ELSEVIER

Biophysical Chemistry 120 (2006) 71 – 79

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

A statistical approach for analyzing structural and regulative information in prokaryotic genomes

Raffaella Paparcone, Stefano Morosetti, Anita Scipioni, Pasquale De Santis *

Dipartimento di Chimica, Università di Roma 'La Sapienza', P.le A. Moro 5, 00185 Rome, Italy

Received 25 August 2005; accepted 11 September 2005
Available online 17 November 2005

Abstract

Although DNA is iconized as a straight double helix, it does not exist in this canonical form in biological systems. Instead, it is characterized by sequence dependent structural and dynamic deviations from the monotonous regularity of the canonical B-DNA. Despite the complexity of the system, we showed that DNA structural and dynamics large-scale properties can be predicted starting from the simple knowledge of nucleotide sequence by adopting a statistical approach. The paper reports the statistical analysis of large pools of different prokaryotic genes in terms of the sequence-dependent curvature and flexibility. Conserved features characterize the regions close to the Start Translation Site, which are related to their function in the regulation system. In addition, regular patterns with three-fold periodicity were found in the coding regions. They were reproduced in terms of the nucleotide frequency expected on the basis of the genetic code and the pertinent occurrence of the aminoacid residues.

Keywords: Genetic code and regulative information; Statistical approach in prokaryotic genomes; Curvature and flexibility in gene sequences; Aminoacid frequency and GC content in prokaryotes

1. Introduction

In the present postgenomic era, DNA sequences files with billions of informational elements are currently accumulating in the data banks. The need of translating the linear information of the base sequences into functional elements is becoming more and more crucial.

In a recent past, DNA chain was considered to be substantially homogeneous in its canonical structure and acting as a simple repository of the genic information. Moreover, its expression and control were supposed to be fully delegated to proteins. On the contrary, DNA is now being recognized more and more extensively as a very complex polymorphic macromolecule, which plays a relevant part in the management of the gene information that it contains.

Therefore, it is now accepted that, as a consequence of the chemical differences of base pairs along the sequence, the whole DNA is characterized by static and dynamic deterministic modulations of the canonical B-structure which influence supercoiling, looping and protein-binding. As a consequence, every function of DNA, including transcription, replication and recombination, is guided by deviations from the monotonous regularity of the straight canonical B-DNA structure. Such structural deviations are an intrinsic property of the sequence and are recognized and amplified by protein binding. This requires a free energy balance that is paid by the protein interactions but minimized by suitable intrinsic DNA structural properties involved in the recognition. This is particularly evident in the case of sequence dependent DNA-histone octamer association, the nucleosome, which provides the packaging and the superstructural organization of the genome as well as gene regulation.

In fact, many DNA binding proteins often induce DNA to bend and twist, in particular, polymerases and topoisomerases around which DNA folds. The integration host factor (IHF), the TATA binding protein (TBT), the catabolite gene activator protein (CAP), the resolvase and CRO repressor operator protein, which participate in transcription, regulation, replica-

* Corresponding author. Tel.: +39 06 49913228; fax: +39 06 4453827.
E-mail address: pasquale.desantis@uniroma1.it (P. De Santis).

0301-4622/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2005.09.020

tion and site-specific recombination, are important examples of induced distortions in the canonical B-DNA structure.

In conclusion, an important part of the DNA information content is not localized on the codonic regions but appears to be related to stereochemical features of large tracts of sequence.

A useful representation of such sequence dependent properties is conveniently given in terms of curvature and flexibility functions, which appear to be involved in mechanisms governing stability and large-scale dynamics of biological systems, as well as in the recognition processes of regulative and structural proteins.

Some years ago, we developed an analytical method to study the effects of the sequence on modeling the three-dimensional superstructure of DNA based on the integration of the theoretically evaluated slight conformational perturbations of the different dinucleotide steps along the sequence. Such a theoretical model is capable of translating the sequence fluctuations in superstructural elements of DNA. In fact, we have developed a statistical mechanics model to derive superstructural properties of DNA from the sequence-dependent curvature and flexibility. This model allows the prediction of the electrophoretic manifestations of the DNA curvature [1–3], the thermodynamic constants of the sequence-dependent circularization reactions [4], the writhing transitions from relaxed to supercoiled circular forms [5], and the nucleosome thermodynamic stability [6,7]. More recently, the model was confirmed by the statistical analysis of AFM images of DNAs [8,9].

In the present paper, we adopt the notion of superstructural homology in the comparative analysis of genomes, a concept which was found very useful in the analysis of proteins where the aminoacid sequences are strongly degenerated with respect to their secondary and tertiary structures, architectures and biological functions.

Adopting such a physical representation of the genome, we analyzed large pools of genes of different prokaryotic genomes by averaging the sequence-dependent curvature and flexibility functions in order to identify the common superstructural properties over their individual or occasional features. This approach, which corresponds to the strategy of enhancing the signals to noise ratio along an informational file, appears to be useful and promising. Since common features are identified and localized along the sequences for a class of genes, it will be possible to detect such features in each sequence by operating with the pertinent correlation function. Such a strategy can be used to identify putative functional superstructures and their associated binding properties to regulative or structural proteins.

