BMC Molecular Biology

BioMed Central

Research article

Sequence periodicity of *Escherichia coli* is concentrated in intergenic regions

Sergey Hosid, Edward N Trifonov and Alexander Bolshoy*

Address: Genome Diversity Center, Institute of Evolution, University of Haifa, Mt. Carmel 31905 ISRAEL

Email: Sergey Hosid - hosid@research.haifa.ac.il; Edward N Trifonov - trifonov@research.haifa.ac.il; Alexander Bolshoy* - bolshoy@research.haifa.ac.il

* Corresponding author

Published: 26 August 2004

BMC Molecular Biology 2004, **5**:14 doi:10.1186/1471-2199-5-14

Received: 31 December 2003
Accepted: 26 August 2004

This article is available from: http://www.biomedcentral.com/1471-2199/5/14

Abstract

Background: Sequence periodicity with a period close to the DNA helical repeat is a very basic genomic property. This genomic feature was demonstrated for many prokaryotic genomes. The *Escherichia coli* sequences display the period close to 11 base pairs.

Results: Here we demonstrate that practically only ApA/TpT dinucleotides contribute to overall dinucleotide periodicity in *Escherichia coli*. The noncoding sequences reveal this periodicity much more prominently compared to protein-coding sequences. The sequence periodicity of ApC/GpT, ApT and GpC dinucleotides along the *Escherichia coli* K-12 is found to be located as well mainly within the intergenic regions.

Conclusions: The observed concentration of the dinucleotide sequence periodicity in the intergenic regions of *E. coli* suggests that the periodicity is a typical property of prokaryotic intergenic regions. We suppose that this preferential distribution of dinucleotide periodicity serves many biological functions; first of all, the regulation of transcription.

Background

DNA sequence periodicity with the period about 10–11 base pairs (bp) has been long known in eukaryotic DNA sequences. It was discovered recently in prokaryotic sequences as well [1-6]. The periodicity in Eubacteria sequences usually shows the period close to 11 bp [1]. This period is clearly different from the structural helical period of 10.5–10.6 bp/turn [7,8]. The difference was interpreted [1,2] as a possible reflection of the sequence dependent writhe of prokaryotic DNA. In the work [9] it was demonstrated that the periodicity in the bacterial genomes, in *E. coli* as well, is distributed in a non-uniform way, in scattered segments of the size 100–150 bases. It was also known for a long time that quite a few DNA promoter regions of *E. coli* possess the sequence periodicity of AA and TT dinucleotides [10].

The sequence periodicity of AA/TT dinucleotides is frequently associated with sequence-dependent DNA curvature, which is known to play an important role in the initiation of transcription of many genes (for reviews, see [11-15]). Using different models and approaches for prediction of intrinsic DNA curvature it was shown that many *E. coli* promoters have upstream curved sequences [16,17]. Pedersen et al. [18] showed that promoter area frequently has an unusual sequence structure. This region possesses higher DNA curvature, more rigid and less stable. Moreover, in our study of prokaryotic terminators of transcription (Hosid and Bolshoy, submitted) we have

found that in *E. coli* DNA curvature peaks are frequently located downstream of the CDS.

Since the dinucleotide periodicity with the period close to the helical repeat is associated with DNA intrinsic curvature [19-23], the curvature distribution along DNA would suggest similar distribution of DNA sequence periodicity.

In this work, the sequence dinucleotide periodicity in *E. coli* and its distribution along the genome are systematically analyzed. A strong preference of intergenic regions to express the sequence periodicity of AA, AC, GC, and TT dinucleotides is discovered.

Results and Discussion

Positional autocorrelation analysis of the nucleotide sequences is an appropriate tool to detect all major characteristic distances in the sequences, the periodicities in particular. The complete genome of *E. coli*, as well as its coding and noncoding regions, was subjected to this procedure. Resulting autocorrelation profiles for all 16 dinucleotides (data not shown) were further analyzed by Fourier transform. In Fig. 1 the corresponding spectra are shown. The analysis demonstrates presence of the sequence periodicity of AA and TT dinucleotides with a period close to 11 bp mostly in intergenic regions, and weaker periodicity of AC and GC notably exclusively in intergenic regions. All 16 dinucleotides show periodicity of 3 bp, a well-known characteristics of the coding sequences, e.g. [24,25]. Weak 2 bp periodicity of AT and TA is also observed in intergenic regions. It indicates, perhaps, presence of tandem ApT repeats. A weak 10 bp periodicity of GC in intergenic regions, probably, corresponds to terminator regions (work in progress). The amplitudes of the 11 bp periodicity of AA and TT are the highest, even comparable with 3 bp coding periodicity. We, thus, focused on AA and TT distributions.

