ELSEVIER

FEMS Microbiology Letters 218 (2003) 23–30

FEMS MICROBIOLOGY Letters

www.fems-microbiology.org

The DNA secondary structure of the *Bacillus subtilis* genome

Valentina Tosato [a], Kresimir Gjuracic [a], Kristian Vlahovicek [b], Sandor Pongor [b], Antoine Danchin [c], Carlo V. Bruschi [a,*]

[a] *Microbiology Group, International Centre for Genetic Engineering and Biotechnology, AREA Science Park, Padriciano 99, 34012 Trieste, Italy*
[b] *Protein Structure Group, International Centre for Genetic Engineering and Biotechnology, AREA Science Park, Padriciano 99, 34012 Trieste, Italy*
[c] *Hong Kong University Pasteur Research Centre, Dexter HC Man Building 8, Sassoon Road, Pokfulam, Hong Kong*

Received 6 May 2002; received in revised form 30 August 2002; accepted 30 September 2002

First published online 7 November 2002

Abstract

The entire genomic DNA sequence of the Gram-positive bacterium *Bacillus subtilis* reported in the SubtiList database has been subjected in this work to a complete bioinformatic analysis of the potential formation of secondary DNA structures such as hairpins and bending. The most significant of these structures have been mapped with respect to their genomic location and compared to those structures already known to have a physiological role, such as the rho-independent transcription terminators. The distribution of these structures along the bacterial chromosome shows two major features: (i) the concentration of the most curved DNA in the intergenic regions rather than within the ORFs, and (ii) a decreasing gradient of large hairpins from the origin towards the *ter*C end of chromosomal DNA replication. Given the increasing biological relevance of secondary DNA structures, these findings should facilitate further studies on the evolution, dynamics and expression of the genetic information stored in bacterial genomes.

Keywords: DNA hairpins; Bending; Bioinformatic analysis; Secondary structures; *Bacillus subtilis* genome

1. Introduction

Experimental evidence is accumulating for various functional meanings of nucleic acids secondary structures in vivo. Regulatory RNA hairpins are involved in transcription termination [1] and attenuation mechanisms, extensively described for the *Bacillus subtilis pyr* [2], *ilv-leu* [3] and *trp* [4] operons, T box (tRNA synthetases) [5] and S box (sulfur metabolism) [6]. DNA hairpins are formed as molecular intermediates during Tn5 [7] as well as Tn10 [8] transposition. In the process of V(D)J recombination they are formed by the RAG protein complex and subsequently opened by the Artemis/DNA-dependent PK complex [9,10]. Moreover, discrete hairpin structures are recognition signals for viral RNA encapsidation in vivo [11] and for bacteriophage DNA packaging reactions [12]. Another secondary DNA structural feature, intrinsic DNA bending, has also gained importance as a potential site for affinity interaction with proteins involved in the regulation of gene expression at the level of transcription initiation [13] and DNA modification [14].

Currently, the description of these higher-order nucleic acid structures inside prokaryotic and eukaryotic genomes focuses prevalently on RNA terminator hairpins, in particular in those microorganisms characterized by rho-independent transcription termination [15]. Recently, Danchin et al., [16] presented the provocative view of a correlation between the distribution of genes on the bacterial chromosome and the physical architecture of the cell. These and other works point towards the existence, within the genome of various organisms, of a new storage layer of information, responsible for the higher level of cellular organization, and surpassing the one of protein encoding.

B. subtilis is the best-characterized member of the Gram-positive bacteria from the genetic and biochemical point of view. The sequencing of its genome, composed of 4 214 810 nucleotides, was completed in 1997 by an inter-

* Corresponding author. Tel.: +39 (040) 375 7304; Fax: +39 (040) 375 7343.
E-mail address: bruschi@icgeb.trieste.it (C.V. Bruschi).