Statistical approaches of genome analysis were attempted by other authors [10–15]. These authors generally report average curvature distributions of genomes and compare the real genomes with the equivalent random sequences with the same base pairs or dinucleotide steps content. Comparisons are also reported with "in silico" evolved genomes by point mutations followed by curvature selection [11]. However, only the global average curvature was considered for the selection. The general results are that the average curvature distribution does not change sensitively and appear to be very similar to those characterizing the pertinent randomized sequences. However, the reported results do not concern the fine features of the coding regions as well as local relevant curvatures in real genomes which could be associated to biological functions. Such local curvatures could be buried in the global distribution because they are plausibly rare in the higher organism genomes. On the contrary, they are more frequent in prokaryotes where e.g. promoters, known to be curved, represent a non-negligible part of the whole genome.

Finally, this paper deeply analyzes the fine features of the curvature and flexibility profiles in the coding regions as well as the base occurrence also in relation with the aminoacid frequency of the expressed proteins.

2. Material and methods

Gene sequences (some thousands per genome) were derived from the complete genomes in the Entrez Genome Database (ftp://ncbi.nlm.nih.gov/genomes/). All the sequences were aligned with respect to the Start Translation Site (STS) in almost all cases represented by the start codon ATG.

In order to analyze the statistical distribution of the base sequences and superstructural properties associated along the genes, a 450 bp window, containing 400 bases upstream and 50 downstream the start codon of each gene, was chosen for the analysis. Furthermore, promoter sequences were distinguished according to the distances between the start of a gene coding region and the end of the previous one. If the distance was less than 400 bp, we considered the two coding regions as part of the same operon [16] and the corresponding promoter was not considered for statistics, since it overlapped with the preceding coding region. On the contrary, if the distance was greater than 400 bp, the promoter sequence was included in the sets.

The list of the investigated genomes is reported in Table 1.

2.1. DNA intrinsic curvature and flexibility

Sequence-dependent curvature profile of each DNA gene has been calculated according to the nearest-neighbor model

Table 1
The list of the investigated genomes

Archaea	
Aeoropyrum pernix	Crenarchaeota
Sulfobolus solfataricus	Crenarchaeota
Archaeoglobus fulgidus	Euryarchaeota
Pyrococcus furiosus	Euryarchaeota
Pyrococcus horikoshii	Euryarchaeota
Bacteria	
Aquifex aeolicus	Aquificae
Agrobacterium tumefaciens	Proteobacteria
Brucella melitensis[a]	Proteobacteria
Escherichia coli	Proteobacteria
Borrelia burgdorferi	Spirochetes
Treponema pallidum	Spirochetes

[a] This proteobacteria has two chromosomes. They are analysed separately and distinguished using the suffixes I and II.

we introduced several years ago [1,2]. DNA curvature represents a sequence-dependent property since it corresponds to the deviation between the local axes of two adjacent helix turns. According to this model, local curvature assigned to the base pair n is evaluated in the complex plane summing up all the local deviations over a turn of the double helix about the position n [1,2]. Curvature profiles calculated adopting other models proposed in literature [17–19] do not show significant changes.

Normalized melting temperatures [20] averaged over a helix turn, $f(n)$, have been used to monitor flexibility changes along the sequence for each gene as we previously proposed [6,7]. In this approach the flexibility is referred to a random sequence and represents also the local differential conformational stability of the DNA structure.

Comparisons were also made with the results obtained adopting other flexibility scales available in literature [18,19]; that derived by X-ray crystal structures of double-helix and protein-DNA complexes [19] are reported in this paper.

These DNA intrinsic properties have been averaged over statistically significant sets of prokaryotic gene sequences.

We evaluated the average curvature, $<C(n)>$, the flexibility factor, $<f(n)>$, and base occurrences at each position along the sequences of genes.

3. Results

3.1. Curvature and flexibility calculations

Fig. 1 shows the profiles of the modulus of the average curvature along the sequence for some of the organisms reported in Table 1. These were obtained summing up the curvature of each gene taking into account the phase relative to the STS position.

We found conserved curvature signals for all the considered promoter regions, due to recurrence of the conserved structural features at about −10 and −35 sequence positions, which characterize the prokaryotic genes. These regions are identified by relatively high and phased curvatures.

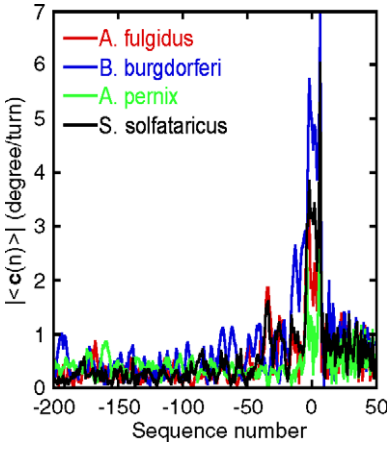

Fig. 1. Profiles of the modulus of the average curvature along the sequence for some of the organisms reported in Table 1. Only some organisms are shown to avoid an excess of superimposition.

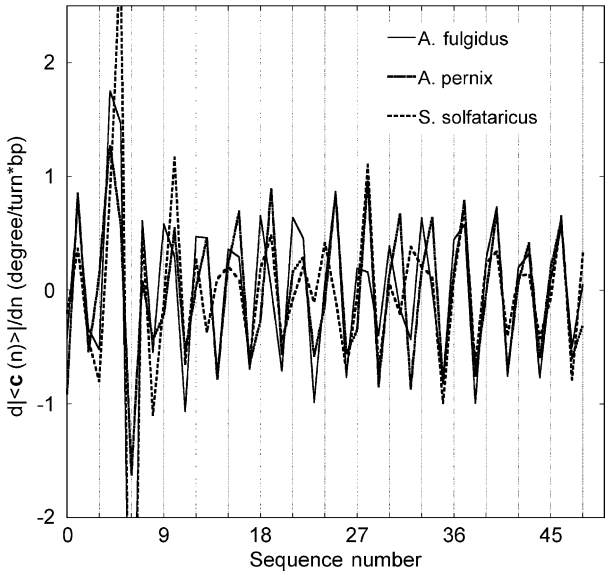

Fig. 2. Profiles of the first derivative of the average curvature along the sequence for the organisms reported in the panel. The three-fold periodicity is particularly evident in the sequence of the minima. Only some organisms of Table 1 are shown to avoid an excess of superimposition.

