2316–2327 *Nucleic Acids Research, 2006, Vol. 34, No. 8*
doi:10.1093/nar/gkl230

Involvement of DNA curvature in intergenic regions of prokaryotes

Limor Kozobay-Avraham, Sergey Hosid and Alexander Bolshoy*

Genome Diversity Center, Institute of Evolution, University of Haifa, Haifa 31905, Israel

Received January 8, 2006; Revised February 3, 2006; Accepted March 27, 2006

ABSTRACT

It is known that DNA curvature plays a certain role in gene regulation. The distribution of curved DNA in promoter regions is evolutionarily preserved, and it is mainly determined by temperature of habitat. However, very little is known on the distribution of DNA curvature in termination sites. Our main objective was to comprehensively analyze distribution of curved sequences upstream and downstream to the coding genes in prokaryotic genomes. We applied CURVATURE software to 170 complete prokaryotic genomes in a search for possible typical distribution of DNA curvature around starts and ends of genes. Performing cluster analyses and other statistical tests, we obtained novel results regarding various factors influencing curvature distribution in intergenic regions, such as growth temperature, A+T composition and genome size. We also analyzed intergenic regions between converging genes in 15 selected genomes. The results show that six genomes presented peaks of curvature excess larger than 3 SDs. Insufficient statistics did not allow us to draw further conclusion. Our hypothesis is that DNA curvature could affect transcription termination in many prokaryotes either directly, through contacts with RNA polymerase, or indirectly, via contacts with some regulatory proteins.

INTRODUCTION

Curved DNA is involved in many biologically important processes, such as transcription initiation (1–4) and termination (5), recombination (6), DNA replication (7), and nucleosome positioning (8). The studies of regulation of transcription, which involve DNA curvature, relate mainly to a specific gene, family of genes or promoters (9–11).

DNA curvature in prokaryotes is usually presented upstream of the promoter sequence (12) but sometimes within the promoter sequence (10,13) or even in two or more locations at the promoter region (12,14). In several studies wherein two upstream elements are observed, each of them stimulates transcription by a different mechanism. For example, in the *argT* promoter of *Escherichia coli* the proximal curved DNA favors RNAP binding, whereas in the same promoter, the distal element facilitating isomerization to the open complex during transcription (14). Other functions of DNA curvature, such as enhancing the affinity of the complex by forming a large loop around RNAP (12,15) and bringing together transcription elements that are located in inaccessible distance (15), were also suggested. These functions indicate that DNA curvature has diversified roles in gene regulation besides serving as a signal for the recognition of regulatory proteins.

The prokaryotes' cell components, including DNA, are influenced by environmental conditions of their habitat. For example, let us look at the three strains of *Prochlorococcus marinus* genomes from the viewpoint of their size. The strain that adapted to high light intensities had the smallest genome of any oxygenic phototroph, while the strain that adapted to low light had the largest genome (16). Likewise, different genomes can have different DNA curvature promoter profiles, reflecting various environmental factors as well as other aspects of genome organization. Promoters of prokaryotes, especially those of mesophilic bacteria, are, in general, preceded by DNA curvature (17–19), and phased A-tracts located upstream of promoter regions have been well documented as a paradigm illustrating the role of the promoter upstream curvature [reviewed in (1,12)]. The influence of temperature on intrinsic DNA curvature, expressed as an electrophoretic anomaly, has been studied previously (20,21). It was universally found that the effect of DNA curvature disappears with rising temperature. For example, Katayama *et al.* (22) found that DNA curvature upstream of the *plc* (phospholipase C gene) promoter in *Clostridium perfingens* stimulates its activity in a low-temperature dependent manner.

Another possible explanation of the influence of temperature on DNA conformation lies in the distribution of DNA

*To whom correspondence should be addressed. Tel./Fax: +972 4 8240382; Email. bolshoy@research.haifa.ac.il topoisomerases in bacteria and archaea. While mesophilic bacteria and archaea use gyrase to introducing negative supercoiled DNA in addition to a local unwinding to initiate DNA activity, hyperthermophiles use reverse gyrase for positive supercoiling. On one hand, in mesophilic prokaryotes, such additional local unwinding can be contributed by the UCS (upstream curved sequences) since curved DNA structures mimic a negative supercoil. On the other hand, in hyperthermophiles the high temperature may provide the strand-opening needed (23).

Comprehensive genome analysis of DNA curvature in regulatory regions was performed by us (5,17–19) and others (4,24,25). It was found that regulatory regions are significantly more curved from their neighboring coding regions and from expectations based on their dinucleotide composition. It is well established that in many prokaryotic genomes DNA intrinsic curvature upstream of promoters is related to the activity of the particular promoter (12,22,26–29). However, there are other genomes that do not present such UCS. The factors influencing distribution of DNA curvature have not been clearly identified and characterized yet. Moreover, evidence on excess of curved DNA downstream of the termination site that may be involved in the transcription termination process has been published only recently (5).

In this study our main objective is to comprehensively analyze downstream curved sequences in prokaryotic genomes and to establish the possible relationship between curvature in promoter and terminator regions. We also studied genomic and environmental factors influencing curvature distribution in these transcription regulation sites. The results showed that curved DNA downstream of the termination site is frequently found in mesophilic bacteria and archaea. Furthermore, genome size and A+T composition have an effect on DNA curvature excess in terminator regions but to a lesser degree than in promoter regions. The results point to correlation in regulation mechanisms between initiation and termination of the transcription, which are determined by environmental and genomic factors. Cluster analyses, backed by other statistical tests, allow us deeper assessment of the factors influencing DNA curvature distribution in promoter and terminator regions.

METHODS

Database

We compiled a database that can be divided into two parts: the first part, data gathered from the literature, and the second part, data obtained by our performed analysis. For every genome, we placed its characteristics in the database such as size in base pairs, number of genes, A+T content averaged over the complete genome, A+T content averaged over all non-coding sequences and A+T content of coding sequences (CDS). Some taxonomic descriptors (Kingdom, Phylum and Class) were also added to the database. Furthermore, results obtained for each genome were also recorded such as: values of curvature excess parameters, cluster belonging and percent of predicted hairpin terminators. The database can be found at our web site http://genome.haifa.ac.il/~limor/curved_prom_term.

Curvature calculation

The prediction of DNA curvature was made using our CURVATURE program. The program is available upon request from A. Bolshoy (bolshoy@research.haifa.ac.il). This program calculates a 3D path of a DNA molecule and estimates the curvature of the axis path (30). The CURVATURE algorithm is based on the stepwise calculation of geometric transformations according to the set of dinucleotide wedge angles (31,32). The whole genome sequence was used as an input to the CURVATURE program and a map of curvature distribution using a given window size of 125 bp along the whole sequence was produced. A curvature value at a position i corresponds to a curvature of the arc approximating to the predicted DNA path, when the arc approximates a path segment of the length of 125 bp with a center of the segment in the position i. The DNA curvature was measured in DNA curvature units (cu) introduced by Trifonov and Ulanovsky (33) and used in all our previous studies. For example, a segment of 125 bp of length with a shape close to a half-circle has a curvature value of about 0.34 cu. Such strongly curved pieces with values of >0.3 cu appear infrequently in genomic sequences.

Averaging predicted DNA curvature in the neighborhood of 5′ ends and 3′ ends of CDS

Automatic procedures of extraction utilized annotations of complete prokaryotic genome sequences in the GenBank, the public genome library of the National Center for Biotechnology Information. In our study of curvature distribution around the 5′ and 3′ ends of the CDS, we only processed CDS longer than 125 nt and flanked by intergenic regions longer than 125 nt. For every gene, we aimed to take a neighborhood with lengths of ±400 bases. However, we took only non-coding segments upstream of the starts and downstream of the ends; and only coding segments downstream of the starts and upstream of the ends. The mean value of predicted DNA curvature (g_i) and its SEs were calculated separately for coding and non-coding regions. The SEs (σ_i) were estimated by bootstrap method using 1000 runs.