0378-1097/02/$22.00
PII: S0378-1097(02)01032-7

national consortium composed of 25 European, seven Japanese and one Korean laboratory together with two biotechnology companies [17]. Several computer analyses were performed, trying to characterize different features of its genome. Washio et al. [18] reported a systematic calculation of the free-energy values of mRNA tertiary structures around stop codons over the entire genome of *B. subtilis* with consequent comparison of these results with the genome of other prokaryotes. Moreover, accurate analysis of coding sequences termini [19], codon usage, lateral gene transfer [20] and long repeats distribution [21] are reported in literature. However, no complete general analysis of higher-order secondary DNA structure has been reported to date for the *B. subtilis* genome, in spite of the increasing importance of understanding from the structural–functional point of view the informational meaning of whole genomes.

In this paper we report a complete bioinformatic analysis of the whole genome of *B. subtilis* concerning the propensity of its DNA to exhibit significant molecular curvatures and to form stem-and-loop (hairpin) structures. To perform this study, we analyzed the entire genomic sequence by the Bend-it algorithm to study its ability to form significant curved DNA motifs. Moreover, we computed the different thermodynamic potential of hairpin formation with respect to their location in the chromosome, within or outside coding sequences of ORFs. We further analyzed coding sequences containing at least one significant hairpin to ascertain any correlation between the hairpin DNA sequence and its encoded peptide motifs and to map their distribution on the chromosome with respect to gene location. The results presented provide evidence for the concentration of bent DNA in the intergenic regions of the genome as well as of the distribution of the significant stem-loops according to a gradient following the direction of chromosomal DNA replication. The present work is a first step in a more in-depth analysis of the informational content of genomes as a whole, since appropriate parameters for determining the 3D structure of nucleic acids should be established. The possible functional implications of these findings on *B. subtilis* genome evolution and DNA transactions are discussed.

2. Materials and methods

2.1. Bioinformatics

The entire genomic DNA sequence of *B. subtilis* 168 [16] was downloaded via ftp from the Pasteur Institute SubtiList (http://bioweb.pasteur.fr/GenoList/SubtiList) server in 21 contigs of 220 kb each, having 20 kb overlap, to be analyzed for its DNA bendability. The term bendability refers to DNA's ability to bend in the direction of the major groove. Numeric measures of bendability have been determined from DNase I digestion data [22]. Based on these parameters, and the related consensus bendability scale [23], the bendability/curvature propensity values were calculated with the Bend-it server (http://www2.icgeb.trieste.it/~dna/bend). Individual DNA contigs ranging in size from 199 960 to 200 040 bp were submitted in raw format, and the predicted curvature data, expressed as degrees/helical turn, were collected through e-mail in ASCII text table format.

The inverted repeats search in each contig was performed by using the Wisconsin Package Version 10.0, Genetics Computer Group (GCG), Madison, WI. In particular, the probability of hairpin formation was analyzed by the StemLoop program and the data plotting to visualize all the inverted-repeat structures was obtained by the DotPlot program. The free-energy values were calculated using the RNA tertiary structure prediction program, RNA fold, the algorithm of which is based on that of mFold by Zucker [24], adapted to the single strand DNA molecules [25]. The optimal and sub-optimal secondary structures for the DNA molecules, predicted by mFold, was computed by PlotFold. Protein motifs search was performed by Motifs with patterns defined in the Prosite dictionary. Reformat and Fetch programs were used to re-write and copy all the 220-kb sequence files.

2.2. Analysis parameters

The significant curved DNA motifs (bent DNA) were derived from the raw data according to known experimental parameters. In the context of a prediction, high predicted values (5° to 25°/helical turn) indicate those segments that are most probably curved, while segments that have low predicted curvature values (below 5°/helical turn) are considered straight motifs. The correlation between the numeric value and the extent of curvature is quite strong [26]. The DNA motifs showing more than 14° per helical turn were chosen in order to obtain most probable curved motifs. In a further analysis, curved motifs with even higher curvature values ($>15°$ and $>16°$/helical turn) were also considered.