The sharp signal at STS position is due to the alignment of all the promoter sequences with respect to the ATG triplet. The average curvature increases in the region that starts at about 50 bp upstream the start translation site for all the studied promoter sets, as shown for some representative cases in Fig. 1. As already proposed by other authors [21], DNA curvature in these regions might not only serve for the recognition of regulatory proteins, but it could also facilitate the formation of the open complex during transcription initiation. These results are in agreement with those published by other authors [13–15], who showed that curvature-mediated transcriptional activation is a common feature that is shared even in phylogenetically distant bacteria.

Furthermore, we found a marked three-fold periodicity that characterizes the coding region. To emphasize this periodicity we calculated the derivative of the average curvature function along the sequence for all the organisms reported in Table 1. The curvature trend shows a rather regular three-fold periodicity for all the investigated genomes. Fig. 2 reports the trend for some of the investigated organisms to avoid an excess of superimposition.

Moreover, we calculated the average curvature modulus ($<|\mathbf{c}(n)|>$) (namely, we averaged the curvature independently of the relative phase) and the corresponding standard deviation ($STD(|\mathbf{c}(n)|)$) for the pool of prokaryotic organisms. A linear correlation between these two calculated functions was always obtained. In Fig. 3 the comparison between them is reported in the case of *A. pernix*.

These findings recall the AFM results we recently published [8,9]. From the analysis of thousand AFM images of DNA we showed that the average curvature modulus and the related standard deviation have very similar trend. This result suggests that the average curvature modulus and the corresponding standard deviation contain the same dependence on static and

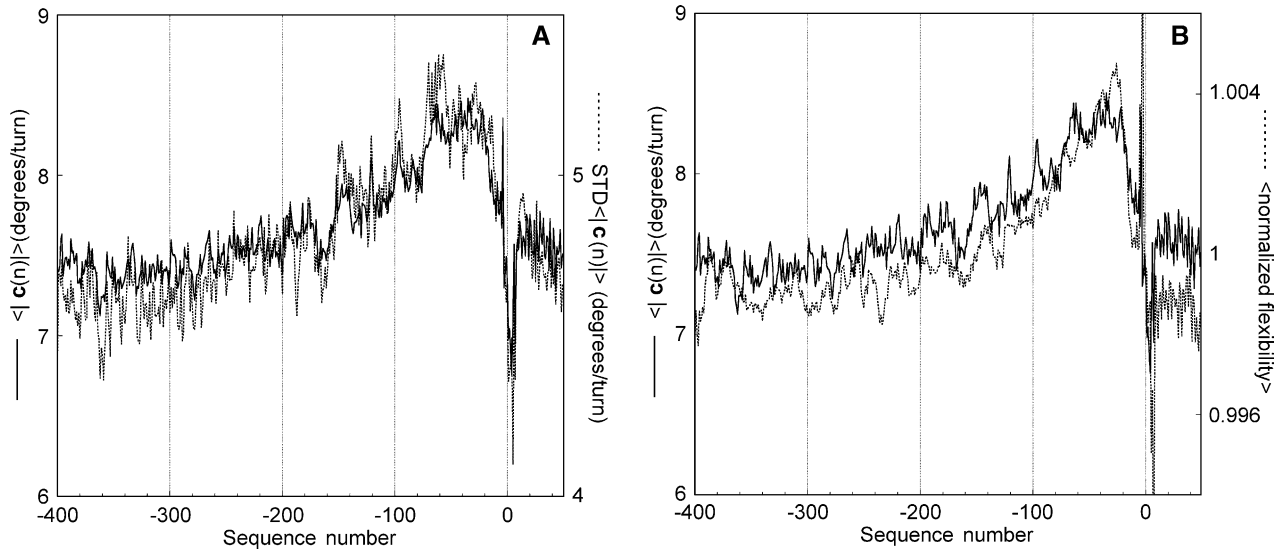

Fig. 3. (A) Comparison between the profiles of the average curvature modulus (<|**c**(*n*)|>) and the corresponding standard deviation (STD(|**c**(*n*)|)). (B) Comparison between the profiles of the average curvature modulus (<|**c**(*n*)|>) and the corresponding flexibility. Both (A) and (B) are referred to the *A. pernix* sequence.