Preparation of randomized genomes

We constructed control genomes with the same dinucleotide composition for genic and intergenic sequences separately. This procedure was made for testing the significance of the results and comparing properties of natural and artificial genomes. The construction procedure consisted of three steps: (i) a genome was cut in separate genic and intergenic pieces at every 5′ and 3′ gene junction; (ii) each piece was separately reshuffled preserving dinucleotide composition and (iii) all the pieces were reassembled in the original order. For every genome, we prepared 10 randomized control genomes using the above-mentioned procedure of shuffling and rejoining randomly reshuffled pieces. We estimated the magnitude of curvature of coding and non-CDS of the artificial genomes by averaging 10 randomized shuffled genomes (r_i). The program is available upon request from S. Hosid (hosid@research.haifa.ac.il). The distributions of the mean curvature of reshuffled sequences (dashed line in Figures 1a and 2a), either around the starts or ends of translation, were plotted for some genomes along with curvature distribution of the real genome.

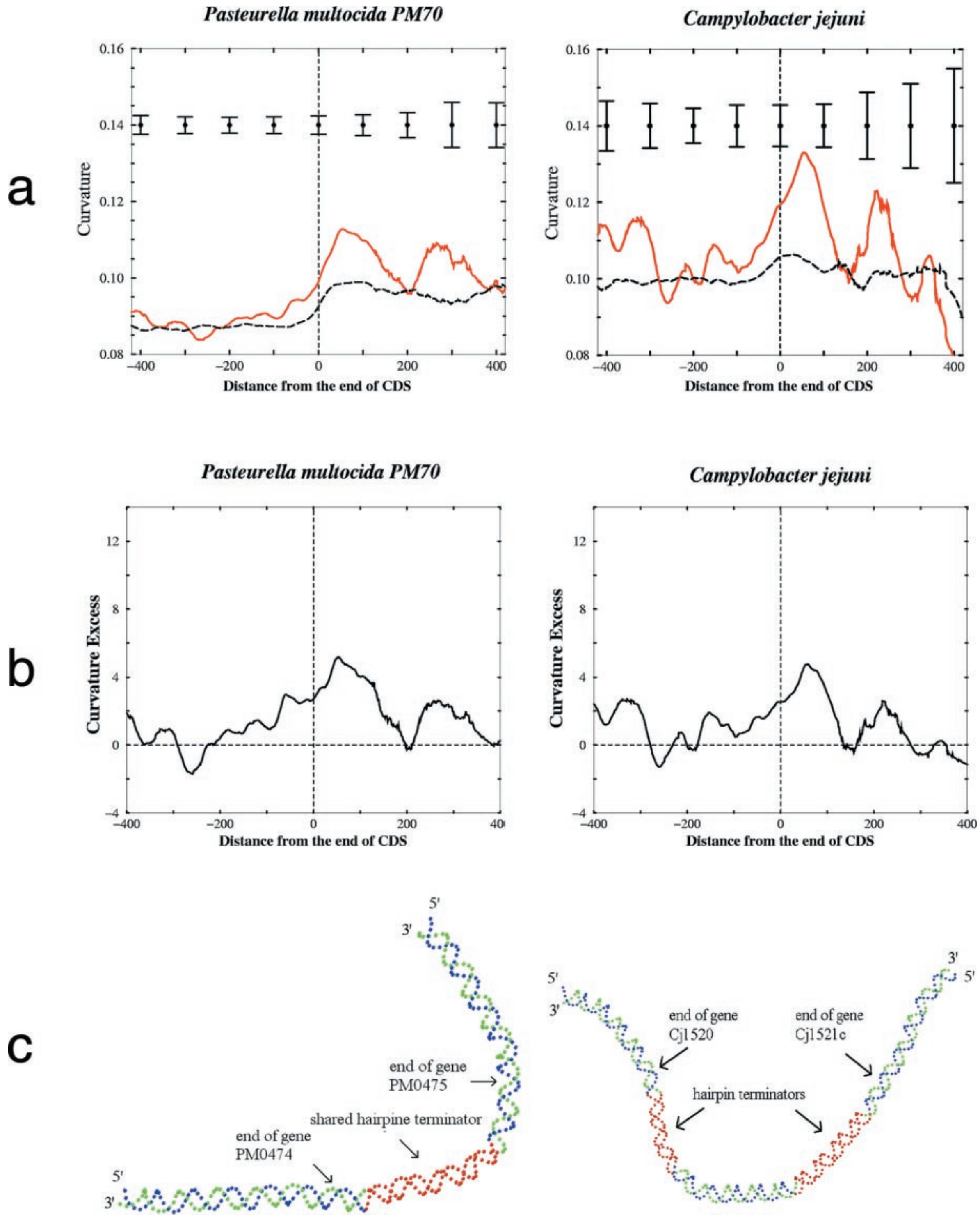

Figure 1. Curvature distributions of big AT-rich mesophilic genomes in the neighborhood of the end of translation. *P.multocida* and *C.jejuni* are representative of the group of mesophilic genomes with A+T composition over 50% in their non-coding regions and with genome size over 1.4 million bp. For each genome the sets of regions ±400 bases in length around the end of translation were compiled. Also, only genes longer than 125 nt and flanked by downstream intergenic regions longer than 125 nt were processed. The program CURVATURE with a window size of 125 nt was used to predict curvature distributions. (**a**) The *y*-axis represents the DNA curvature measured in curvature units (0.08/0.16 cu) and the *x*-axis represents the position around the end of translation. The red line represents a profile obtained by averaging the distributions of all fragments from the same genome. The SEs were estimated by the bootstrap method using 1000 runs. For better visibility, error bars corresponding to several distances around the 3′ end, are shown separately from the curvature maps. The dashed lines represent curvature distributions obtained by averaging the distribution of analogous shuffled fragments as explained in Methods. (**b**) The *y*-axis represents the curvature excess in standard deviation units (−4/13 SD). Curvature excess was obtained by estimation apparent deviation between genomic and random curvature values. (**c**) The 3D trajectories of DNA in terminator regions of chosen convergent genes are shown. Arrows indicate the positions of the stop codons. The coding strands are colored in green and the complementary strand in blue. In the region of the hairpin terminator predicted by Mfold program, the two strands of DNA were colored in red.

