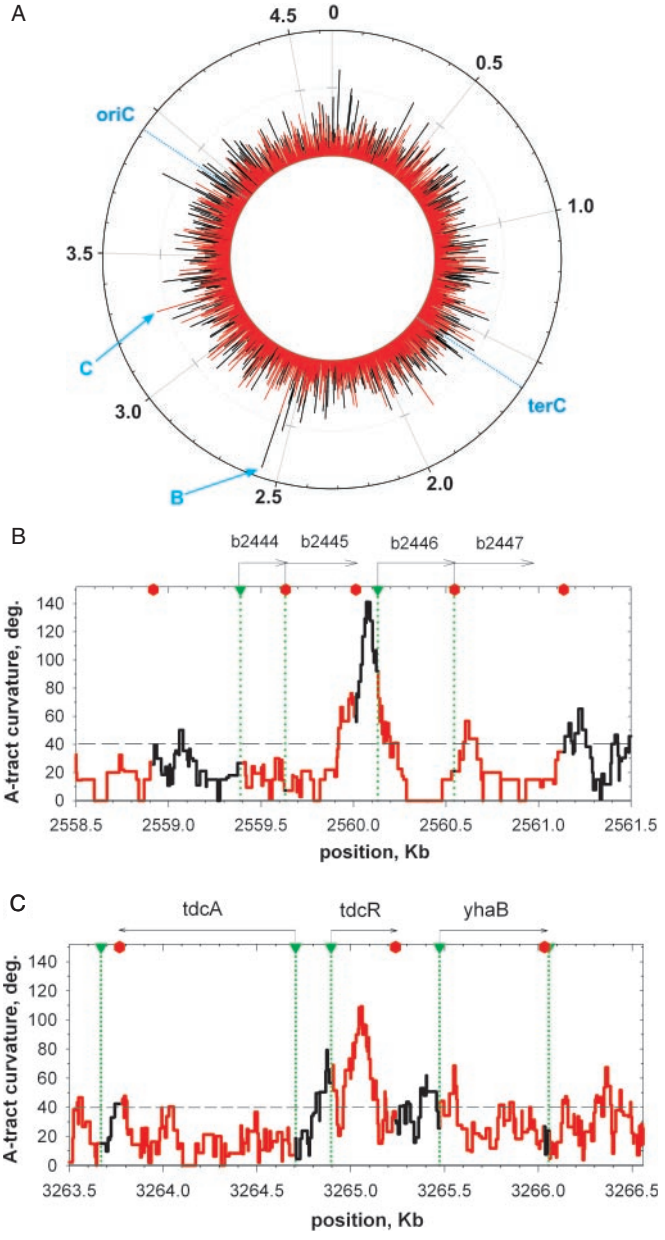

Figure 7. Distribution of the A-tract curvature in the *E.coli* genome. The A-tract curvature was calculated in the 100 bp sliding window with 1 bp step, according to Equation 1 and using the set of the twist angles from (47). (**A**) Circular diagram represents the A-tract curvature distribution in the entire *E.coli* genome (red, CDS; black, intergenic sequences). Origin and terminus of replication (63) are indicated. The numbers on the outmost circle indicate the position in the genome in millions of base pairs. The A-tract curvature values start at 40° (the points on the innermost circle). The cross bars at the radial lines correspond to 100° of curvature. (**B** and **C**) Two detailed pictures of the A-tract curvature distribution in the 3000 bp regions, indicated with arrows in (A). The start and termination sites of transcription are indicated with green triangles and red hexagons, respectively. The gene directions are shown with the hooked arrows. Notice that the strong peaks in A-tract curvature (up to 100°) are located both in the intergenic regions (B) and in the CDS (C).

A-tracts in other bacteria

Distribution of A-tracts in several other bacterial genomes with intermediate GC-content, *Bacillus halodurans* (44% GC), *Yersinia pestis* (48% GC), *Salmonella typhimurium* and *Neisseria meningitides* (52% GC), is similar to that in *E.coli*. First, the A-tracts are 2- to 3-fold over-represented in these genomes as compared with the random sequences of the same base composition. Second, the 10–12 bp periodicity in the A-tract distribution is pronounced. Third, the characteristic length of the A-tract phasing is in the range 80–130 bp in each of the analyzed genomes.