We ran StemLoop with varying thermodynamic parameters: the minimum number of bonds per stem (stringency) varied from 40 to 78 (with G–T, A–T/U, and G–C scored as 1, 2, and 3 bonds, respectively) and the correspondent minimum stem length from 20 to 50 nt. As a general procedure, we arbitrarily varied stringency parameters in order to yield a hairpin content per contig smaller than eight. The loop size always ranged from a minimum of 3 to a maximum of 50 nt. Since the maximum processing program length is 300 000 nt, we used the 21 220-kb DNA sequences with a 20-kb overlap. The sequence homology search was carried out by the Ncbi Blast e-mail server (http://www.ncbi.nlm.nih.gov/blast/blast.cgi). We mapped the location of the hairpins using the complete *B. subtilis* genome submission in EMBL/GenBank/DDBJ databases (accession numbers from Z99104 to Z99124).

3. Results and discussion

3.1. Curvature

The first step towards a computational analysis of any structural parameter is to study its distribution within the genome. We note that the predicted curvature values are assigned to individual nucleotides by a sliding window along the genome. The resulting values can best be pictured as a long plot along the genome sequence that can be correlated with the known features of the genome. Fig. 1A shows a circular view of the *B. subtilis* genome with curved motifs greater than 14°/helical turn. These data have been used to answer two related questions concerning the percentage of the genome in curved motifs and how many ORFs have curved segments. In response to the first, one can count those residues that are curved motifs and express them as a percentage of the total genome length. The resulting plot will show a distribution of DNA curvature within the *B. subtilis* genome (Fig. 2). The values follow a typical distribution reminiscent of a gamma function, previously observed in practically all bacterial genomes. The most curved segments have curvature values above 14° per helical turn, typically less than 1% of the sequence is curved above this level. On the other hand, the most curved motifs have a tendency to avoid ORFs (Fig. 2, inset). In partial agreement with the findings of Bolshoy and Nevo [13], we found that intrinsic DNA curvature distribution in the genome varies depends upon the cut-off degree value applied to the analysis. With a cut-off of 14° the majority of the curved DNA is located within ORFs, with 15° the distribution is nearly equal for ORFs and intergenic regions (IGR), while with 16° the distribution is reversed, with up to nearly 64% of the curved DNA located within the IGR (Fig. 2, inset). Given the small percentage of the *B. subtilis* genome occupied by IGR with respect to ORFs, this distribution seems even more significant although, of the six most curved DNA regions mentioned in Fig. 1, four are found inside ORFs and two within IGRs. The location of the curvatures found within the coding sequence of the genes did not show any regularity, while the intrinsic high curvature of the chromosomal IGR sequence is in agreement with current views of this feature [27]. The flexibility of the DNA at promoters and terminator sequences has also been observed in eukaryotic DNA of the yeast *Saccharomyces cerevisiae* (Bruschi, personal communication) and has been reported for chromosome I, III and IV of *Leishmania major* at the site of the transcription switching point ([28, 29], http://www.cbs.dtu.dk/services/GenomeAtlas/).

In response to the second question, we counted those ORFs that have at least one nucleotide with a curvature above a certain threshold value. Fig. 3A shows the number of ORFs with at least one curved motif. A total of 256 ORFs overlap with motifs that are curved above 14° per helical turn are listed in Table 1. Interestingly, only 6.2% of all ORFs contain at least one significant curvature (>14°/helical turn). The table also shows that they are equally distributed among all functional categories, with the exception of ORFs encoding proteins similar to unknown proteins and ORFs with no similarity, for which there is a slight under- and over-representation, respectively. From these data we can conclude that a very small proportion of the ORFs contains curved domains, speculating that this could be a result of the selection for rigid DNA sequences or that curvature within ORFs is a spurious event with no particular regulatory meaning. The regions with a curvature greater than 16° are represented in Fig. 4. The distribution of these 47 DNA regions scattered throughout the genome shows randomness for those having a higher degree of curvature. With a 17° bending cut-off there are only six DNA regions found clustered in two areas of the genome, ranging between 888 and 1500 and 3075 and 3260 kb. The possibility of a structural–functional relation between bending and protein-binding sites [26] or chromosome arrangements in bacteria, where folding may preserve the linear order of genes in the DNA [30], should be further investigated.