dynamic curvature. They are proportional and their ratio is $(2/(\pi-2))^{1/2}$. A close ratio is also found in the analysis of promoter sequences. The agreement with AFM results suggests a very interesting parallelism: the changes in kinetic energy due to thermal and stochastic perturbations caused by the environment on DNA molecules in the AFM images mirror the changes of potential energy due to the stochastic mutations of base pair at each position. This suggests that the intrinsic curvature of the genes required by the biological functions is a DNA "phenotype" [22] selected from stochastic mutations which produced the actual pool of genes in different genomes. It should be noted that the average curvature modulus increases over long promoter tracts (about 200 bp) whereas the modulus of the average curvature, evaluated taking into account the relative phases, shows changes restricted to about 50 bp near STS position (see Fig. 3B). This means that in more extended regions locally curved but not properly phased tracts are present. This result can be explained on the basis of the transcription complex formation which should be stabilized by the proper phasing of the curved tracts. It is noteworthy that a torsion distributed over a relatively long sequence requires low energy cost. On the contrary, some extent of coherence would be required on regions nearby the STS position. Fig. 3B shows the comparison between the profiles of the flexibility factor and the average curvature modulus in the case of *A. pernix*. They show very similar trends; an analogous correlation is found in all the prokaryotic genomes we analyzed. This is a very impressive result since flexibility is experimentally evaluated measuring melting temperatures of dinucleotide residues [20], while the average curvature modulus is calculated adopting our theoretical model [1,2]. It is possible to hypothesize that curvature and flexibility cooperatively work in order to bind

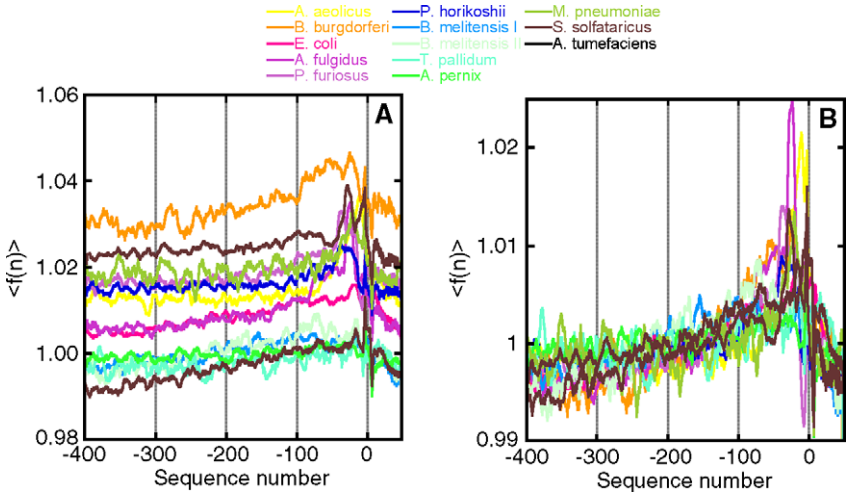

Fig. 4. (A) Profiles of the average flexibility along the sequence for the organisms of Table 1. (B) Profiles of the average flexibility as in A but normalized with respect to the corresponding average values.

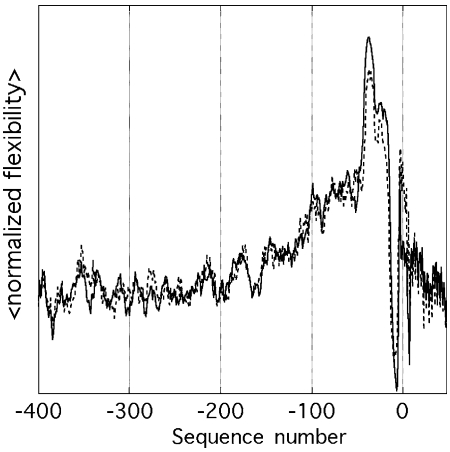

Fig. 5. Profiles of the average normalized flexibility profiles for *P. furiosus* calculated using our (solid line) [6] and Olson scale (dashed line) [19].

RNA-polymerase and the related transcription factors. In fact, while curvature acts favoring the complex formation by reducing the elastic free energy required to distort the sequence, flexibility in the same region allows the adaptation of the sequence to the new conformation of the complex.

Finally, average flexibility profiles were calculated for different prokaryotes, adopting the flexibility scale based on the normalized melting temperature contributions assigned to each dinucleotide residues [6]. Their profiles are reported along the sequence in Fig. 4A. In Fig. 4B they are normalized with respect to the corresponding average values to emphasize the common trend.

For all the investigated organisms, flexibility progressively increases in the promoter region approaching STS and decreases in the coding one. Even though the average flexibility factor relative to each organism is quite different, the corresponding profiles are very similar (Fig. 4B). This is probably due to their specific function in binding the transcription protein complexes. The flexibility could be important to allow these regions to form large transcriptional loop bringing near in the space components of transcriptional complex that would be otherwise more distant in a straight DNA sequence. Then, the behavior of average flexibility along the sequence seems to enhance the function of DNA in the biological processes, which it is involved in.

However, it should be noted that the evaluation of DNA flexibility in terms of the sequence is still controversial. In fact, there are different flexibility scales available in literature. For this reason, we calculated average normalized flexibility profiles also adopting different scales [18,19] and found that the resulting profiles of average normalized flexibility are similar for all the studied organisms (even though the different scales are not fully correlated). Fig. 5 shows the comparison of the average normalized flexibility of promoters from *P. furiosus*, calculated using our [6] and Olson scale [19].

As for the average curvature profiles, also the normalized flexibility profiles of all the investigated genomes show a marked three-fold periodicity, above all in the coding region, that plausibly depends on the genetic code. This result was confirmed calculating the relative Fourier transform relative to the coding regions of all the profiles.