The bacteria with GC-rich genomes naturally have lower A-tract occurrences. However, in some genomes these occurrences are 2-fold higher than in random sequences, and the 10–12 bp periodicity remains noticeable (e.g. *Pseudomonas aeruginosa*, 58% GC). However, in the extremely GC-rich genomes (e.g. *Mycobacterium tuberculosis*, 66% GC and *Streptomyces coelicolor*, 72% GC) the A-tract occurrences are close to or even lower than those in random sequences. Finally, in the extremely GC-poor genomes (*Bacillus anthracis*, 35% GC and *Staphylococcus aureus*, 33% GC) the A-tract occurrences are very high and, as a result, the periodicity peaks are widened significantly (data not shown). More detailed analysis of the A-tract distribution in various bacterial genomes is in progress and will be presented elsewhere.

DISCUSSION

Validation of the results

To ensure that our findings reflect the actual features of the genome organization and do not depend on the particular methods used in the analysis, we performed several tests. Specifically, we applied 'running' 3 bp averaging to exclude the 3 bp periodicity from the A-tract distance autocorrelations (18,45). Such a procedure neither led to the emergence of new peaks in the intensity versus period plot (Figure 5A) nor affected the characteristic length of the 10–12 bp periodicity for the A-tract distribution.

In addition, we analyzed the distribution of the A-tracts containing only adenines or thymines in one strand (such non-extendable sequences A_n:T_n comprise a particular case of the A_nT_m-tracts and are denoted here as A-runs, see Table 2). Naturally, the occurrences of A-runs were 2- to 3-fold lower than the occurrences of the A_nT_m-tracts (Table 1) and we observed smaller magnitudes of the peaks in the autocorrelation function calculated for A-runs (data not shown). Nevertheless, the results obtained for A-runs are qualitatively the same as the results for A_nT_m-tracts reported here. For comparison, results for the G-runs (sequences G_n:C_n) are also given in Table 2.

Throughout this study, we compared the A-tract distribution in bacterial genomes with that in the random sequences of the same base composition. In addition to this parameter, natural sequences are distinguished by the sets of dinucleotide relative abundances or 'genomic signatures' [reviewed in (46)]. In particular, for the *E.coli* genome, the ratio of (observed/expected) frequencies is 1.2 for the AA:TT dinucleotide. Therefore, to check the statistical significance of our results on the over-representation of A-tracts, the random sequences with the *E.coli* dinucleotide composition were generated and used for control (see Methods for further details).

As expected, the relative abundance of genomic A-tracts diminished compared with evaluations based on the individual nucleotide frequencies (cf. the 'genome/random' ratios for A_nT_m-tracts based on the dinucleotide content given in Table 2 with the corresponding data based on the individual nucleotide frequencies given in Figure 3). Nevertheless, deviations from random remain statistically significant (Table 2). The occurrences of the 4–8 bp long A-tracts in the genome differ from those in the random sequences with high level of significance ($P < 0.001$ for 4, 5, 6 and 8 bp long tracts; $P < 0.01$ for 7 bp tracts; *t*-test). In most of the cases, the A-tract occurrences are higher than expected by 9–14% [Table 2; $(n + m) = 5$ to 8]. Only the number of 4 bp long A-tracts is less than expected by 12%.

The above consideration is related to A-tracts in the form A_nT_m. However, if one considers the A-tracts containing only adenines or thymines in one strand (A-runs), then the A_4:T_4 runs are over-represented in the *E.coli* genome (Table 2). The A_n:T_n runs produce stronger curvature than do A_nT_m-tracts (28,30); therefore, their relative abundance would be critical for the DNA folding. Importantly, the (observed/expected) ratio increases up to 1.3–1.5 for the 5–7 bp A-runs, where the DNA curvature is the strongest and drops below unity for the longer A-runs. These data, taken together, strongly suggest that over-representation of the A-tracts is not a direct consequence of the well-known biases in the dinucleotide composition but rather reflects 'new' hitherto unreported features of the *E.coli* genome organization.

Finally, to test our results on the distribution of the phased A-tracts in the genome, we used three different sets of the twist angles (47–49) to calculate the A-tract curvature [note that the results shown in Figure 7 were obtained with the twist angles calculated by Kabsch *et al*. (47)]. All the three sets produced similar results, both in terms of locations of the phased A-tract clusters and in terms of the A-tract curvature values.