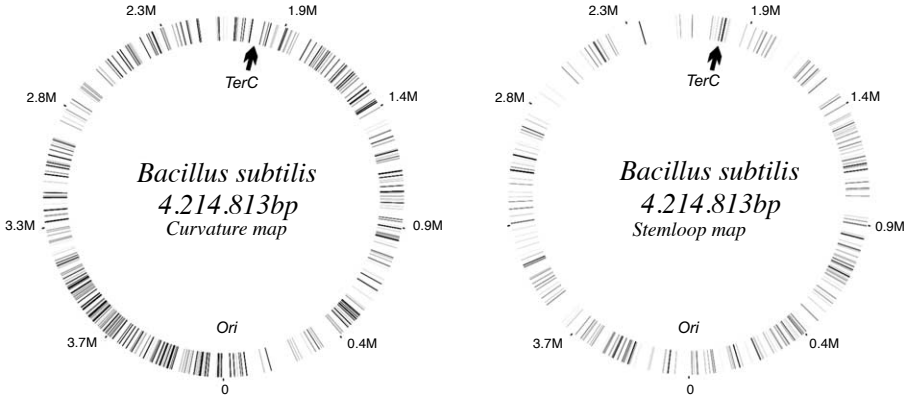

Fig. 1. Graphic representation of the curved motifs (A) and hairpin structures (B) in the *B. subtilis* genome. For the analysis of curvature, regions between 14 and 19 degrees/helical turn were represented by shades of grey. Hairpins represented are included in the range between 70 and 155.

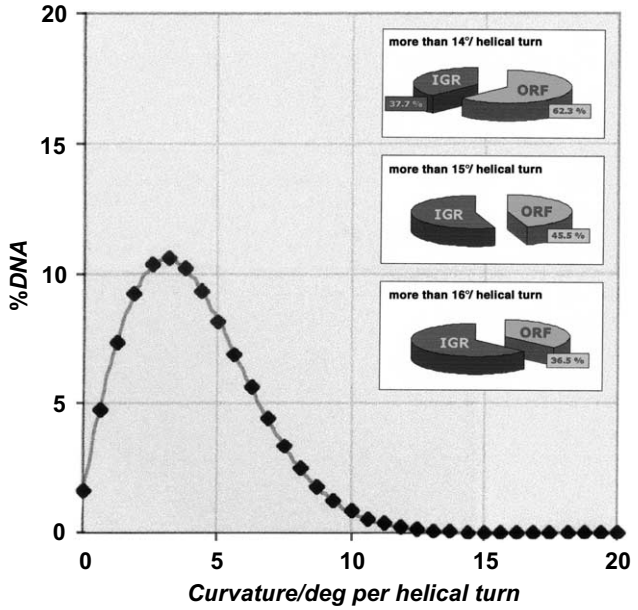

Fig. 2. Distribution of curved motifs in the *B. subtilis* genome. The curvature was calculated in windows of three helical turns (31 nucleotides) and the calculated value was assigned to the central nucleotide of the window.