3.2. Periodicity and base occurrence

The presence of regularities in the GC distribution and consequent DNA physical properties were first shown in

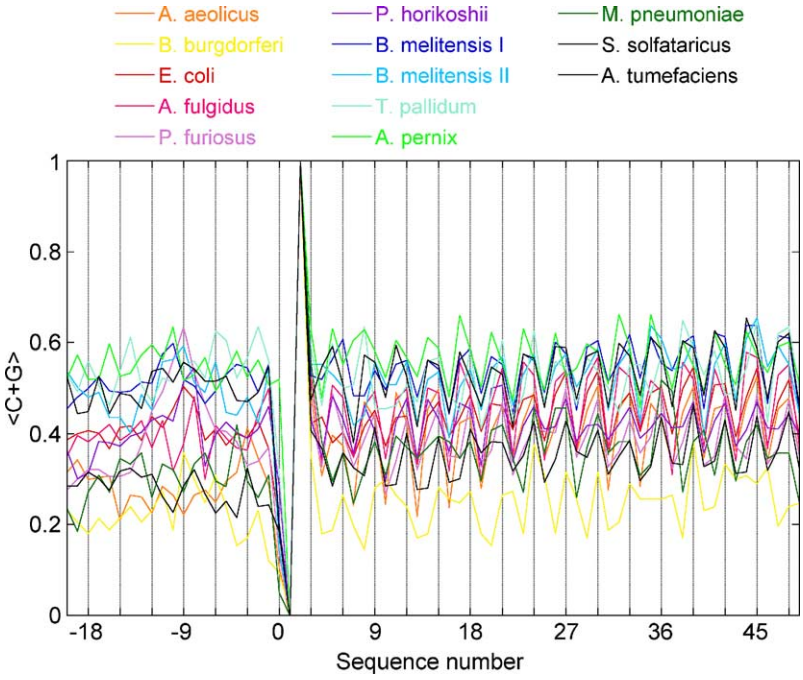

Fig. 6. Profiles of the average content of cytosine and guanine (<C+G>) along the sequence. The line simply joins consecutive points to underline the trend along the sequence.

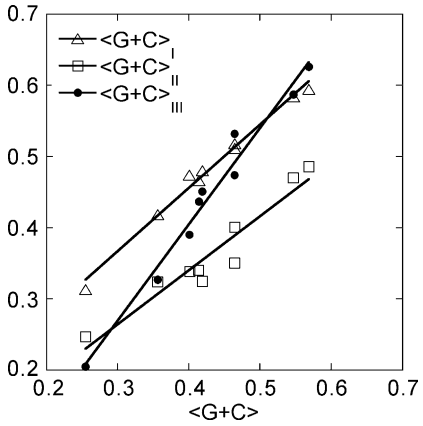

Fig. 7. Experimental correlation between <C+G> content averaged on all the considered genomes and its average values in the first, second and third codonic position.

vertebrates and plants by Bernardi and coworkers and connected to evolutionary hypotheses [23–27].

More recently, statistical analyses of GC content distributions in different vertebrates and invertebrates were reported showing the presence of significant signals at the transcription start and stop sites as well as some regularity in GC content in coding with respect to non-coding regions [12,14,28].

In this paper, we analyze the GC statistical distribution in prokaryotic genomes within a large range of GC content, in which the coding regions represent the major proportion (about 80–90%) of the integral genomes.

As already shown in Fig. 4, the genomes are characterized by flexibility fluctuations along the sequence around different average values. Since the DNA flexibility is a manifestation of the base pair composition, we analyzed base occurrence along the sequence and obtained the corresponding profiles for the four bases. These are characterized in the codon regions by coherent fluctuations with three-fold periodicity. In particular, G and T show the larger amplitudes and regularity.

Fig. 6 illustrates the occurrence of G+C which effectively discriminates the set of genomes analyzed and physically determines the differential thermodynamic stability of the DNAs.

The presence of a three-fold periodicity points out that the probability of finding a cytosine or a guanine is not the same in the three codonic positions. Therefore, we calculated the average occurrence of cytosine and guanine in the three codonic positions along the sequence, ($<C+G>_{I,\ II,\ III}$) and we reported them versus the total average content for each organism (<C+G>). As illustrated in Fig. 7, the increase in G+C content differently affects the corresponding average content in the three codonic positions.

In all the cases we found good linear correlations ($|R|>0.96$) with the stronger variation for the third position. Because changing the third base of the codons does not generally modify the majority of the codified aminoacid, the major variability in the third position represents a way to favor the conservation of aminoacid pool in response to evolution pressures.

For a deeper analysis of periodicity and mutations tolerance of genetic code in terms of DNA sequence, we calculated the average purines occurrence (<A+G>) and reported it along the coding region as illustrated in Fig. 8. Beyond the three-fold periodicity, it is interesting to note that purines are always more abundant in the first codonic position.

In addition, Fig. 9 shows the profile of <A+G> average dispersion along the coding sequence, calculated for all the analyzed organisms. The highest average dispersion is in the third codonic position, while the second one results to be the most conserved.