Table 2. Ratios of occurrences of the A- and G-tracts of different length in the *E.coli* genome and in random sequences with the same dinucleotide content[a]

Tract length (bp)	A-tracts[b] A_nT_m	A_n:T_n	G-tracts[b] G_nC_m	G_n:C_n
4	0.88***	1.05***	0.83***	0.71***
5	1.10***	1.34***	0.64***	0.55***
6	1.14***	1.48***	0.44***	0.34***
7	1.09**	1.26***	0.32***	0.28***
8	1.14***	0.95	0.15***	0.20***
9	0.78***	0.22***	0.05***	0.09
10	0.28***	0.00	0.12**	0.46

[a]The ratios 'genome/random' are averaged over 10 implementations of a random sequence with the same dinucleotide composition as in the genome. **Significance level $P < 0.01$ and ***significance level $P < 0.001$ (*t*-test).
[b]The results for tracts in the form A_nT_m or G_nC_m are shown separately from the tracts containing purines in one strand and pyrimidines in the other strand (A_n:T_n or G_n:C_n).

Comparison with the results of previous sequence analyses

Our observation that the short A-tracts are over-represented in the *E.coli* genome is consistent with the codon usage frequencies (50). It was shown that the codons constituting '3 bp A-tracts' are among the most frequently used ones in *E.coli*; namely, codons AAA, UUU, AUU and AAU occur at frequencies 33.6, 22.4, 30.4 and 17.7 per 1000 of used codons, respectively (the average frequency is 18.3). Also, one of the most frequent codons, GAA (39.6/1000), contains the AA dinucleotide as well. This preference in the codon usage may explain how a genome with 88% coding sequence can harbor such a large number of the A-tracts.

The high percentage of the coding sequences in the genome also explains the presence of the 3 bp periodicity in the distributions of both A- and G-tracts. However, the pronounced 10–12 bp periodicity was observed for the A-tracts only. This

is in accord with the previous *E.coli* genome studies (15,18,51) where the 11 bp periodicity was observed for the WW dinucleotides (W stands for A or T) but not for the GG dinucleotides, while the 3 bp periodicity was observed for both WW and GG dinucleotides. Another important observation by Herzel *et al*. (18) was that the 11 bp periodicity in the distribution of WW dinucleotides extended up to 100 bp, although the upper limit of this periodicity was not discussed. Here, for the first time, we provide a quantitative estimate of the periodicity length.

The genomic distribution of the phased A-tracts can be compared with the distribution of the DNA curvature estimated with the 'wedge' model (33). Before comparing the two sets of data, note that both approaches, ours and that based on the wedge model, have advantages and disadvantages. On the one side, the wedge model (33) takes into account all 16 DNA dimeric steps, but the dimeric 'wedge angles' are not consistent with the X-ray and NMR data, and the predicted magnitude of curvature is too high [for recent reviews see (25,32)]. On the other side, the empirical parameter 'A-tract curvature' used here gives the values of DNA bending consistent with experimental data (43,44), but accounts only for the bends caused by the A-tracts interspersed with 'pseudo-random' sequences. Possible contribution from the other sequences is ignored. However, the phased A-tracts make the most pronounced contribution to DNA curvature. Furthermore, the G-tracts, the other sequence motifs capable of producing DNA curvature, are distributed quasi-randomly in the genome. Therefore, we believe that the 'A-tract curvature' correctly reflects the main tendencies in the 'real' curvature distribution in genomic DNA.

Based on the wedge model, it has been predicted earlier (21,22) that the intergenic regions, specifically promoters, are the most curved regions in bacterial genomes. We also observe that the 'concentration' of the phased A-tracts is higher in the intergenic regions compared with the coding regions in *E.coli* genome. In addition, we show that the phased A-tracts clusters are abundant in the coding sequences, resulting in a significantly higher A-tract curvature of the coding DNA than it would be expected from the base composition. This novel observation indicates that the DNA curvature due to A-tracts is a general characteristic of the *E.coli* genomic DNA, not limited to the regulatory regions.

Implications in bacterial genome packaging

The non-random distribution of A-tracts may provide a structural basis for the bacterial chromosome packaging. As shown above, the A-tracts are over-represented in the genome and demonstrate pronounced 10–12 bp periodicity. This periodicity is close to the DNA helical repeat and therefore is likely to produce the DNA intrinsic curvature (11). More importantly, the phased A-tracts are grouped into clusters, thereby greatly increasing the 'local concentration' of these A-tracts in genomic DNA. Accordingly, the energy cost of DNA looping is significantly lower for the A-tract containing fragments compared with the intrinsically straight ones (25).