To screen the genome of *E. coli* and find out where the periodical regions are located, we chose the period 11.2 bp [1,2,5] and this study (Fig. 1); and the window of 150 bp [9,26]. We used periodical AA and TT probes with the above periodicity to correlate with the *E. coli* genome sequence and to detect the periodical sites. This calculation shows that the periodicity is not evenly distributed along the *E. coli* genome.

In Fig. 2, the typical maps for several large segments of the *E. coli* genome are shown. The periodicity is distinctly located in certain regions. Many of the peaks observed are found to correspond to the intergenic regions (indicated by the black bars at the top). For example, two such peaks of periodicity in Fig. 2a correspond to the intergenic regions. Three such maxima are observed in Fig. 2b, three in Fig. 2c, and two in Fig. 2d. For the genome sections in Fig. 2 about 2/3 of the intergenic regions are associated with the local periodicity.

To verify the apparent strong correlation between the intergenic regions and AA/TT periodicity, we split intergenic regions in several families by size and analyzed the subsets separately by aligning (centering) the regions and summing up the respective local periodicity distributions. The combined maps for intergenic regions with a size from 50 to 150 bp, from 150 to 250 bp, from 250 to 350 bp, from 350 to 450 bp, and from 450 to 550 bp are shown in Fig. 3. This figure demonstrates, indeed, that intergenic regions are typically periodic, irrespective of the size. The average amplitudes of the observed periodicities – 0.1–0.25 units – are comparable with the amplitudes in Fig. 2, which indicates that, indeed a large proportion of the intergenic regions are periodical.

To verify the choice of the period 11.2 bases, we calculated the periodicity maps for highly populated group of the regions of the size 200 ± 50 bp, by assuming different periods in the range 10.5–12.5 bases. The resonance 3D plot in Fig. 4 indicates that the best-fit period is 11.3 ± 0.4 bp, which confirms earlier estimates of the *E. coli* DNA sequence periodicity.

The spectral analysis (Fig. 1) and examples of the periodicity distribution maps (Fig. 2) show that apart from described correlation among the intergenic regions and AA/TT periodicity, there are numerous sites of periodicity located within coding sequences. Work is in progress to find out the functional relevance, if any, of these sites.

Conclusions

The observed concentration of the sequence periodicity in the intergenic regions corroborates earlier results and suggests that the periodicity is a typical property of the intergenic regions.

Methods

Genome data

The sequence of the whole genome of *Escherichia coli* K-12 MG1655, locus U00096, 4639221 base pairs, was taken from the National Center of Biotechnology Information ftp://ftp.ncbi.nih.gov/genbank/genomes. Intergenic regions were identified in accordance with the annotation to this genome of *E. coli* and gathered in a separate dataset.

Fourier transform of positional autocorrelation function

Autocorrelation profile X was calculated for each dinucleotide separately. For the calculation of ApA autocorrelation, for example, we calculated the number of occurrences of pairs ApA – ApA in a distance k, and designated it by X_k. Spectral analysis of autocorrelation profile X was obtained using the following formulae:

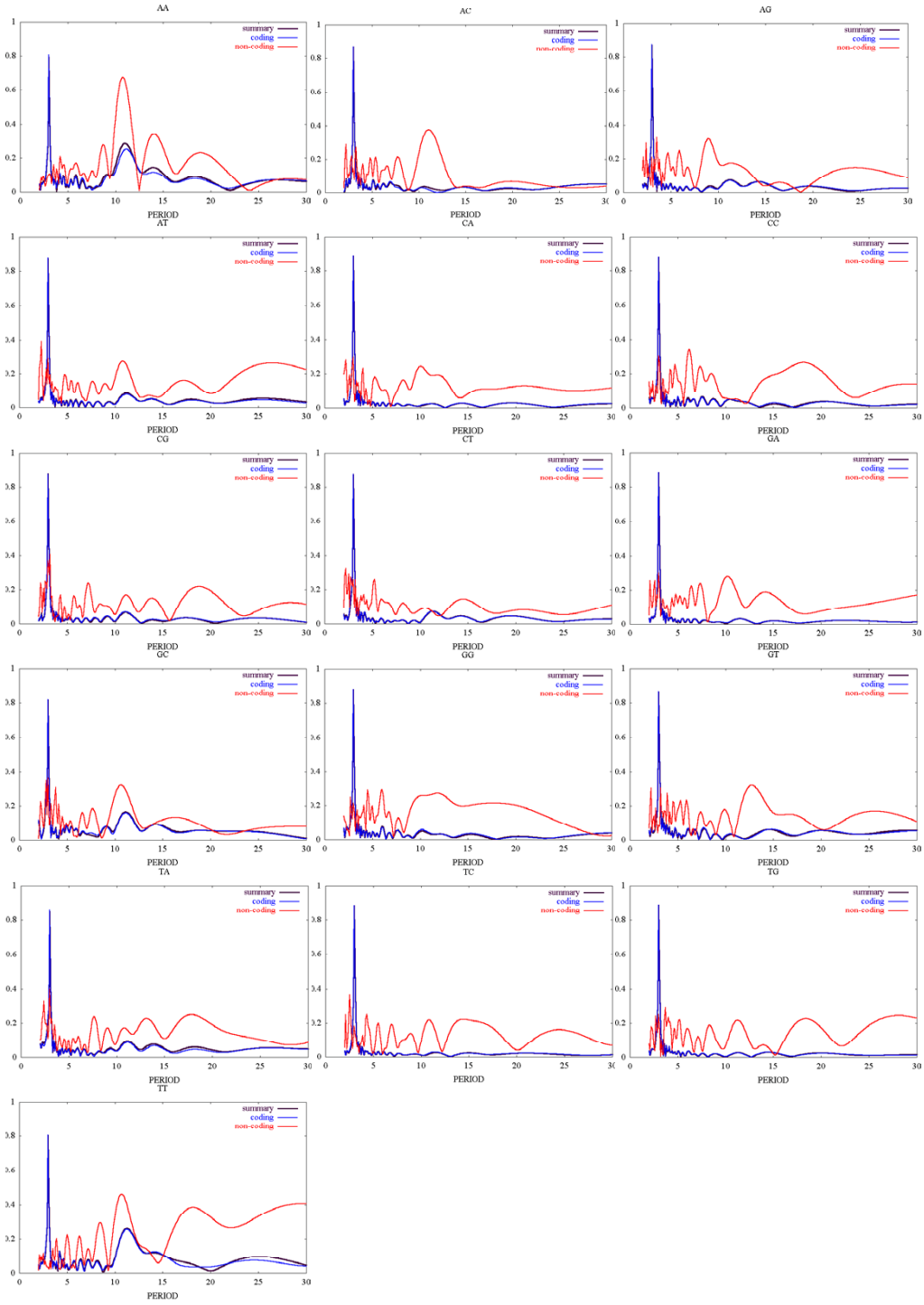

Figure 1
Periodograms of the distance distributions of 16 dinucleotides in *E. coli* genome. The complete nucleotide sequence of *E. coli,* as well as subsets of its coding and noncoding regions, was subjected to the positional autocorrelation analysis for all 16 dinucleotides separately. Resulting autocorrelation profiles were after that analyzed by Fourier transform. The black lines correspond to the whole genome, the blue curves – to the coding sequences, and the red curves – to the noncoding sequences.