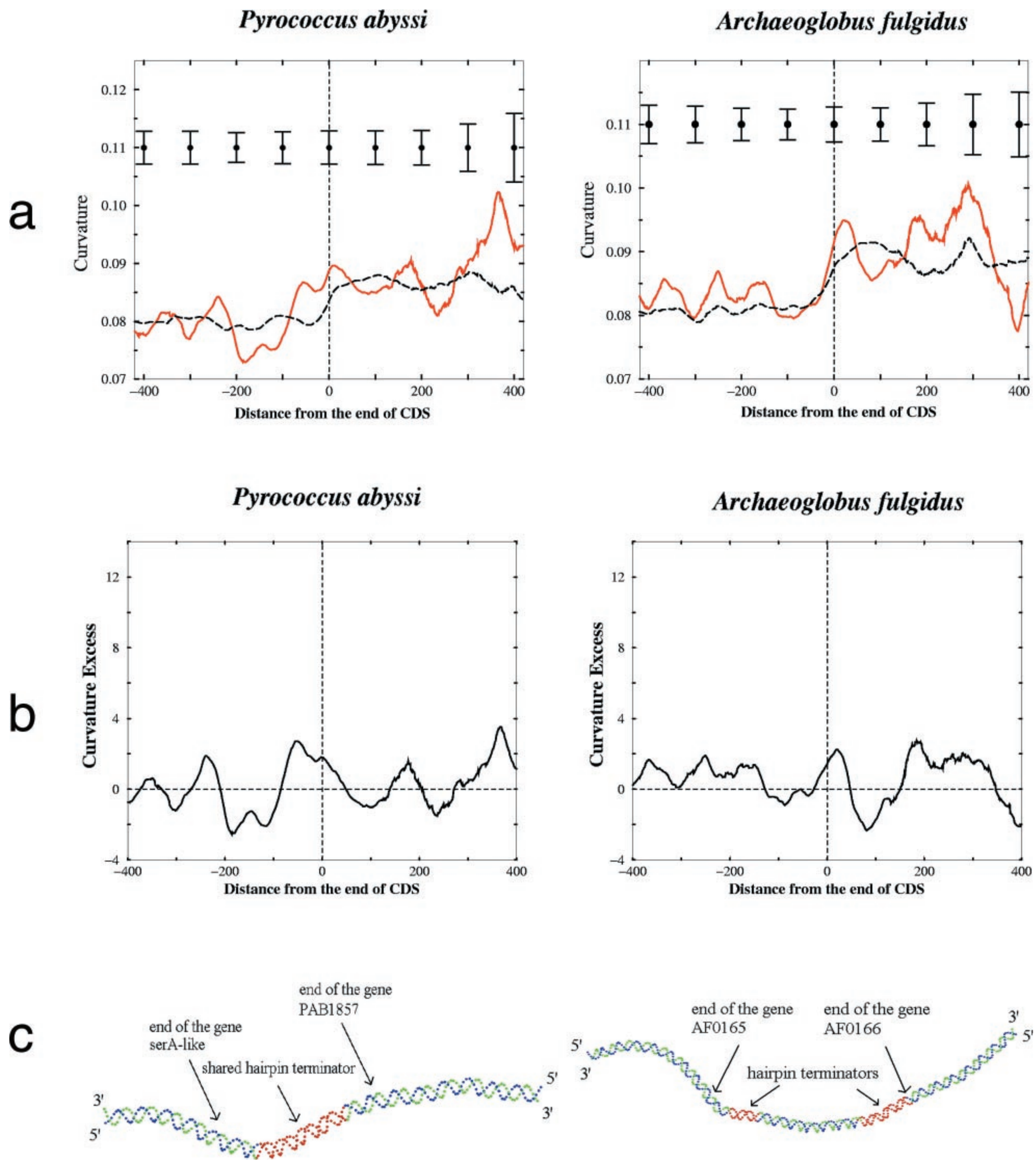

Figure 2. Curvature distributions of hyperthermophilic archaea and bacteria in the neighborhood of the end of translation. *Archaeoglobus fulgidus* and *P.abyssi* were representative of the group of hyperthermophilic genomes. DNA curvature calculations for the real and the reshuffled genomes were performed as described in the legend to Figure 1. (**a**) The *y*-axis represents the DNA curvature measured in curvature units (0.07/0.13 cu) and the *x*-axis represents the position around the end of translation. (**b**) The *y*-axis represents the curvature excess in SD units (−4/13 SD). Curvature excess was obtained by estimation apparent deviation between genomic and random curvature values. (**c**) The 3D trajectories of DNA in terminator region of chosen genes are shown. Arrows indicate the positions of the stop codons. The coding strands are colored in green and the complementary strand in blue. In the region of the hairpin terminator predicted by Mfold program, the two strands of DNA were colored in red.

Calculation of comparison parameters

Three comparison parameters of curvature excess (CE) were calculated and used for clustering analyses: maximal curvature excess (MCE), upstream integral excess (UIE) and downstream integral excess (DIE).

Curvature excess is an apparent deviation between genomic (g_i) and random (r_i) curvature values measured in SD units and calculated as follows:

$$\mathrm{CE} = \frac{(\mathrm{g}_i - \mathrm{r}_i)}{\sigma_i}.$$

The parameter MCE was determined by detecting the maximal CE value, either in 400 bp upstream of 5′ ends or downstream of 3′ ends of CDS, for all 170 genomes.

The two other parameters, UIE and DIE, represent the average in CE over X base pairs:

$$IE = \frac{\sum_{i=1}^{i=X} (g_i - r_i)/\sigma_i}{X}.$$

X in the parameter UIE was determined as 125 bp (between −63 and −188 nt) upstream to the 5′ ends of CDS, wherein promoters are usually located. However, X for the parameter DIE determined to be 100 bp, (between +1 and +100 nt) downstream to the 3′ end of CDS, wherein terminators sites are usually located.

Cluster analysis

K-means clustering algorithm (34) was operated over three clusters using the program SAS Enterprise. The parameter MCE, UIE and DIE calculated as explained above were used for cluster analysis procedures.

Convergent genes extraction

We chose 15 genomes with high curvature excess in terminator regions (DIE over 3 SD units) and extracted only their convergent genes. Furthermore, for every such genome we performed DNA curvature map around the ends of translation, as explained above, but only on genes with an intergenic region longer than 50 bp.

Prediction of rho-independent terminators

We used the GeSTer program based on the algorithm described in (35) for extracting the fraction of intrinsic hairpin terminators. The authors kindly made the installable version of the software free for non-commercial use at ftp.bork.embl-heidelberg.de/pub/users/suyama/GeSTer.

RESULTS

Genomic DNA curvature profiles in the neighborhood of the starts and ends of translation

We applied the software CURVATURE on 170 completed genomes to describe distributions of the DNA curvature around starts and ends of translation according to the wedge model (31). The plots that represent the curvature excess distributions of these regions for each genome are presented in our web site http://genome.haifa.ac.il/~limor/curved_prom_term.

Figures 1 and 2 show a few examples of curvature distribution and curvature excess around the ends of translation in typical mesophilic and hyperthermophilic genomes. The upper plots of each figure (Figures 1a and 2a) show the curvature profile of the genome around the ends of translations, along with curvature profile expected from the dinucleotide compositions (constructions of such profiles explained in Methods). The middle plots (Figures 1b and 2b) show the distribution of curvature excess in standard deviation units calculated as explained in Methods. In addition, for every genome a 3D trajectory of DNA around terminator of a chosen gene was performed (Figures 1c and 2c). The genes that were chosen to represent the 3D pathway of DNA in typical terminator regions are convergent genes: PM0474 and PM0475 of *Pasteurella multocida*, Cj1520 and Cj1521c of *Campylobacter jejuni*, SerA-like (Figure 1c) and PAB1857 of *Pyrococcus abyssi* and AF0165 and AF0166 of *Archaeglobus fulgidis* (Figure 2c). The intergenic regions of these pairs of genes exclusively contain terminators. Positions of putative hairpin terminators were predicted by the Mfold program (36). The results show that DNA curvature excess in terminator regions, as in promoter regions, is determined by temperature of habitat.

Cluster analysis of DNA curvature excess in promoter and termination regions of genomic sequences

We applied K-means algorithm over three clusters using the parameters MCE, UIE and DIE. Cluster analysis showed that the data variations in the clusters are smaller while using the UIE parameter comparing with the MCE parameter. This result indicates that clustering based on UIE (or DIE) parameters is less biased and more reliable than MCE. For this reason, and also because a similar comparative study using MCE was previously performed by Kozobay-Avraham *et al.* (19), we do not present clustering analyses results based on MCE here.

The mean values of UIE and DIE, and the amount of the genomes in every cluster are summarized in Figure 3.

A comparison of the corresponding clusters show that the mean curvature excess values are higher for upstream regions compared with downstream areas. For example, the mean value of cluster 3 using the parameter UIE is 7.8 SD units compared with 4.9 U using the DIE parameter. These results indicate that promoter regions are more curved, in average, than terminators. However, the amounts of genomes in the clusters, using promoter or terminator analysis, revealed very similar picture. Cluster 3 contains the smallest amount of the genomes: 16 genomes using UIE (from 10.4 to 6.4 SD units) and 30 genomes using DIE (from 7.3 to 2.0 SD units). Each of the other clusters, using the parameters UIE or DIE, contains about the same amount of genomes (from 70 to 78).