When DNA is supercoiled (as in the bacterial nucleoid), the presence of the intrinsically curved fragments would facilitate branching of the plectonemic superhelix (52,53). Owing to the A-tracts, such branching would occur at the specific positions and, thus, it may direct the DNA compaction in a pre-arranged manner (Figure 8). In frame of this model, the promoters, as the most curved fragments of genomic DNA, would frequently appear at the apexes of the superhelical branches, which could increase the accessibility of the promoters for transcription factors and RNA polymerase (54,55).

Furthermore, we found that the clusters of phased A-tracts are omnipresent in the genome, providing both the intergenic and the coding regions with 'looping potential'. Thus, the highly non-random distribution of the A-tracts could constitute a genome-wide structural code for DNA packaging in *E.coli*. The same traits of the A-tract distribution occur in several other bacteria (see below), indicating that such structural code may represent a widespread molecular mechanism involved in the bacterial nucleoid organization.

In turn, the under-representation of the G-tracts in the *E.coli* genome (Figure 3) may be related to the increased propensity of these sequence motifs to adopt the A-DNA conformation (56,57). A-DNA is less flexible than B-DNA in terms of both bending and twisting and, therefore, the excessive number of G-tracts may hinder the nucleoid compaction by producing the extended stretches of the A-conformation in genomic DNA. Another reason for the small number and 'quasi-random' distribution of G-tracts may be that the requirement for protein coding leaves room for only one 'folding code' hidden in the *E.coli* genome, namely, the A-tract phasing.

The structural code observed in the present study is consistent with the experimental data on the MNase digestion, which provides information on the characteristic sizes of the most abundant structural components in the bacterial nucleoid. The results of the previous studies (58,59) and those of a recent systematic digestion analysis (K.M. Virnik, M.Y. Tolstorukov, V.B. Zhurkin and S. Adhya, manuscript in preparation) suggest that the elementary structural unit of the nucleoid from actively dividing cells has the size of 100–120 bp. This value is close to the characteristic size of the phased A-tract clusters estimated here. This size, $\sim$100 bp, is also very close to the sizes of the regulatory *gal*- and *lac*-loops (113 and 92 bp, respectively). Therefore, we suggest that folding and regulatory loops may have similar structural organization and, moreover, the same loops may play both regulatory and packaging roles in the nucleoid (Figure 8B).

However, the structural properties of the DNA alone are hardly sufficient for the 1000-fold DNA compaction. Rather, the DNA structural code may complement other packaging mechanisms, such as the involvement of architectural proteins (HU, H-NS, Hfq, CbpA, etc.), some of which have specific affinity to the curved DNA (60). Thus, the clusters of phased A-tracts are likely to be operative in the packaging of the bacterial chromosome in two ways: (i) facilitate forming of the small DNA loops and (ii) serve as 'binding sites' for the nucleoid-associated proteins (22).

Comparison with other bacteria

Analysis of the A-tract distribution in different bacteria allows drawing several inferences. (i) The A-tract phasing and clustering in *S.typhimurium* are nearly the same as in *E.coli*. In addition, these closely related bacteria have similar sets of the DNA-binding proteins (61). Therefore, it is likely that their chromosomes are packaged similarly. Overall, the similarity of the A-tract distributions in *E.coli* and *S.typhimurium* substantiates our hypothesis on the importance of A-tracts