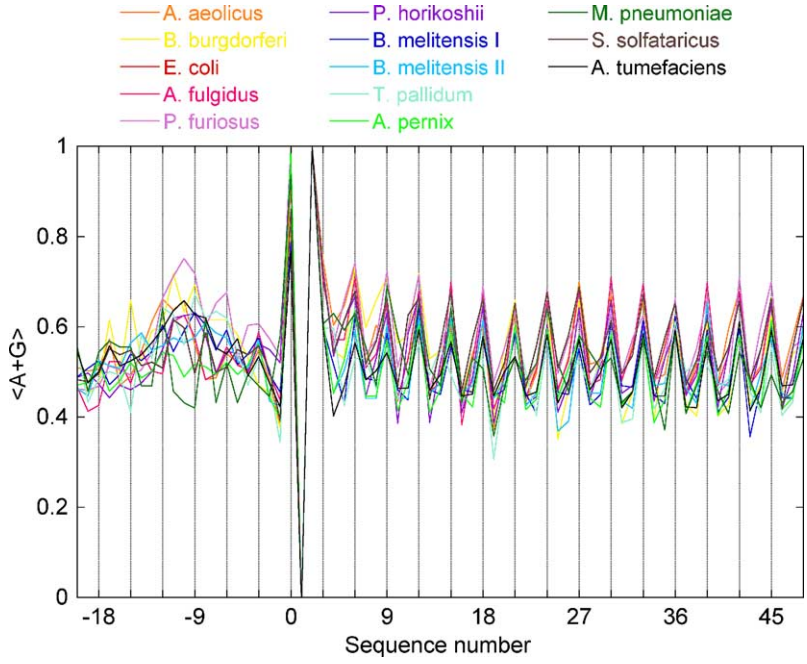

Fig. 8. Profiles of the average content of purines (<A+G>) along the sequence. The lines simply join consecutive points to underline the trend along the sequence.

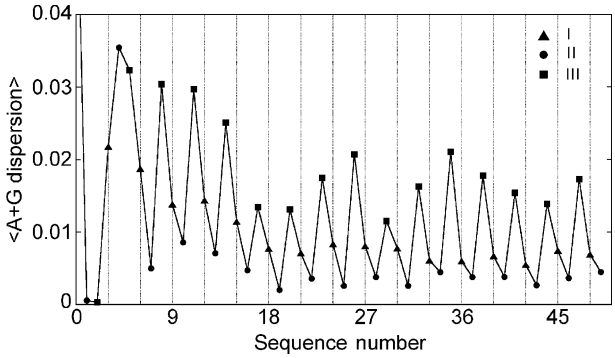

Fig. 9. Profile of the average dispersion of <A+G> calculated over all the analyzed organisms in the coding region. The lines simply join consecutive points to underline the trend along the sequence.

Indeed, according to the genetic code, interchanging purines in the second position often retain the hydrophobic/hydrophilic character of codified aminoacids, while a mutation in the first position generally codifies a different class of aminoacids and could change the nature of the resulting protein. Moreover, the low <A+G> dispersion in the second position involves a low variance in the ratio of hydrophilic and hydrophobic aminoacid residues in the codified proteins.

The regularity of these profiles prompted us to predict the average base distribution in prokaryotes as a function of their average G+C content.

Adopting *E. coli* as representative of prokaryotic organisms, we used its aminoacid average composition data available in literature (http://ww2.ebi.a4.uk/integr8/EBI-Integr8-HomePage.do) and the genetic code to evaluate the expected frequency of codons when degenerated codons are used indifferently as reported in Table 2. Afterwards, the data of Table 2 were normalized to take into account the pertinent preferential codon usage. The comparison between the profiles of expected and experimental <C+G> content is reported in Fig. 10. There is a good agreement in both the cases until C+G average value remains almost constant but the degenerate codon usage seems to reproduce the experimental profile with higher accuracy. However, close to the start translation site, the experimental profile significantly deviates from the periodical expected profiles due to the relative decrease of average G+C content associated to an increase of the curvature and flexibility required for the function of this DNA tract.

Table 2
Degenerate codon usage derived from the experimental aminoacid occurrence in *E. coli* (http://ww2.ebi.a4.uk/integr8/EBI-Integr8-HomePage.do)

	A(III)	C(III)	G(III)	T(III)
AA	3.10	2.30	3.10	2.30
AC	1.42	1.42	1.42	1.42
AG	0.89	1.22	0.89	1.22
AT	1.90	1.90	2.20	1.90
CA	2.40	1.05	2.40	1.05
CC	1.30	1.30	1.30	1.30
CG	0.89	0.89	0.89	0.89
CT	1.60	1.60	1.60	1.60
GA	3.50	2.65	3.50	2.65
GC	1.95	1.95	1.95	1.95
GG	1.83	1.83	1.83	1.83
GT	1.60	1.60	1.60	1.60
TA	0.00	1.55	0.00	1.55
TC	1.22	1.22	1.22	1.22
TG	0.00	0.90	1.30	0.90
TT	1.60	2.00	1.60	2.00

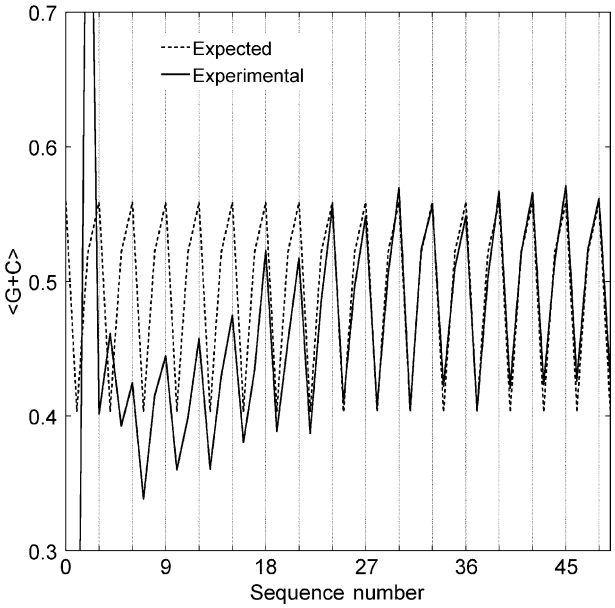

Fig. 10. Comparison between the profiles of <C+G> experimental content and the expected obtained starting from the amino acid composition of *E. coli* (http://ww2.ebi.a4.uk/integr8/EBI-Integr8-HomePage.do). The expected profiles were calculated considering both preferential and degenerate codon usage. The lines simply join consecutive points to underline the trend along the sequence.