An average curvature excess profile was obtained for every cluster either in the neighborhood of the start of genes (Figure 4a) or end of genes (Figure 4b). The results show that the curvature profiles related to terminator regions are different from mirror reflections of the curvature profiles at the promoter regions. The differences exist both in the distances of the peaks from a reference point (5′ or 3′ end of a gene) and in the shape of the profile. Indeed, while the curvature excess profiles of clusters 2 and 3 in promoter regions present peaks located ~150 bp upstream to the start of genes the corresponding profiles in terminator regions present peaks located very close to the end of genes (~50 bp downstream to the end). Also, a rise of a curvature excess in promoter regions is spread over >250 bp followed by descent close to the start of the genes, while in terminator regions a rise starts before the end of gene.

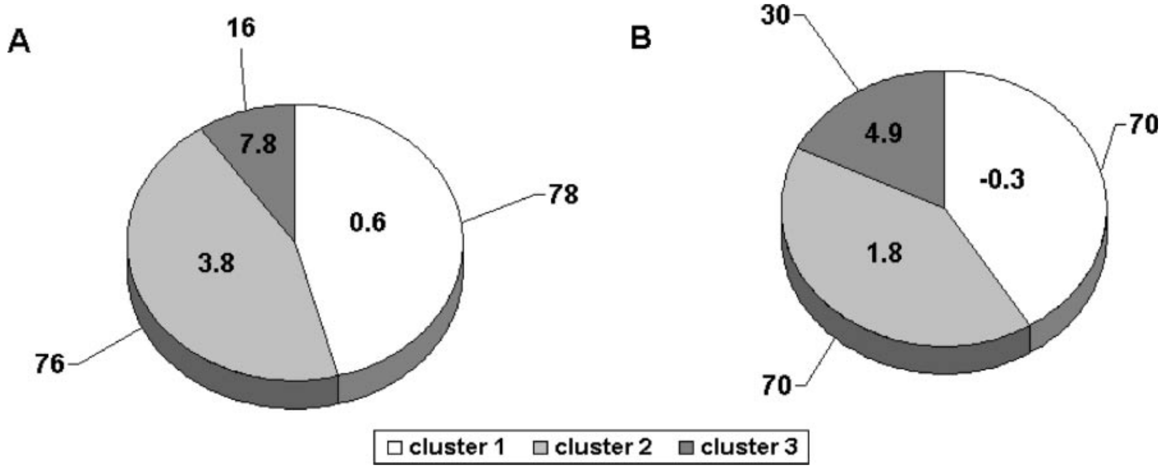

Figure 3. Cluster analysis. K-means algorithm over three clusters was carried out using the parameter Integral Excess in (**A**) promoter (UIE) and (**B**) terminator (DIE) regions of all 170 genomes. The numbers inside each of the pie's pieces represent the mean value of curvature excess of each cluster. An amount of the genomes in each cluster is indicated outside of the pie's pieces.

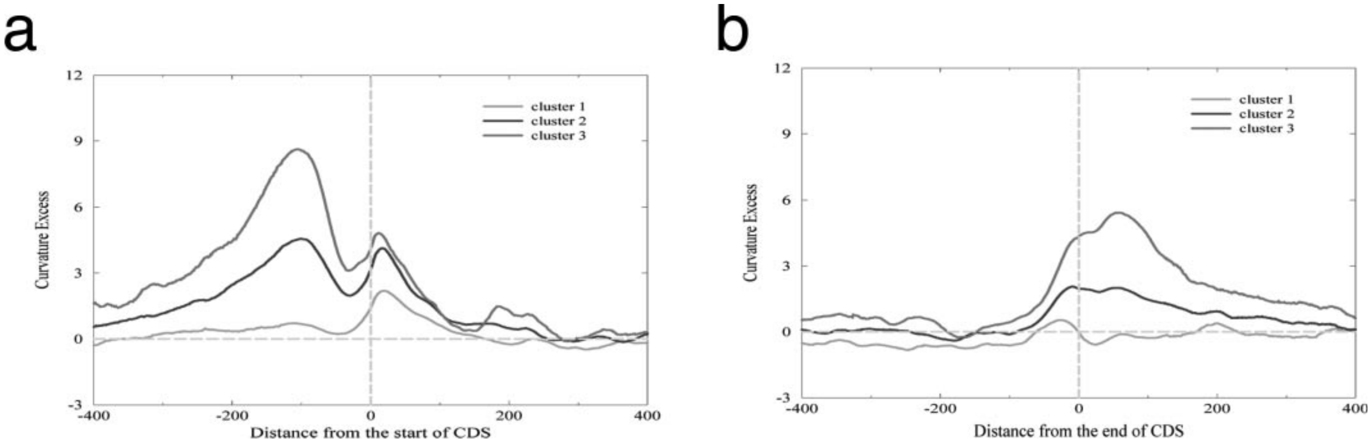

Figure 4. Mean profiles (centroids) of clusters. Genomic profiles based on curvature excess distribution in the neighborhood of the starts (**a**) and ends (**b**) of genes. The centroids are obtained by averaging all profiles related to each of the three clusters obtained by K-means algorithm. The *y*-axis represents the curvature excess in standard deviation units and the *x*-axis represents the position around the start or end of translation. The highest profiles are related to the cluster 3, and the lowest profiles correspond to the cluster 1.

Correlation analysis between environmental and genomic characteristics and curvature excess values in promoter and terminator regions

In the current study we tried to verify our previous qualitative findings regarding the factors that influence curvature distribution in promoter regions in a quantitative manner. We also set out to study the factors that influence curvature distribution in termination regions of different genomes.

From the data obtained by cluster analyses, we constructed three histograms of curvature excess for promoter (UIE) or terminator (DIE) regions, respectively (Figures 5 and 6). Each histogram is colored according to one characteristic: (i) growth temperature, (ii) genome size and (iii) A+T composition. Each pie plot presents the distribution of one characteristic in a particular cluster.

Cluster 3 has the highest mean values, either based on the curvature excess in upstream (UIE) or downstream (DIE) regions, and include only mesophilic bacteria that have a relatively 'big' genome size [over 1.4 Mb as we arbitrarily determined in our previous publication (19)] except chromosome II of *Vibrio cholerae* when using DIE. Moreover, all of the genomes in this cluster are AT-rich (above 50% A+T content in the non-coding region). Interestingly, none of these genomes belong to the 16 very AT-rich genomes (above 70% A+T). These 16 genomes are distributed between the other two clusters and most of them are 'small' genomes (Figures 5 and 6).

In Cluster 2 based on clustering using curvature excess in promoter regions (UIE), 75 out of the 76 organisms are mesophiles and 1 is a thermophile (Figure 5A). In this cluster, ~70% are 'big AT-rich' and the rest are either GC-rich or 'small' (Figure 5B and C). Cluster 1, which represents the lowest mean value of curvature excesses in promoter regions, contains the lowest percents of mesophiles compared to the other clusters (Figure 5A). The data based on the UIE show that this cluster contains all the hyperthermophiles and 99% percent of the thermophiles. Moreover, among the 51 mesophiles only 7 are 'big AT-rich'.