3.2. Hairpin

The analysis of the hairpins was carried out with the StemLoop program of the GCG package. An obvious technical problem associated with this calculation is that in any DNA sequence there is a very large number of potential small hairpins, so that a lower threshold must be chosen in terms of the number of hydrogen bonds per hairpin. We have chosen a lower threshold of 80 hydrogen bonds and a loop size of between 3 to 50 nucleotides. The StemLoop program was then run by a series of automatically generated overlapping fragments of the genome and the results combined into a unique list. To visualize the results we projected the hairpins back to the genomic sequence and represented them as a circular plot (Fig. 1B). The result illustrates that, even though there is a large number of stem-loops in the genome, both the oriC and the terC, the region containing the terminus of DNA replication of the *Bacillus* chromosome, are in a stem-loop-free region. The *ter*C region is characterized by transcription 'gray-holes' clusters and high A–T content [17], and it contains only a few putative *chi*-like sequences involved in AddAB-mediated recombination [31]. Because of this, it is a region believed to be poorly recombinogenic, with the exception of the *ter*C locus (contig XI), characterized by the long inverted repeats A and B which are typical features of the replication terminus. This distribution could be explained by the evolutionary selection of chromosomal DNA molecules replicating with progressively less energy-consuming DNA elongation kinetics, to ensure completion of duplication. Indeed, long palindromic structures are known to cause genomic instability during replication in yeast [32]. Moreover, the low level of homologous recombination, reported for contigs X to XV (2110 to 2960 kb) [31,33], correlates with the scarcity of long hairpins in this region, in agreement with the fact that the presence of hairpins may favor the initiation of recombination [34]. The distribution of the hairpins overlapping with ORFs is shown in Fig. 3B. For each of these stem-loops a homology search was carried out, using both the Ncbi and the SubtiList server. Using these thermodynamic parameters, we discovered that 79.1% of these large hairpins (68

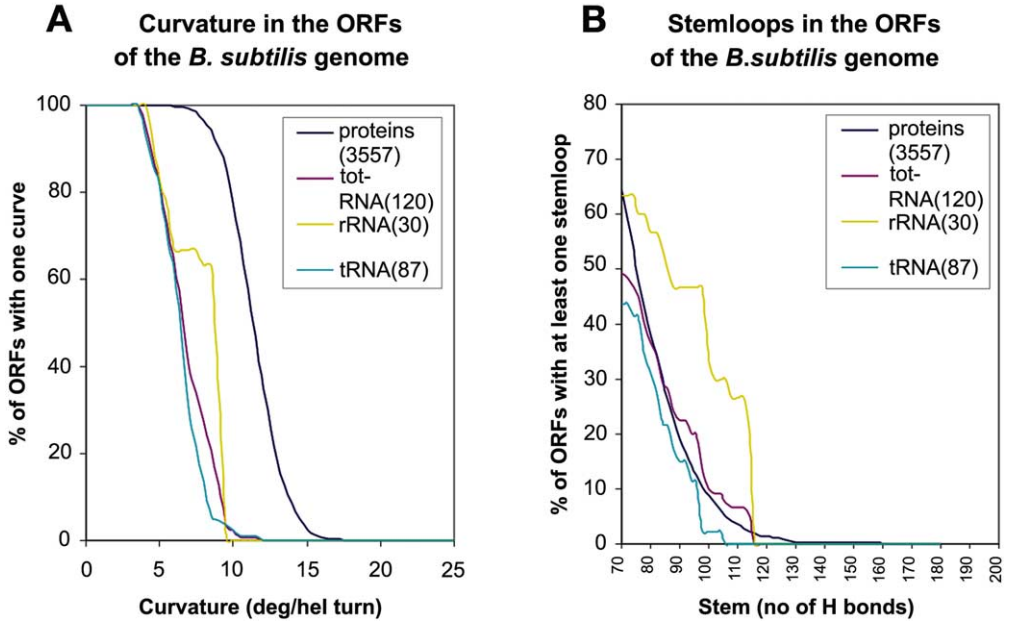

Fig. 3. Percentage of ORFs overlapping with curved motifs (A) and hairpin structures (B) in the*B. subtilis* genome. The total number of protein, RNA, rRNA and tRNA ORFs is given in parentheses.