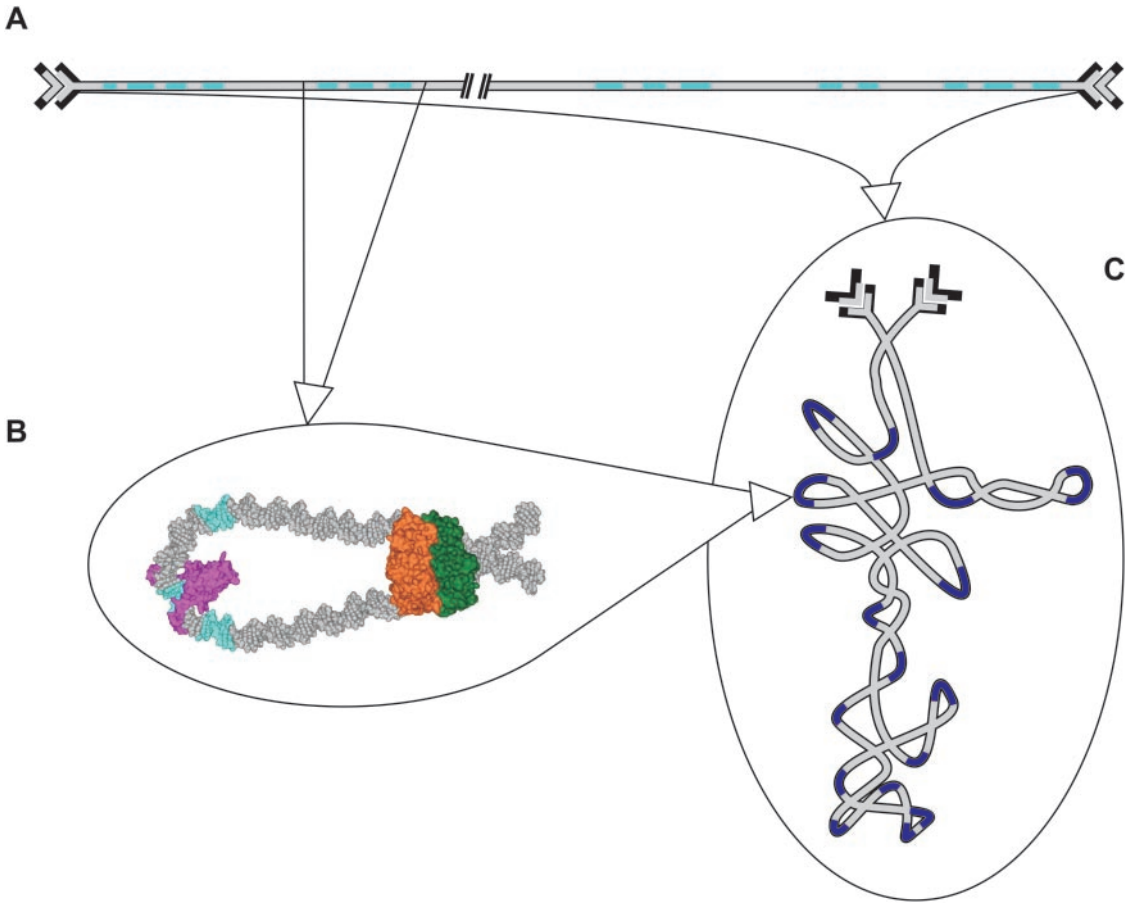

Figure 8. Model of the A-tract assisted compaction of the bacterial chromosome. (**A**) Schematic representation of the A-tract distribution in a fragment of the bacterial genome. A-tracts are shown in cyan and non-A-tract DNA is shown in gray. Note that A-tracts are grouped in clusters. (**B**) Putative model of the bacterial 'compactosome' (64): a cluster of the phased A-tracts introduces DNA curvature, facilitating DNA looping by the architectural proteins. The color code used for the A-tracts and non-A-tract DNA is the same as in (A). Proteins assist in DNA bending (magenta) and secure the loop closure (ochre and green). This scheme is drawn by analogy with the *gal*-loop, where HU protein facilitates DNA bending, and Gal repressors secure the loop closure (65,66). (**C**) Under superhelical stress, the clusters of phased A-tracts (dark blue) would facilitate branching of the plectonemically supercoiled DNA, appearing at the apexes of the branches. Thus, the phased A-tract clusters may constitute a code for the sequence-directed packaging of the bacterial chromosome within the domains of supercoiling.

for DNA packaging. (ii) Clustering of the phased A-tracts is inherent not only to genomes of *E.coli* and closely related bacteria. We observed similar A-tract distribution in *B.halodurans* (fermicutes) and *N.meningitides* (β-proteobacteria), which are quite distant from *E.coli* (γ-proteobacteria).

It is obvious that the proposed A-tract-related mechanism of the DNA packaging is not operative in all bacteria, specifically in those with extremely high GC-content. It is conceivable that these bacteria use different mechanisms of chromosome condensation. Indeed, it is known that bacteria differ significantly in the composition of DNA-binding proteins (61,62), and in the DNA sequence organization near gene starts (22). Thus, the structural signals in DNA sequence, if any, are also likely to be different.

CONCLUSIONS

We investigated distributions of the A- and G-tracts in *E.coli* and several other bacterial genomes and observed that: (i) short A-tracts are 2- to 3-fold over-represented in the genomes, compared with random sequences, while the G-tracts are under-represented; (ii) the A-tract distribution demonstrates periodicity of 10–12 bp, indicating that the A-tracts are often phased with the DNA helical repeat; (iii) the phased A-tracts are grouped into ∼100 bp clusters; (iv) such clusters are present throughout the genomes, including the coding sequences. Our results suggest that the non-random distribution of A-tracts in a bacterial genome may constitute the structural code for DNA condensation into a nucleoid. The genomic DNA fragments, which are intrinsically curved due to the presence of the A-tract clusters, would serve as basic elements of DNA packaging in bacteria, facilitating DNA looping and directing interactions with architectural proteins.