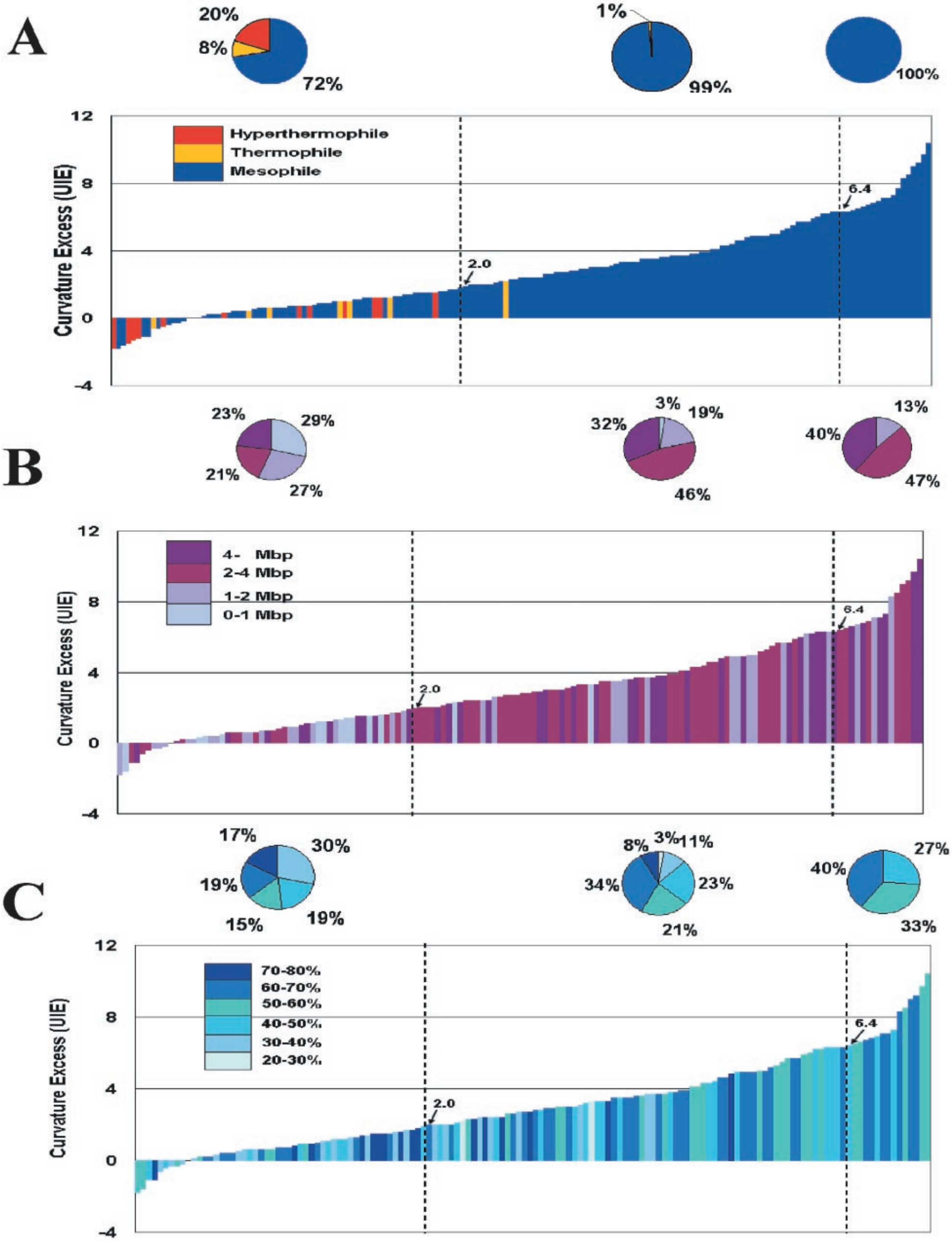

Figure 5. Histograms of curvature excess in promoter regions. Coloring of each histogram represent distributions of one characteristic along the clusters: (**A**) optimal growth temperature (OGT), (**B**) genome size and (**C**) A+T composition. The parameter UIE, which was calculated for the purely upstream window (from nt −63 to −188), was used to build the histograms. In histogram A all the genomes are represented, excluding seven genomes with unknown OGT. In histograms (B) and (C) only mesophilic genomes are represented. Curvature excess threshold of each cluster is indicated in the left side of the vertical dashed line. Above each histogram three pie plots are presented for better visibility of the character distribution in each cluster.

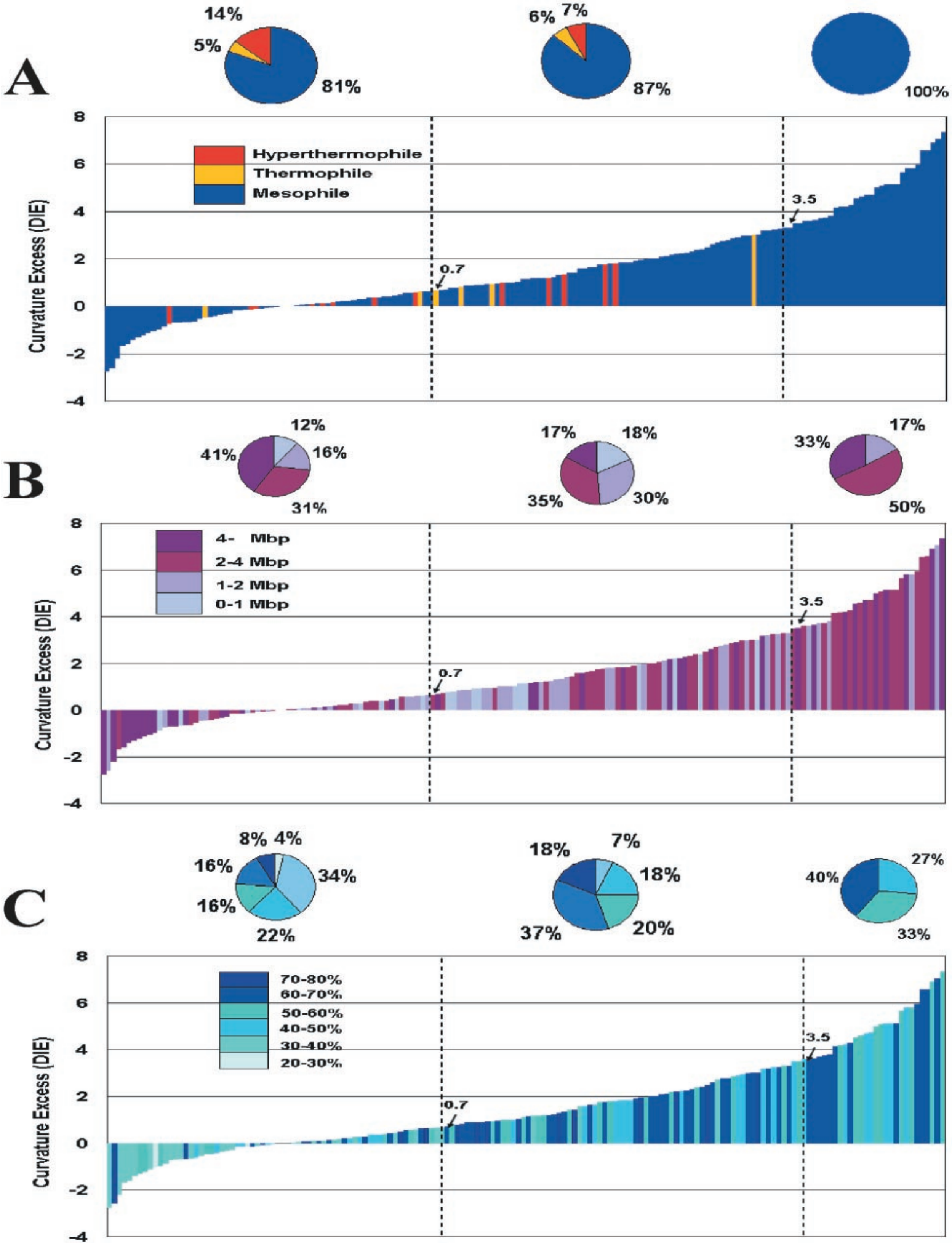

Figure 6. Histograms of curvature excess in terminators regions. Colorings of each histogram represent distribution of one characteristic along the three clusters: (**A**) optimal growth temperature (OGT), (**B**) genome size and (**C**) A+T composition. The parameter DIE, which was calculated for the first 100 bp after the stop codon, was used to build the histograms. In histogram A all the genomes are represented, except excluding genomes with unknown OGT. In histograms (B) and (C) only mesophilic genomes are represented. Curvature excess threshold of each cluster is indicated in the left side of the vertical dashed line. Above each histogram three pie plots are presented for better visibility of the character distribution in each cluster.