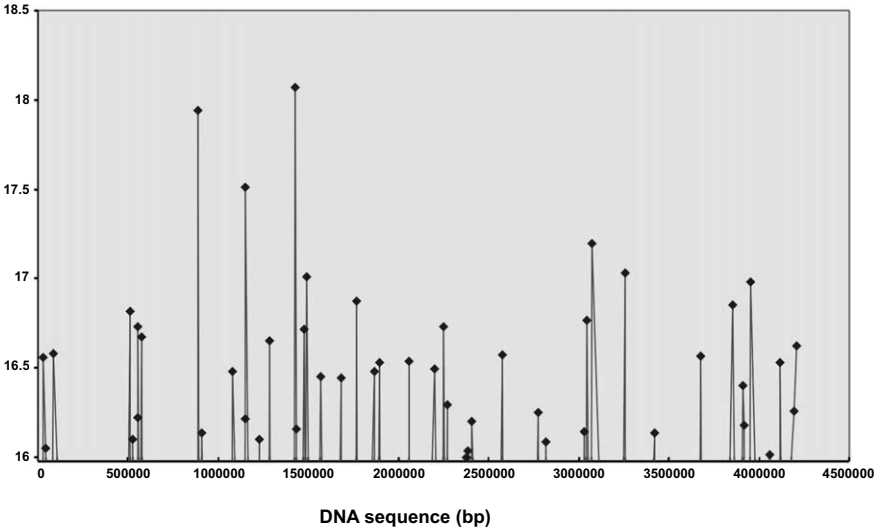

Fig. 4. Distribution of the most curved regions within the *B. subtilis* genome.

out of 86) was located inside a protein-coding DNA sequence. The interesting result was that in 53 cases (61.6%) this hairpin was located at the 3′ end of the ORF, corresponding to the carboxy terminal of the peptide. A second analysis was performed on every 220-kb contig with a minimum stem length of 35 nt and with a minimum number of bonds/stem of 60–62. As before, the loop size varied from 3 to 50 nt. In this analysis, we found 12 new hairpins inside ORF DNA regions and 11 putative terminators, six of which are located within divergons. The last set of parameters used to search for potential DNA secondary structures in the *B. subtilis* genome was a minimum stem length of 20 and 42–43 as minimum number of bonds/stem. The decreasing stem length led to the identification of numerous DNA regions corresponding to putative RNA transcription terminators. By using lower stem length values the number of putative rho-independent terminators found with this methodology increases further. Sixty-six putative terminators were found, 53 of which were already mentioned in SubtiList and 43 of which were located in divergons. Only 10 of the hairpins revealed with these parameters map within an ORF region. Since