ACKNOWLEDGEMENTS

The authors are grateful to Edward Trifonov and Amos Oppenheim for valuable discussions. Funding to pay the Open Access publication charges for this article was provided by National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Pettijohn,D.E. and Sinden,R.R. (1985) Structure of the isolated nucleoid. In Nanninga,N. (ed.), *Molecular Cytology of Escherichia coli*. Academic Press, London, pp. 199–227.
2. Drlica,K. and Rouviere-Yaniv,J. (1987) Histonelike proteins of bacteria. *Microbiol. Rev.*, **51**, 301–319.
3. Robinow,C. and Kellenberger,E. (1994) The bacterial nucleoid revisited. *Microbiol. Rev.*, **58**, 211–232.
4. Case,R.B., Chang,Y.P., Smith,S.B., Gore,J., Cozzarelli,N.R. and Bustamante,C. (2004) The bacterial condensin MukBEF compacts DNA into a repetitive, stable structure. *Science*, **305**, 222–227.
5. Kleppe,K., Ovrebo,S. and Lossius,I. (1979) The bacterial nucleoid. *J. Gen. Microbiol.*, **112**, 1–13.
6. Worcel,A. and Burgi,E. (1972) On the structure of the folded chromosome of *Escherichia coli*. *J. Mol. Biol.*, **71**, 127–147.
7. Holmes,V.F. and Cozzarelli,N.R. (2000) Closing the ring: links between SMC proteins and chromosome partitioning, condensation, and supercoiling. *Proc. Natl Acad. Sci. USA*, **97**, 1322–1324.
8. Zhuang,X. (2004) Molecular biology. Unraveling DNA condensation with optical tweezers. *Science*, **305**, 188–190.
9. Trifonov,E.N. (1998) 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Physica A*, **249**, 511–516.
10. Rozenberg,H., Rabinovich,D., Frolow,F., Hegde,R.S. and Shakked,Z. (1998) Structural code for DNA recognition revealed in crystal structures of papillomavirus E2-DNA targets. *Proc. Natl Acad. Sci. USA*, **95**, 15194–15199.
11. Trifonov,E.N. (1985) Curved DNA. *CRC Crit. Rev. Biochem.*, **19**, 89–106.
12. Matthews,K.S. (1992) DNA looping. *Microbiol. Rev.*, **56**, 123–136.
13. Adhya,S., Geanacopoulos,M., Lewis,D.E., Roy,S. and Aki,T. (1998) Transcription regulation by repressosome and by RNA polymerase contact. *Cold Spring Harb. Symp. Quant. Biol.*, **63**, 1–9.
14. Trifonov,E.N. and Sussman,J.L. (1980) The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl Acad. Sci. USA*, **77**, 3816–3820.
15. Zhurkin,V.B. (1981) Periodicity in DNA primary structure is defined by secondary structure of the coded protein. *Nucleic Acids Res.*, **9**, 1963–1971.
16. Widom,J. (1996) Short-range order in two eukaryotic genomes: relation to chromosome structure. *J. Mol. Biol.*, **259**, 579–588.
17. Herzel,H., Trifonov,E.N., Weiss,O. and Grosse,I. (1998) Interpreting correlations in biosequences. *Physica A*, **249**, 95–120.
18. Herzel,H., Weiss,O. and Trifonov,E.N. (1999) 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics*, **15**, 187–193.
19. VanWye,J.D., Bronson,E.C. and Anderson,J.N. (1991) Species-specific patterns of DNA bending and sequence. *Nucleic Acids Res.*, **19**, 5253–5261.
20. Jauregui,R., O'Reilly,F., Bolivar,F. and Merino,E. (1998) Relationship between codon usage and sequence-dependent curvature of genomes. *Microb. Comp. Genomics*, **3**, 243–253.
21. Gabrielian,A.E., Landsman,D. and Bolshoy,A. (1999–2000) Curved DNA in promoter sequences. *In Silico Biol.*, **1**, 183–196.
22. Bolshoy,A. and Nevo,E. (2000) Ecologic genomics of DNA: upstream bending in prokaryotic promoters. *Genome Res.*, **10**, 1185–1193.
23. Hagerman,P.J. (1990) Sequence-directed curvature of DNA. *Annu. Rev. Biochem.*, **59**, 755–781.
24. Harrington,R.E. (1992) DNA curving and bending in protein–DNA recognition. *Mol. Microbiol.*, **6**, 2549–2555.
25. Zhurkin,V.B., Tolstorukov,M.Y., Xu,F., Colasanti,A.V. and Olson,W.K. (2005) Sequence-dependent variability of B-DNA: An update on bending and curvature. In Ohyama,T. (ed.), *DNA Conformation and Transcription*. Landes Bioscience, Texas and Springer, NY, pp. 18–34. http://www.eurekah.com/abstract.php?chapid=2024&bookid=151&catid=30.
26. Marini,J.C., Effron,P.N., Goodman,T.C., Singleton,C.K., Wells,R.D., Wartell,R.M. and Englund,P.T. (1984) Physical characterization of a kinetoplast DNA fragment with unusual properties. *J. Biol. Chem.*, **259**, 8974–8979.
27. Hagerman,P.J. (1986) Sequence-directed curvature of DNA. *Nature*, **321**, 449–450.