Clusters 1 and 2 based on clustering using curvature excess in terminator regions (DIE) consist of 70 genomes each. While examining these clusters, the most prominent difference from the clustering based on promoters is the distribution of the thermophilic and hyperthermophilic genomes in the clusters. This group shows homogeneous distribution: 12 genomes were included in cluster 1 (which represents the lowest mean value) and 9 genomes in cluster 2 (the medium). Another difference can be seen in the distribution of the 'small' mesophilic genomes. While performing cluster analysis on the promoters' curvature (UIE), cluster 1 includes smaller mesophilic genomes than cluster 2, which corresponds with the expectations mentioned in Kozobay-Avraham *et al.* (19). Performing the same analysis on terminators, we found that the relationship between genome size and curvature in terminator regions is weaker than in promoter regions.

Another interesting picture can be seen when examining the phyla distribution at every cluster. Clustering based on UIE curvature forms cluster 3 with 14 genomes out of 16 belonged to three groups that were known to be AT-rich: *Firmicutes* and *Proteobacteria* gamma and epsilon subdivisions, which are more AT-rich than the other members of the *Proteobacteria* phyla. A similar picture can be seen using the parameter DIE; ~70% of the genomes presented in cluster 3 belong to these phyla. At clusters 2 and 1, based on the parameter UIE, the picture is more heterogenic. However, some interested points can still be drawn, i.e. all the seven members of the phylum *Chlamydia* belong to cluster 1. In this phylum, the variation in size is relatively very small—from about 1 to 2.4 Mb. On the contrary, if we look at all 7 genomes of *Cyanobacteria* wherein the variation of size is very big—from about 1.6 to 6.4, or the 13 genomes of the *Actinobacteria*—from 0.9 to 9 Mb, the cluster distribution is less homogeneous. This phenomenon indicates a relationship between the size of the genomes and clustering, based on curvature excess at the promoter regions. As we expected, all the members of the kingdom Archaea, except the two mesophilic *Methanosarcinas*, *Methanosarcinas acetivorans* and *Methanosarcinas mazei*, are included in cluster 1, which represents the lowest mean value using UIE or DIE.

Curvature distribution in the vicinity of 'pure' terminators

There are three types of gene pairs relatively to the directions of transcription: unidirectional, convergent and divergent. These different gene arrangements generate three classes of spacers that differ in terms of the types of regulatory sites they contain. Spacers between unidirectional genes may include both a terminator for the upstream gene and promoter signals for the downstream gene; spacers between divergent genes have only promoter; and spacers between convergent genes exclusively contain terminators. Therefore, to eliminate any possible influence of curved promoters on terminator regions, we studies curvature excess in spacers between the convergent genes of 15 different genomes. These genomes were chosen due to relatively high curvature excess—above 3 SD units—presented in their termination sites. Distribution of curvature around the ends of convergent genes is presented in Figure 7. The results show that the statistics of convergent genes is much poorer than of all intergenic regions indicated by the relatively large errors-bars (shown separately from the curvature graphs). While we estimated the differences between observed maximal curvature values and expected values over all 3′ end regions, the differences in six genomes (*Bacillus subtilis*, *Bacteroides thetaiotaom VPI-5482*, *Corynebacterium glutamicum ATCC 13032*, *E.coli K12*, *M.acetivorans* and *Neisseria meningiti Z2491)* out of the 15 are still significant—above 3 SEs.

Correlation between DNA curvature in promoter and terminator regions

The Pearson correlation coefficient showed strong positive linear correlation between curvature excess in promoter (UIE) and terminator (DIE) regions, $r = 0.8$ with $P < 0.05$. A natural assumption would be that a hyperthermophilicity is a major reason of the correlation. The elimination of the thermophilic and hyperthermophilic genomes decreases the correlation to 0.74, indeed. Tests show that genome sizes and G+C content influence a magnitude of the correlation as well. Whether a mechanism of a termination of transcription is a factor?

In light of insubstantial evidence that sequences downstream of factor-independent termination site may influence transcription termination, we investigated the relationship between this termination mechanism and curvature in promoter regions. For this purpose we used the software GeSTer (35) over most of the genomes to predict putative hairpin terminators in the genomes. Then, we calculated the correlation between the amount of such hairpin terminators in the genomes and curvature excess in promoter regions (UIE). The result showed a positive correlation between the UIE and the percent of hairpin terminators ($r = 0.47$). The amount of the predicted hairpin terminators as well as the numbers of genes in each genome is also placed in our database at http://genome.haifa.ac.il/~limor/curved_prom_term.

DISCUSSION

The effect of intrinsic curvature upstream of a bacterial promoter on the efficiency of transcription was first reported in the early 1980s. Today, a countless number of examples are known indicating the importance of curved DNA sequence during different steps of transcription. Most of the examples show the vital role of DNA curvature in regulating the transcription initiation process (4,12) and only a few studies showed the role of DNA curvature on the termination process. These studies mainly discuss the role of DNA curvature in termination of reverse transcription in retroviruses (37,38). To the best of our knowledge, only one computational study showed that DNA curvature is very common in termination regions of the 'model' bacteria *E.coli and B.subtilis* (5). The role of such a phenomenon in this regulation site remains unclear. Nevertheless, a strong positive correlation between DNA curvature in promoter and terminator regions and other analyses leads us to the conclusion that this phenomenon is influenced by the same factors (optimal growth temperature, genome size and A+T composition) as those of promoter regions, but to a lesser extent way. This study aimed to answer the question whether there is any connection between DNA curvature in promoters and terminators or, in other words,