A new profile was obtained starting from <C+G> local content along the sequence as obtained by averaging three adjacent positions, and using the fitting equations for the three codon positions as illustrated in Fig. 7. In this way we suppose that the same mechanism acts among different organisms and locally along the sequence, changing preferentially the third codonic position, and therefore generally buffering the expressed aminoacids pool. Even though the strong variation of average C+G frequency (0.2÷0.6), the obtained distributions reproduce the experimental ones very satisfactory. Fig. 11 shows the comparison between <C+G> experimental profiles and those predicted for some organisms largely differentiated for their <G+C> content.

In all the cases the use of the locally averaged profiles allows us to reproduce not only the three-fold periodicity, but also the shape of the profiles. It is worth noting that the <G+C> content in the third codonic positions increases significantly faster than the average <G+C> content specific for each organism.

In this case, differently from the data shown in Fig. 10, the profile of <G+C> content is very satisfactorily reproduced also in the region nearby the STS, where the changes of G+C frequency are needed to modify both DNA flexibility and

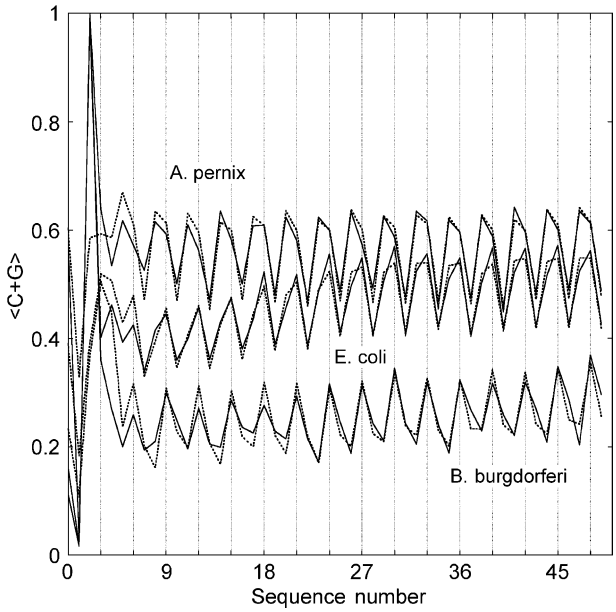

Fig. 11. Comparison between <C+G> experimental profiles (solid line) and those predicted (dashed line) on the basis of the different linear dependence on <G+C> in the three codon positions (see Fig. 7), for some organisms. The lines simply join consecutive points to underline the trend along the sequence.

curvature in order to favor the association and function of the transcription complex.

4. Conclusions

The analysis of a pool of prokaryotic genomes, sorted to cover a rather large interval of average GC content, from 25% up to 60%, shows the presence of conserved features of curvature and flexibility in regulatory regions. Curvature increasing in the regions immediately close to the STS suggests the need of some superstructural features for transcription complexes.

In the same regions an increase of flexibility, compared to the average value, is found. In fact, even though the average flexibility of each organism is quite different, the corresponding normalized profiles are very similar. This is plausibly related to the role of DNA flexibility in binding and functioning of transcription proteins.

It should be noted that the average curvature modulus and flexibility increase over long tracts (about 200 bp) of the promoters, whereas the modulus of the average curvature, evaluated taking into account the relative phases, shows changes restricted to about 50 bp near the STS position. This means that in this region phased curved tracts are required whereas in more extended regions not phased curved tracts, due to their higher flexibility and length, can be properly phased with low energy cost by interactions with transcription proteins.

Both curvature and flexibility show a three-fold periodicity in the coding regions, which mirrors the nature of the genetic code and as a consequence the differential base frequencies along the sequence. The analysis of the profiles of the statistical occurrence of the bases in the three periodically equivalent positions along the sequence of the different organisms indicates a strict correlation with the G+C content. The rate of the increment in terms of the average G+C content in the third codonic position is larger than those of the first and second positions, according to the relative indifference of the coded aminoacids toward the third codon base.

Moreover, the G+A statistical occurrence in the second code position shows the lowest dispersion along the coding sequence due to the relative insensitivity of the hydrophobic or hydrophilic character of the coded aminoacids to the interchange of purines in the genetic code.

Therefore, the gene coding sequences in the different organisms could be considered as the result of mutations which tend to the conservation of the tertiary structure of most proteins whereas the thermal stability of DNA increases with the G+C content. In fact, starting from the average frequencies of the different aminoacids of *E. coli* it is possible to accurately reproduce the corresponding DNA coding region sequence profile adopting the simple equipartition of the experimental frequency of each aminoacid among the degenerate codons. Furthermore, the profiles of the other prokaryotes, characterized by different <G+C> content, are also reproduced taking into account the different linear relations that give the statistical occurrence of bases in terms of the average G+C content.

Acknowledgements

This work was supported by: “Progetto 60% Ateneo 2003–2004” of University “La Sapienza”, “Programma MIUR — Progetti di Ricerca di Interesse Nazionale 2005–2007” and Istituto Pasteur, Fondazione Cenci-Bolognetti.