28. Hagerman,P.J. (1988) Sequence-dependent curvature of DNA. In Wells,R.D. and Harvey,S.C. (eds), *Unusual DNA Structures*. Springer-Verlag, NY, pp. 225–236.
29. Koo,H.S., Wu,H.M. and Crothers,D.M. (1986) DNA bending at adenine thymine tracts. *Nature*, **320**, 501–506.
30. Koo,H.S. and Crothers,D.M. (1988) Calibration of DNA curvature and a unified description of sequence-directed bending. *Proc. Natl Acad. Sci. USA*, **85**, 1763–1767.
31. Crothers,D.M., Haran,T.E. and Nadeau,J.G. (1990) Intrinsically bent DNA. *J. Biol. Chem.*, **265**, 7093–7096.
32. Beveridge,D.L., Dixit,S.B., Barreiro,G. and Thayer,K.M. (2004) Molecular dynamics simulations of DNA curvature and flexibility: helix phasing and premelting. *Biopolymers*, **73**, 380–403.
33. Bolshoy,A., McNamara,P., Harrington,R.E. and 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.
34. Maroun,R.C. and Olson,W.K. (1988) Base sequence effects in double-helical DNA. III. Average properties of curved DNA. *Biopolymers*, **27**, 585–603.
35. Calladine,C.R., Drew,H.R. and McCall,M.J. (1988) The intrinsic curvature of DNA in solution. *J. Mol. Biol.*, **201**, 127–137.
36. Goodsell,D.S., Kaczor-Grzeskowiak,M. and Dickerson,R.E. (1994) The crystal structure of C-C-A-T-T-A-A-T-G-G. Implications for bending of B-DNA at T-A steps. *J. Mol. Biol.*, **239**, 79–96.
37. Brukner,I., Dlakic,M., Savic,A., Susic,S., Pongor,S. and Suck,D. (1993) Evidence for opposite groove-directed curvature of GGGCCC and AAAAA sequence elements. *Nucleic Acids Res.*, **21**, 1025–1029.
38. Goodsell,D.S., Kopka,M.L., Cascio,D. and Dickerson,R.E. (1993) Crystal structure of CATGGCCATG and its implications for A-tract bending models. *Proc. Natl Acad. Sci. USA*, **90**, 2930–2934.
39. Hardy,G.H. and Rogosinski,W.W. (1999) *Fourier Series*. Dover Publications, Mineola, NY.
40. Gabrielian,A., Simoncsits,A. and Pongor,S. (1996) Distribution of bending propensity in DNA sequences. *FEBS Lett.*, **393**, 124–130.
41. Eisenberg,D., Schwarz,E., Komaromy,M. and Wall,R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.*, **179**, 125–142.
42. Koo,H.S., Drak,J., Rice,J.A. and Crothers,D.M. (1990) Determination of the extent of DNA bending by an adenine-thymine tract. *Biochemistry*, **29**, 4227–4234.
43. Tchernaenko,V., Halvorson,H.R. and Lutter,L.C. (2004) Topological measurement of an A-tract bend angle: effect of magnesium. *J. Mol. Biol.*, **341**, 55–63.
44. Tchernaenko,V., Radlinska,M., Drabik,C., Bujnicki,J., Halvorson,H.R. and Lutter,L.C. (2003) Topological measurement of an A-tract bend angle: comparison of the bent and straightened states. *J. Mol. Biol.*, **326**, 737–749.
45. Kolker,E., Tjaden,B.C., Hubley,R., Trifonov,E.N. and Siegel,A.F. (2002) Spectral analysis of distributions: finding periodic components in eukaryotic enzyme length data. *OMICS*, **6**, 123–130.
46. Karlin,S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr. Opin. Microbiol.*, **1**, 598–610.
47. Kabsch,W., Sander,C. and Trifonov,E.N. (1982) The ten helical twist angles of B-DNA. *Nucleic Acids Res.*, **10**, 1097–1104.
48. Gorin,A.A., Zhurkin,V.B. and Olson,W.K. (1995) B-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.*, **247**, 34–48.
49. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
50. Nakamura,Y., Gojobori,T. and Ikemura,T. (2000) Codon usage tabulated from the international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.*, **28**, 292.
51. Herzel,H., Weiss,O. and Trifonov,E.N. (1998) Sequence periodicity in complete genomes of archaea suggests positive supercoiling. *J. Biomol. Struct. Dyn.*, **16**, 341–345.
52. Laundon,C.H. and Griffith,J.D. (1988) Curved helix segments can uniquely orient the topology of supertwisted DNA. *Cell*, **52**, 545–549.
53. Rippe,K., von Hippel,P.H. and Langowski,J. (1995) Action at a distance: DNA-looping and initiation of transcription. *Trends Biochem. Sci.*, **20**, 500–506.
54. Schneider,R., Lurz,R., Luder,G., Tolksdorf,C., Travers,A. and Muskhelishvili,G. (2001) An architectural role of the *Escherichia coli* chromatin protein FIS in organising DNA. *Nucleic Acids Res.*, **29**, 5107–5114.
55. Dorman,C.J. and Deighan,P. (2003) Regulation of gene expression by histone-like proteins in bacteria. *Curr. Opin. Genet. Dev.*, **13**, 179–184.