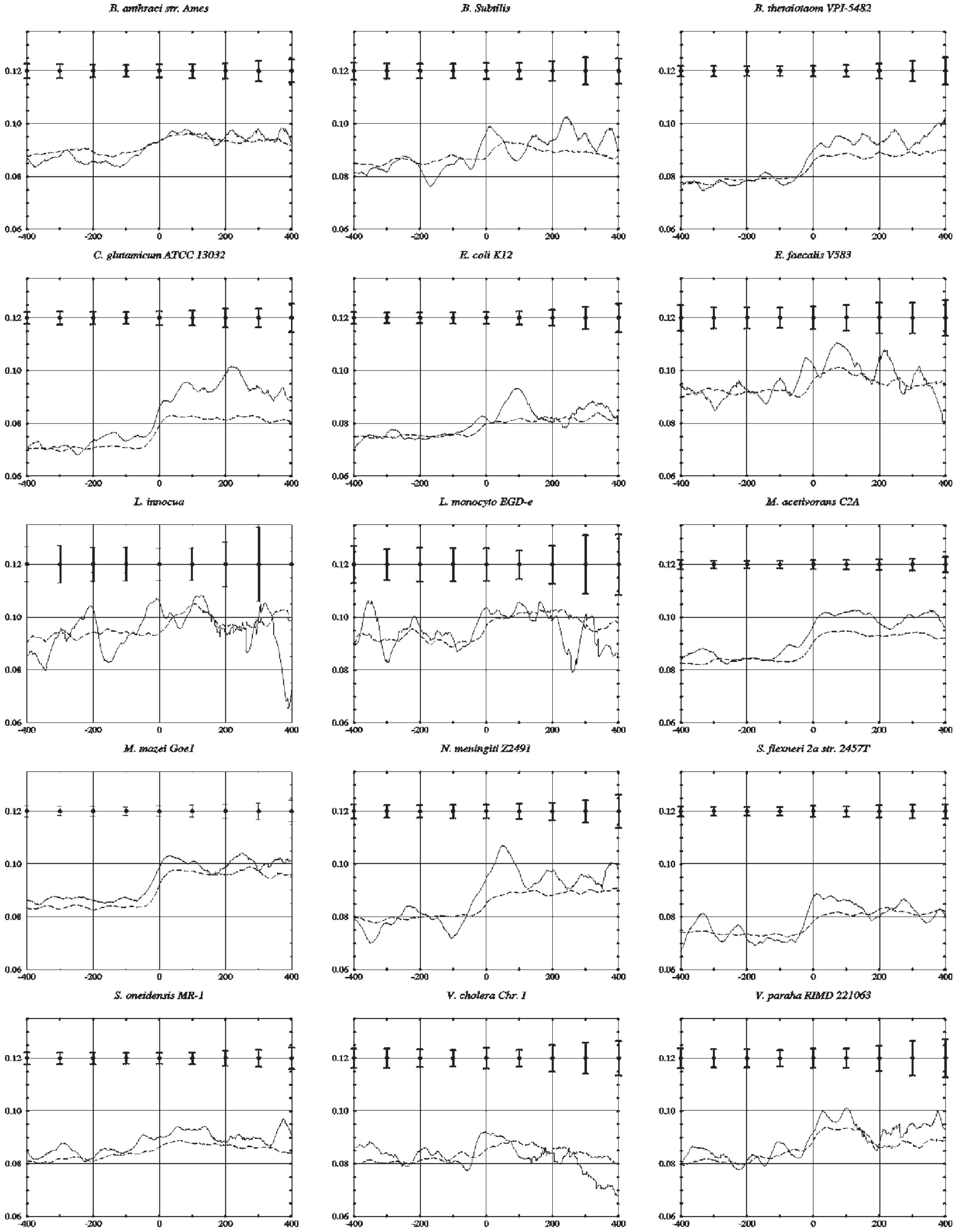

Figure 7. Curvature distributions in the neighborhood of the end of translation in convergent genes. From 15 selected genomes, which had relatively high curvature excess (DIE), only convergent genes were extracted. The program CURVATURE with a window size of 125 nt was used to predict curvature distributions. Only genes with intergenic region longer than 50 nt were processed. *Bacillus anthraci str. Ames* (781 genes), *B.subtilis* (543 genes), *B.thetaiotaom VPI-5482* (578 genes), *C.glutamicum ATCC 13032* (488 genes), *E.coli K12* (632 genes), *Enterococcus faecalis V583* (327 genes), *Listeria innocua* (290 genes), *Listeria monocyto EGD-e* (301 genes), *M.acetivorans C2A* (755 genes), *M.mazei Goe1* (534 genes), *N.meningiti Z2491* (277 genes), *Shigella flexneri 2a str. 2457T* (626 genes), *Shewanella oneidensis MR-1* (778 genes), *V.cholera* chromosome I (781 genes) and *V.paraha RIMD 221063* chromosome I (477 genes). DNA curvature calculations for the real and reshuffled convergent genes of every genome were performed as described in the legend to Figure 1.

whether regulation of transcription initiation and termination has something in common. Owing to an extensive amount of data we were able to define the factors influencing the curvature distribution in promoter and terminator regions.

Two hypotheses related to the role of curvature in the termination mechanism may be proposed. One possible hypothesis is that curved DNA directly slows down the RNA polymerase progression. The second hypothesis suggests an indirect role of the downstream curved sequences that may serve as binding sites to some proteins, histone-like protein (H-NS) or a similar auxiliary protein, contribute to a more effective termination of transcription. A recent publication by Tolstorukov *et al.* (39) showed a long-range DNA intrinsic curvature in *E.coli* genome upholding the hypotheses regarding the possible role of DNA curvature in bacterial genomes. The authors suggested that not only does DNA curvature serve as a packaging code but it also directly interacts with architectural proteins and facilitates DNA looping. In the publication of Azam and Ishihama (40) dual function of DNA curvature was also presented wherein some curved DNA-binding proteins, such as HNS, serve as global regulators of gene functions as well as structural proteins for compacting genomes. Sequences downstream of the termination site, which have not yet been transcribed, can considerably effect termination. The exact way in which downstream sequences effect intrinsic termination is unclear; however, since no consensus elements can be drawn, this finding has led to the conclusion that the stability and/or the conformation of the DNA downstream to the transcript hairpin region can also affect termination (41). Intrinsic DNA curvature, which is well known to activate promoter and efficient transcription in prokaryotes, may be a good candidate in affecting termination efficiency as well. Hosid and Bolshoy (5) were the first to show that DNA curvature is very common in termination site in both kinds of terminators, rho-dependent or independent, in *E.coli* and *B.subtilis*. Here, we found that most of the 'big' AT-rich mesophilic genomes, bacteria and archaea, showed non-random distribution of DNA curvature in the vicinity of the termination site. This evolutionary conservation pattern may provide new evidence for a biological signal in regulating termination of transcription.

In every prokaryotic genome a majority of non-coding regions are spacers between genes oriented in the same direction. Is it possible that a rise of the curvature excess around a 3′ end is related to a following 5′ end? In the case that only convergent genes were analyzed for curvature distribution in 'pure' terminators, for some genomes the statistics was too poor for drawing conclusions. However, beside the two genomes (*E.coli* and *B.subtilis)*, which were mentioned in Hosid and Bolshoy (5) having non-random distribution of curvature in termination sites of convergent genes, we found 4 more genomes out of 15 that were analyzed (*B.thetaiotaom VPI-5482*, *C.glutamicum ATCC 13032* and *N.meningiti Z2491*) presenting very clearly the same phenomenon. Moreover, for all genomes of cluster 3 (Figure 4) we can mention that aligning the curvature excess profiles according to 3′ ends, we obtain rather narrow peaks that is a clear indication that such an alignment is meaningful. The difference in the profiles is also resulted in the location of the peaks. While the downstream peaks of profiles are located very close to the ends of translation, the peaks of the promoter regions are not located immediately before the start of translation. This finding can be explained by the variation in the distance between transcription and translation points of start versus end of genes in prokaryotes. These differences were taken into consideration while calculating integrative values we considered distinct neighborhoods for upstream and downstream regions (UIE versus DIE).

In our previous publication we showed that the most prominent phenomenon is that an abundance of UCS in a genome was determined by temperature of its habitat. Other characteristics, such as genome size and A+T composition, also influence this phenomenon. In the current study, cluster analyses and other statistical tests, which were applied on extensive data of predicted curvature excess distributions, allow us to determine the contribution of such characteristics in promoter and termination regions. The results verified our arbitrary sorting of the genomes according to the following order: growth temperature, genome size and A+T composition. When we take into consideration the three factors that influence curvature in regulation sites, peculiar observations can be explained. For example, there was no representation of very AT-rich (above 70%) genomes found in cluster 3 (with the higher mean value), but underlying genome size revealed that most of these are 'small' genomes. The characterization of the factors influencing curved DNA at the regulation sites of transcription, in addition to sequence motifs, can greatly improve the currently available promoter and gene prediction algorithm.

ACKNOWLEDGEMENTS

This work was partially performed during the sabbatical stay of A.B. in Indiana University School of Informatics. The research was supported by a grant from the United States-Israel Binational Science Foundation BSF 2003291. We would like to thank Ms. Robin Permut for the editing. Funding to pay the Open Access publication charges for this article was provided by University of Haifa.

Conflict of interest statement. None declared.