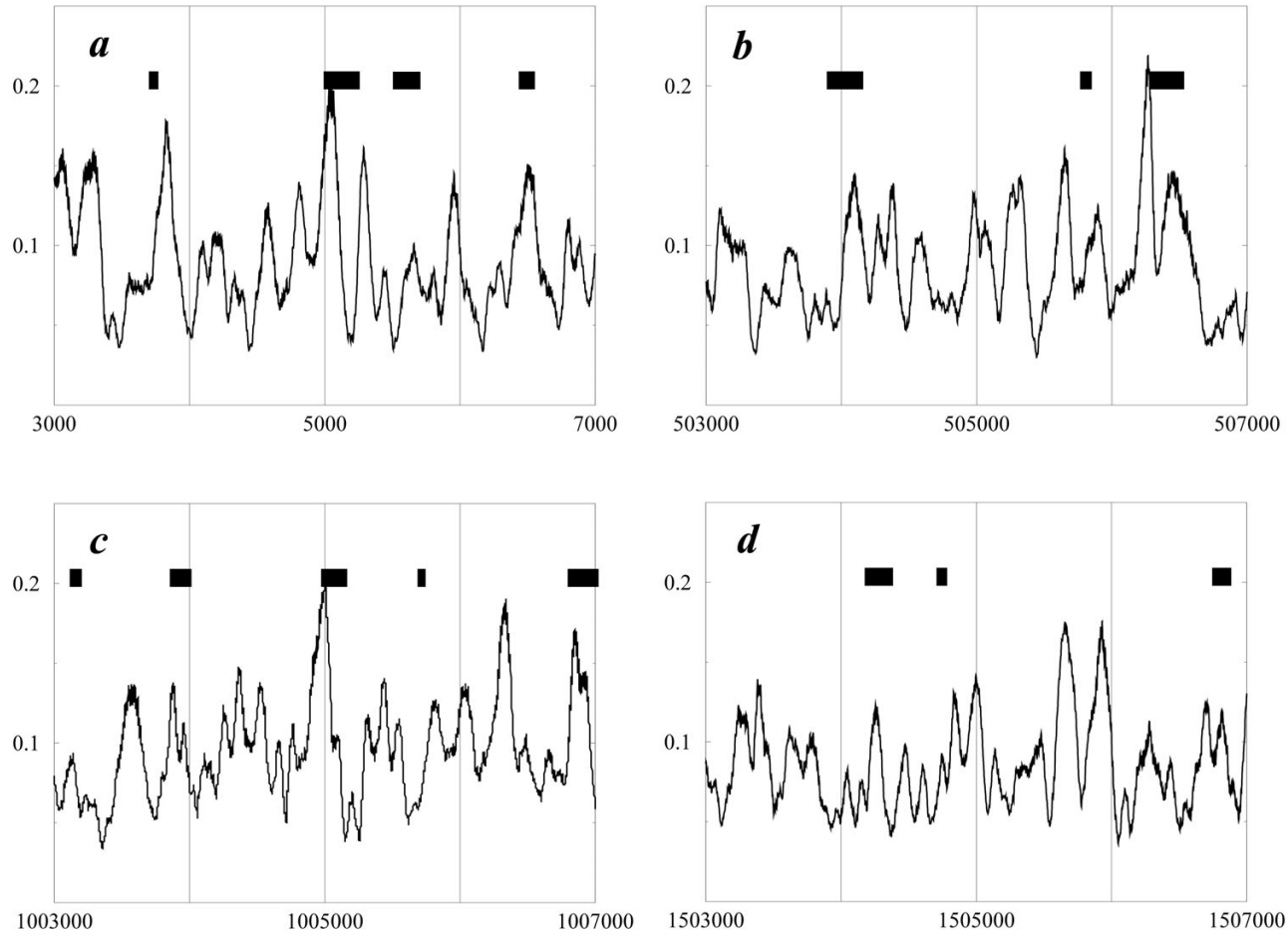

Figure 2
Four examples of periodicity maps for fragments of *E. coli* genome. The maps were smoothed by running average with window 51 bp. The black bars on the top of the plot correspond to positions of intergenic regions.

$$f_p = \frac{\sqrt{\left(\sum_{i=1}^{i=W-2} \sin\left(2\pi\frac{i}{p}\right)\left(X_i - \overline{X}\right)\right)^2 + \left(\sum_{i=1}^{i=W-2} \cos\left(2\pi\frac{i}{p}\right)\left(X_i - \overline{X}\right)\right)^2}}{2\pi \sum_{i=1}^{i=W-2}\left(X_i - \overline{X}\right)^2}$$

where f_p is normalized wave-function amplitude of period p, X is an autocorrelation profile for one chosen dinucleotide, X_i is its value in position i, $\overline{X}$ is its average value, and W is a maximal considered autocorrelation distance (in our case 100 bp).