References

[1] P. De Santis, A. Palleschi, M. Savino, A. Scipioni, A theoretical model of DNA curvature, Biophys. Chem. 32 (1988) 305–317.
[2] P. De Santis, A. Palleschi, M. Savino, A. Scipioni, Validity of the nearest-neighbor approximation in the evaluation of the electrophoretic manifestations of DNA curvature, Biochemistry 29 (1990) 9269–9273.
[3] C. Anselmi, P. De Santis, R. Paparcone, M. Savino, A. Scipioni, From the sequence to the superstructural properties of DNAs, Biophys. Chem. 95 (2002) 23–47.
[4] P. De Santis, M. Fuà, M. Savino, C. Anselmi, G. Bocchinfuso, Sequence-dependent circularization of DNAs: a physical model to predict the DNA sequence dependent propensity to circularization and its changes in the presence of protein-induced bending, J. Phys. Chem. 100 (1996) 9968–9976.
[5] C. Anselmi, G. Bocchinfuso, P. De Santis, M. Fuà, A. Scipioni, M. Savino, Statistical thermodynamic approach for evaluating the writhe transformations in circular DNAs, J. Phys. Chem., B 102 (1998) 5704–5714.
[6] C. Anselmi, G. Bocchinfuso, P. De Santis, M. Savino, A. Scipioni, Dual role of DNA intrinsic curvature and flexibility in determining nucleosome stability, J. Mol. Biol. 286 (1999) 1293–1301.
[7] C. Anselmi, G. Bocchinfuso, P. De Santis, M. Savino, A. Scipioni, A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability, Biophys. J. 79 (2000) 601–613.
[8] G. Zuccheri, A. Scipioni, V. Cavaliere, G. Gargiulo, P. De Santis, B. Samorì, Mapping the intrinsic curvature and flexibility along the DNA chain, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 3074–3079.

[9] A. Scipioni, C. Anselmi, G. Zuccheri, B. Samorì, P. De Santis, Sequence-dependent DNA curvature and flexibility from scanning force microscopy images, Biophys. J. 83 (2002) 2408–2418.
[10] A. Gabrielan, K. Vlahovicek, S. Pongor, Distribution of sequence-dependent curvature in genomic DNA sequences, FEBS Lett. 406 (1997) 69–74.
[11] E. Merino, A. Garciarrubio, The global intrinsic curvature of archaeal and eubacterial genomes is mostly contained in their dinucleotide composition and is probably not an adaptation, Nucleic Acids Res. 28 (2000) 2431–2438.
[12] A.E. Vinogradov, Bendable genes of warm-blooded vertebrates, Mol. Biol. Evol. 18 (2001) 2195–2200.
[13] R. Jàuregui, C. Abreu-Goodger, G. Moreno-Hagelsieb, J. Collado-Vides, E. Merino, Conservation of DNA curvature signals in regulatory regions of prokaryotic genes, Nucleic Acids Res. 31 (2003) 6770–6777.
[14] P.F. Hallin, D.W. Ussery, CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data, Bioinformatics 20 (2004) 3682–3686.
[15] A. Kanhere, M. Bansal, Structural properties of promoters: similarities and differences between prokariotes and eukaryotes, Nucleic Acids Res. 33 (2005) 3165–3175.
[16] M.J.L. De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, S. Miyano, Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information, Pac. Symp. Biocomput. (2004) 276–287.
[17] A. Bolshoy, P. McNamara, R.E. Harrington, E.N. Trifonov, Curved DNA without AA: experimental estimation of all 16 DNA wedge angles, Proc. Natl. Acad. Sci. U. S. A. 88 (1991) 2312–2316.
[18] A.A. Gorin, V.B. Zhurkin, W.K. Olson, B-DNA twisting correlates with base-pair morphology, J. Mol. Biol. 247 (1995) 34–48.
[19] W.K. Olson, A.A. Gorin, X.-J. Lu, L.M. Hock, V.B. Zhurkin, DNA sequence-dependent deformability deduced from protein–DNA crystal complexes, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 11163–11168.
[20] O. Gotoh, Y. Tagashira, Stabilities of nearest-neighbor doublets in double-helical DNA determined by fitting calculated melting profiles to observed profiles, Biopolymers 20 (1981) 1033–1042.
[21] O.N. De Souza, R.L. Ornstein, Inherent DNA curvature and flexibility correlate with TATA box functionality, Biopolymers 46 (2000) 403–415.
[22] C. Anselmi, P. De Santis, R. Paparcone, M. Savino, A. Scipioni, A possible role of DNA superstructures in genome evolution, Orig. Life Evol. Biosph. 34 (2004) 143–149.
[23] G. Bernardi, G. Bernardi, Compositional pattern in the nuclear genome of cold-blooded vertebrates, J. Mol. Evol. 31 (1990) 265–281.
[24] G. D'Onofrio, K. Jabbari, H. Musto, F. Alvares-Valin, S. Cruiveiller, G. Bernardi, Evolutionary genomics of vertebrates and its implications, Ann. N.Y. Acad. Sci. 870 (1999) 81–94.
[25] G. Bernardi, Isochores and the evolutionary genomics of vertebrates, Gene 241 (2000) 3–17.
[26] N. Carels, P. Hatey, K. Jabbary, G. Bernardi, Compositional properties of homologous coding sequences from plants, J. Mol. Evol. 46 (1998) 45–53.
[27] N. Carels, G. Bernardi, Two classes of genes in plants, Genetics 154 (2000) 1819–1825.
[28] L. Zhang, S. Kasif, C.R. Cantor, N. Broude, GC/AT-content spikes as genomic punctuation marks, Proc. Natl. Acad. Sci. U. S. A. 48 (2004) 16855–16860.