REFERENCES

1. Ohyama,T. (2001) Intrinsic DNA bends: an organizer of local chromatin structure for transcription. *Bioessays*, **23**, 708–715.
2. Schroth,G.P., Siino,J.S., Cooney,C.A., Thng,J.P.H., Ho,P.S. and Bradbury,E.M. (1992) Intrinsically bent DNA flanks both sides of an RNA polymerase-I transcription start site—both regions display novel electrophoretic mobility. *J. Biol. Chem.*, **267**, 9958–9964.
3. Angermayr,M., Oechsner,U., Gregor,K., Schroth,G.P. and Bandlow,W. (2002) Transcription initiation *in vivo* without classical transactivators: DNA kinks flanking the core promoter of the housekeeping yeast adenylate kinase gene, AKY2, position nucleosomes and constitutively activate transcription. *Nucleic Acids Res.*, **30**, 4199–4207.
4. Olivares-Zavaleta,N., Jauregui,R. and Merino,E. (2006) Genome analysis of *Escherichia coli* promoter sequences evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated. *Genomics*, **87**, 329–337.
5. Hosid,S. and Bolshoy,A. (2004) New elements of the termination of transcription in prokaryotes. *J. Biomol. Struct. Dyn.*, **22**, 347–354.
6. Mazin,A., Milot,E., Devoret,R. and Chartrand,P. (1994) Kin17, a mouse nuclear-protein, binds to bent DNA fragments that are found at illegitimate recombination junctions in mammalian-cells. *Mol. Gen. Genet.*, **244**, 435–438.

7. Ueguchi,C., Kakeda,M., Yamada,H. and Mizuno,T. (1994) An analog of the Dnaj molecular chaperone in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **91**, 1054–1058.
8. Kiyama,R. and Trifonov,E.N. (2002) What positions nucleosomes?—a model. *FEBS Lett.*, **523**, 7–11.
9. Atlung,T., Knudsen,K., Heerfordt,L. and Brondsted,L. (1997) Effects of sigma(S) and the transcriptional activator AppY on induction of the *Escherichia coli* hya and cbdAB-appA operons in response to carbon and phosphate starvation. *J. Bacteriol.*, **179**, 2141–2146.
10. Espinosaurgel,M. and Tormo,A. (1993) Sigma(S)-dependent promoters in *Escherichia coli* are located in DNA regions with intrinsic curvature. *Nucleic Acids Res.*, **21**, 3667–3670.
11. Carmona,M. and Magasanik,B. (1996) Activation of transcription at sigma 54-dependent promoters on linear templates requires intrinsic or induced bending of the DNA. *J. Mol. Biol.*, **261**, 348–356.
12. Perez-Martin,J., Rojo,F. and de Lorenzo,V. (1994) Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. *Microbiol. Rev.*, **58**, 268–290.
13. Kaji,M., Matsushita,O., Tamai,E., Miyata,S., Taniguchi,Y., Shimamoto,S., Katayama,S., Morita,S. and Okabe,A. (2003) A novel type of DNA curvature present in a Clostridium perfringens ferredoxin gene: characterization and role in gene expression. *Microbiol. Sgm*, **149**, 3083–3091.
14. Hsu,L.M., Giannini,J.K., Leung,T.W.C. and Crosthwaite,J.C. (1991) Upstream sequence activation of *Escherichia coli* Argt promoter *in vivo* and *in vitro*. *Biochemistry*, **30**, 813–822.
15. Matthews,K.S. (1992) DNA looping. *Microbiol. Rev.*, **56**, 123–136.
16. Ussery,D.W. and Hallin,P.F. (2004) Genome update: length distributions of sequenced prokaryotic genomes. *Microbiol. Sgm*, **150**, 513–516.
17. Gabrielian,A.E., Landsman,D. and Bolshoy,A. (1999-2000) Curved DNA in promoter sequences. *In Silico Biol.*, **1**, 183–196.
18. Bolshoy,A. and Nevo,E. (2000) Ecologic genomics of DNA: upstream bending in prokaryotic promoters. *Genome Res.*, **10**, 1185–1193.
19. Kozobay-Avraham,L., Hosid,S. and Bolshoy,A. (2004) Curvature distribution in prokaryotic genomes. *In Silico Biol.*, **4**, 0029.
20. Diekmann,S. (1987) Temperature and salt dependence of the gel migration anomaly of curved DNA fragments. *Nucleic Acids Res.*, **15**, 247–265.
21. Ussery,D.W., Higgins,C.F. and Bolshoy,A. (1999) Environmental influences on DNA curvature. *J. Biomol. Struct. Dyn.*, **16**, 811–823.
22. Katayama,S., Matsushita,O., Jung,C.M., Minami,J. and Okabe,A. (1999) Promoter upstream bent DNA activates the transcription of the Clostridium perfringens phospholipase C gene in a low temperature-dependent manner. *EMBO J.*, **18**, 3442–3450.
23. Lopez-Garcia,P. (1999) DNA supercoiling and temperature adaptation: a clue to early diversification of life? *J. Mol. Evol.*, **49**, 439–452.
24. 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.
25. Jauregui,R., Abreu-Goodger,C., Moreno-Hagelsieb,G., Collado-Vides,J. and Merino,E. (2003) Conservation of DNA curvature signals in regulatory regions of prokaryotic genes. *Nucleic Acids Res.*, **31**, 6770–6777.
26. Bracco,L., Kotlarz,D., Kolb,A., Diekmann,S. and Buc,H. (1989) Synthetic curved DNA sequences can act as transcriptional activators in *Escherichia coli*. *EMBO J.*, **8**, 4289–4296.
27. McAllister,C.F. and Achberger,E.C. (1989) Rotational orientation of upstream curved DNA affects promoter function in *Bacillus subtilis*. *J. Biol. Chem.*, **264**, 10451–10456.
28. Travers,A.A. (1990) Why bend DNA. *Cell*, **60**, 177–180.
29. Gartenberg,M.R. and Crothers,D.M. (1991) Synthetic DNA bending sequences increase the rate of *in vitro* transcription initiation at the *Escherichia coli* Lac-promoter. *J. Mol. Biol.*, **219**, 217–230.
30. Shpigelman,E.S., Trifonov,E.N. and Bolshoy,A. (1993) CURVATURE: software for the analysis of curved DNA. *Comput. Appl. Biosci.*, **9**, 435–440.
31. 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.
32. Kabsch,W., Sander,C. and Trifonov,E.N. (1982) The 10 helical twist angles of B-DNA. *Nucleic Acids Res.*, **10**, 1097–1104.
33. Trifonov,E.N. and Ulanovsky,L.E. (1987) In Wells,R.D. and Harvey,S.C. (eds), *Unusual DNA Structures*. Springer–Verlag, Berlin, pp. 173–187.
34. MacQueen,J.B. (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, Vol. 1, pp. 281–297.
35. Unniraman,S., Prakash,R. and Nagaraja,V. (2002) Conserved economics of transcription termination in eubacteria. *Nucleic Acids Res*, **30**, 675–684.
36. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
37. Lavigne,M., Roux,P., Buc,H. and Schaeffer,F. (1997) DNA curvature controls termination of plus strand DNA synthesis at the centre of HIV-1 genome. *J. Mol. Biol.*, **266**, 507–524.
38. Lavigne,M. and Buc,H. (1999) Compression of the DNA minor groove is responsible for termination of RNA synthesis by HIV-1 reverse transcriptase. *J. Mol. Biol.*, **285**, 977–995.
39. Tolstorukov,M.Y., Virnik,K.M., Adhya,S. and Zhurkin,V.B. (2005) A-tract clusters may facilitate DNA packaging in bacterial nucleoid. *Nucleic Acids Res.*, **33**, 3907–3918.
40. 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.
41. Wagner,R. (2000) *Transcription Regulation in Prokaryotes*. Oxford University Press Inc., NY.