Sequence periodicity

As a probe of periodicity the sine waves with period T were taken to describe idealized periodical distribution of AA and TT dinucleotides within window W. The probes were correlated with *E. coli* sequences by moving the probes along the sequences and calculating the value C for every position.

$$C = \left|\sum_{i=1}^{W} K_i \sin\left(\frac{2\pi i}{T}\right)\right| \Big/ C_{\max}$$

where i is an index of a dinucleotide position in the window W and

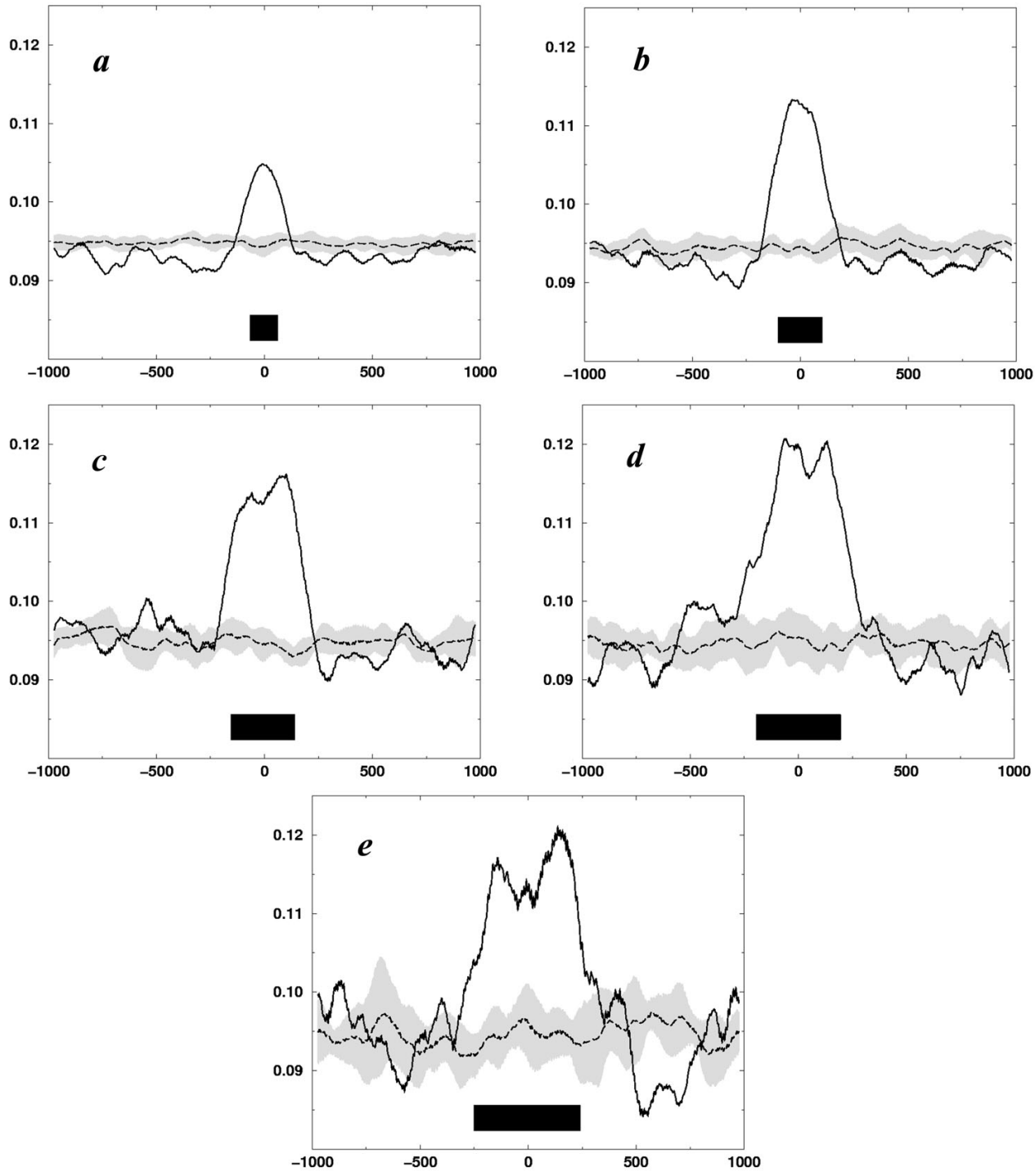

Figure 3
The averaged maps of periodicity are synchronized at the centers of intergenic regions and smoothed by a running average of 51 bp. Five families of the intergenic regions with different lengths are presented: a) 100 ± 50, bp 1073 sequences, b) 200 ± 50 bp, 602 sequences, c) 300 ± 50 bp, 319 sequences, d) 400 ± 50 bp, 160 sequences, and e) 500 ± 50 bp, 78 sequences. The black bars at the bottom of the each figure correspond to the average intergenic region. The gray bands around black dashed lines correspond to standard deviations around randomized background.