56. Minchenkova,L.E., Schyolkina,A.K., Chernov,B.K. and Ivanov,V.I. (1986) CC/GG Contacts facilitate the B to A transition of DNA in solution. *J. Biomol. Struct. Dyn.*, **4**, 463–476.
57. Tolstorukov,M.Y., Ivanov,V.I., Malenkov,G.G., Jernigan,R.L. and Zhurkin,V.B. (2001) Sequence-dependent B↔A transition in DNA evaluated with dimeric and trimeric scales. *Biophys. J.*, **81**, 3409–3421.
58. Varshavsky,A.J., Nedospasov,S.A., Bakayev,V.V., Bakayeva,T.G. and Georgiev,G.P. (1977) Histone-like proteins in the purified *Escherichia coli* deoxyribonucleoprotein. *Nucleic Acids Res.*, **4**, 2725–2745.
59. Lossius,I., Holck,A., Aasland,R., Haarr,L. and Kleppe,K. (1986) Proteins associated with chromatin from *Escherichia coli*. In Gualerzi,C.O. and Pon,C.L. (eds), *Bacterial Chromatin*. Springer-Verlag, Berlin, pp. 91–100.
60. Azam,T.A. and Ishihama,A. (1999) Twelve species of the nucleoid-associated protein from *Escherichia coli*. Sequence recognition specificity and DNA binding affinity. *J. Biol. Chem.*, **274**, 33105–33113.
61. Takeyasu,K., Kim,J., Ohniwa,R.L., Kobori,T., Inose,Y., Morikawa,K., Ohta,T., Ishihama,A. and Yoshimura,S.H. (2004) Genome architecture studied by nanoscale imaging: analyses among bacterial phyla and their implication to eukaryotic genome folding. *Cytogenet. Genome Res.*, **107**, 38–48.
62. Sandman,K., Pereira,S.L. and Reeve,J.N. (1998) Diversity of prokaryotic chromosomal proteins and the origin of the nucleosome. *Cell. Mol. Life Sci.*, **54**, 1350–1364.
63. Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
64. Kellenberger,E. (1991) Functional consequences of improved structural information on bacterial nucleoids. *Res. Microbiol.*, **142**, 229–238.
65. Aki,T. and Adhya,S. (1997) Repressor induced site-specific binding of HU for transcriptional regulation. *EMBO J.*, **16**, 3666–3674.
66. Geanacopoulos,M., Vasmatzis,G., Zhurkin,V.B. and Adhya,S. (2001) Gal repressosome contains an antiparallel DNA loop. *Nature Struct. Biol.*, **8**, 432–436.