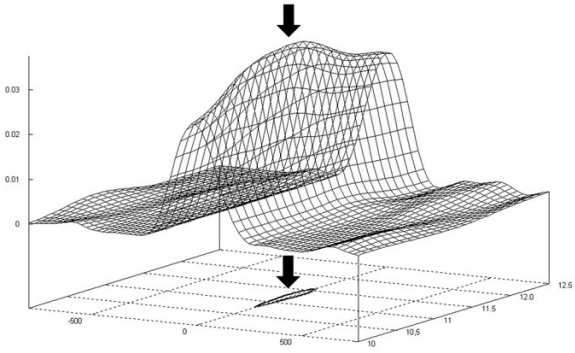

Figure 4
The 3D resonance plot for the intergenic regions of length 200 ± 50 bp. The maximum of resonance the plot corresponds to period 11.3 ± 0.4 bp. The contour around the maximum is also shown as a projection at the base line level.

$$K_i = \begin{cases} 1 & if\ AA\ or\ TT\ dinucleotides \\ 0 & other\ dinucleotides \end{cases}$$

The value $C_{\max}$ is introduced for the normalization purposes. It is calculated as follows:

$$C_{\max} = \sum_{i=1}^{W} L_i \sin\left(\frac{2\pi i}{T}\right)$$

where i is a position in the window W and

$$L_i = \begin{cases} 1 & if\ \sin\left(\frac{2\pi i}{T}\right) \geq 0 \\ 0 & otherwise \end{cases}$$

Ideally periodical sequence segments would be, therefore, described by $C = 1$, while segments with no periodicity would correspond to $C = 0$. The results of these calculations are presented as maps of the sequence periodicity. The four sample maps are shown in Fig. 2a,2b,2c,2d.

Synchronization of the maps

The maps around intergenic regions were combined (summed) separately for the groups of similar sizes of the intergenic regions. Five such groups were analyzed: 100 ± 50 bp, 200 ± 50 bp, 300 ± 50 bp, 400 ± 50 bp, and 500 ± 50 bp. For each group the maps were synchronized at the respective intergenic centers and the sums of the maps were calculated and smoothed by a running average within 51 bp. The standard deviations for the combined plots were estimated by generating random sequences of the same size and dinucleotides composition for each group separately and averaging the respective periodicity maps.

The resonance plot

The resonance 3D plot for the intergenic regions of length 200 ± 50 bp was built from calculations with different periods T in the interval 10–12.5 bp. One-third (202) of the most periodic maps of this group was taken for the calculation. The maps for different periods T were smoothed five times by a running average over 51 bp. The baselines were set to 0. The surface of 3D plot was smoothed 3 times by a running average over 9 point square elements, on the grid with separations 0.1 bp for T, and 20 bp for sequence position.

Competing interests

None declared.

Authors' contributions

SH carried out all graphics. ENT and AB participated in the design of the study and analysis of results. All authors drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We thank V. Kirzhner and all members of the Genome Diversity Center for fruitful discussions and critical comments on the paper. A.B. is grateful to Professors T. Ratiu and J.H. Maddocks from the Bernoulli Institute at the Swiss Federal Institute of Technology for the kind invitation to visit "Centre Bernoulli" for two months. S.H. and A.B. are partially supported by the FIRST Foundation of the Israel Academy of Science and Humanities.

References

1. Herzel H, Weiss O, Trifonov EN: **Sequence periodicity in complete genomes of Archaea suggests positive supercoiling.** *J Biomol Struct Dyn* 1998, **16:**341-345.
2. Herzel H, Weiss O, Trifonov EN: **10–11 bp periodicities in complete genomes reflect protein structure and DNA folding.** *Bioinformatics* 1999, **15:**187-193.
3. Ozoline ON, Deev AA, Trifonov EN: **DNA bendability – a novel feature in E-coli promoter recognition.** *J Biomol Struct Dyn* 1999, **16:**825-831.
4. Tomita M, Wada M, Kawashima Y: **ApA dinucleotide periodicity in prokaryote, eukaryote, and organelle genomes.** *J Mol Evol* 1999, **49:**182-192.
5. Worning P, Jensen LJ, Nelson KE, Brunak S, Ussery DW: **Structural analysis of DNA sequence: evidence for lateral gene transfer in Thermotoga maritima.** *Nucleic Acids Res* 2000, **28:**706-709.
6. Petersen L, Larsen TS, Ussery DW, On SLW, Krogh A: **RpoD promoters in Campylobacter jejuni exhibit a strong periodic signal instead of a-35 box.** *J Mol Biol* 2003, **326:**1361-1372.
7. Strauss F, Gaillard C, Prunell A: **Helical Periodicity of DNA, Poly(Da) . Poly(Dt) and Poly(Da-Dt) . Poly(Da-Dt) in Solution.** *Eur J Biochem* 1981, **118:**215-222.
8. Peck LJ, Wang JC: **Sequence Dependence of the Helical Repeat of DNA in Solution.** *Nature* 1981, **292:**375-378.

9. Tolstorukov MY, Virnik K, Adhya S, Zhurkin VB: **Genome-wide A-tract distribution and DNA packaging in pro-and eukaryotes [abstract].** *J Biomol Struct Dyn* 2003, **20:**869-870.
10. Plaskon RR, Wartell RM: **Sequence Distributions Associated with DNA Curvature Are Found Upstream of Strong Escherichia-Coli Promoters.** *Nucleic Acids Res* 1987, **15:**785-796.
11. Trifonov EN: **Curved DNA.** *Crc Critical Reviews in Biochemistry* 1985, **19:**89-106.
12. Hagerman PJ: **Sequence-directed curvature of DNA.** *Annu Rev Biochem* 1990, **59:**755-781.
13. Harrington RE: **DNA Curving and Bending in Protein DNA Recognition.** *Mol Microbiol* 1992, **6(18):**2549-2555.
14. Perez-Martin J, Rojo F, de Lorenzo V: **Promoters responsive to DNA bending: a common theme in prokaryotic gene expression.** *Microbiol Rev* 1994, **58:**268-290.
15. Gabrielian A, Pongor S: **Correlation of intrinsic DNA curvature with DNA property periodicity.** *FEBS Lett* 1996, **393:**65-68.
16. Lisser S, Margalit H: **Determination of Common Structural Features in Escherichia-Coli Promoters by Computer-Analysis.** *Eur J Biochem* 1994, **223:**823-830.
17. Gabrielian AE, Landsman D, Bolshoy A: **Curved DNA in promoter sequences.** In *Silico Biol* 2000, **1:**183-196.
18. Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW: **A DNA structural atlas for Escherichia coli.** *J Mol Biol* 2000, **299:**907-930.
19. Marini JC, Levene SD, Crothers DM, Englund PT: **Bent Helical Structure in Kinetoplast DNA.** *Proc Natl Acad Sci U S A* 1982, **79(24):**7664-7668.
20. Hagerman PJ: **Evidence for the existence of stable curvature of DNA in solution.** *Proc Natl Acad Sci U S A* 1984, **81(15):**4632-4636.
21. Hagerman PJ: **Sequence dependence of the curvature of DNA – a test of the phasing hypothesis.** *Biochemistry* 1985, **24(25):**7033-7037.
22. Koo HS, Wu HM, Crothers DM: **DNA bending at adenine . thymine tracts.** *Nature* 1986, **320(6062):**501-506.
23. Ulanovsky L, Bodner M, Trifonov EN, Choder M: **Curved DNA – Design, Synthesis, and Circularization.** *Proc Natl Acad Sci U S A* 1986, **83:**862-866.
24. Trifonov EN: **Translation Framing Code and Frame-Monitoring Mechanism as Suggested by the Analysis of Messenger-RNA and 16 S Ribosomal-RNA Nucleotide-Sequences.** *J Mol Biol* 1987, **194:**643-652.
25. Trifonov EN: **3-, 10.5-, 200-and 400-base periodicities in genome sequences.** *Physica A* 1998, **249:**511-516.
26. Bolshoy A, Nevo E: **Ecologic genomics of DNA: upstream bending in prokaryotic promoters.** *Genome Res* 2000, **10:**1185-1193.