Table 1
ORFs containing significantly curved DNA motifs

Functional classification	ORFs containing curved motifs[a]	Number (%) of ORFs with curved motifs[a]		
		>14°/h.t.	>15°/h.t.	>16°/h.t.
1 Cell envelope and cellular processes	**tagF**, tagB, pbpd, ytcC, cwlC, **wprA**, cwlD, area, yfh, yvcS, yunK, yurN, ytgb, yubd, sunT, appD, **exuT**, yhdH, yitg, yfnA, yfkL, **glvC**, yfiC, yfiG, **yfiZ**, ygaD, expZ, ydiF, **gabP**, ydfO, **feuC**, **araN**, yufL, ctaA, fliJ, flhF, prsA, **ftsL**, **rapG**, spsD, cotG, spoIID, spoIVFB, spoIIP, cotC, spoVR, **gerBB**, gerD, comGE	51 (27.1)	11 (22.0)	2 (11.1)
2 Intermediary metabolism	**yoaD**, yoaE, yngE, ykcC, glpK, glpD, yesY, gntK, **gntZ**, yvaM, yngT, acsA, acuB, **acuC**, ytcB, **glgD**, araB, fbpI, ywkA, odhA, yqjN, yqjE, ypwA, hisC, trpB aroF, ctpA, **yncD**, yloW, **ymxG**, dal, ycgM, nasB, yirY, yusQ, accC, yfjR, **yueK**, nadB, **gerCB**	40 (21.3)	8 (16.0)	4 (22.2)
3 Information pathways	dnaI, yprA, yrvE, addB, addA, parC, **topA**, sigL, **ccpB**, ywae, yvhJ, **ywrC**, **ypoP**, exuR, yerO, **lrpC**, greA, yfmL, glyQ, **alaS**, yuiE	20 (10.6)	6 (12.0)	2 (11.1)
4 Other functions	yveQ, yveK, **mmr**, **ytnJ**, katA, ycbR, pksM, **pksP**, pksK, srfAA, srfAB, yomS, yolB, **ymaC**, ydcQ, ydcR, yrvO, **eag**	18 (9.6)	5 (10.0)	3 (16.7)
5 Similar to unknown proteins	yweA, yugS, **ykuG**, yeaD, **yetI**, **yddS**, ywqL, yvgJ, yvdK, ytbJ, **ytfP**, ygcK, yrkO, yqgE, ypeB, yojI, ymcB, **ykrU**, ylxX, yhaN, ydaO, ydfC, **yaaK**, **yabC**, **yabN**	25 (13.3)	8 (16.0)	5 (27.8)
6 No similarity	yxiD, **yydB**, ywcD, yxjh, yvfA, ytlQ, yuaF, **yrzD**, **ytcG**, ygaF, yrkE, **ypuE**, yghG, **yopM**, ypmB, yndK, ykoS, ykoA, yktD, **ylbE**, yjzA, yjbK, **yhdL**, yhaW, **yfmH**, **yfjB**, **yfhO**, ygaE, **yerB**, ydaL, ydbT, yddI, ycdC, **yabT**	34 (18.1)	12 (24.0)	2 (11.1)
Total		188	50	18

[a]ORFs found to contain curved DNA motifs above 14° per helical turn; boldface denotes curvatures above 15°; boldface and underlined denotes curvatures above 16°.

Table 2
ORFs overlapping with at least one hairpin structure

Functional classification	ORFs containing at least one hairpin structure
1 Cell envelope and cellular process	ndhF, ycgO, ycgT, ybgH, nasA, gerKA, yflS, yhcL, yfkF, citS, yheH, ctaE, ylmH, yhvW, resE, bmr, comER, yqeE, spoVID, ytlD, yunJ, atpA, secA, bglP
2 Intermediary metabolism	ydaP, ssuD, xylA, dapG, proJ, ggt, ctpA, mmgB, yrhA, adhB, gapB, pelB, pckA, dhbC, yutB, yunH, yvgR, acdA, vpr, sacX
3 Information pathways	ydbG, yjcD, yvfI
4 Other functions	srfAB, srfAD, pksF, ppsA, ppsC, ppsD, pksR, pksM, yukL
5 Similar to unknown proteins	yazC, yaaE, ycbH, ycgA, yvfW, ydaO, ycsI, yeeK, yesW, yetI, yjcF, yknZ, yloO, yoaN, yqeM, ysnB, ytiP, yveS, yumA, yxjB, ywbG
6 No similarity	YbcS, ykoS, ykuW, ykoS, ymfB, yomI, ypbR, ypbB, yqhO, yqaL, yrkM, yvdQ, yxxB

the free-energy analysis reveals that this type of short hairpin is very stable, their presence inside ORF can represent an impairment on DNA transcription and replication; therefore, it is expected that their number be infrequent.

The total number of DNA palindromic structures and their characteristics are reported in Table 1. We distinguished between structures prone to make hairpins which code for putative terminators (usually found 3–30 nt after the stop signal between two genes in the same orientation) and hairpins, generally longer than the previous ones, which are inside ORFs. Only three hairpins out of 92 are located in chromosomal IGRs, and one hairpin, near the *ndhF* gene, is located within a promoter sequence. None of these four hairpins exist between 'head on' genes (↔), according to the description of IGRs given by Washio et al. [18]. Surprisingly, no stem-loop structure discovered in this study seems to be linked to the autogenous transcriptional attenuation mechanism of *B. subtilis*. Indeed, the StemLoop program did not detect palindromic structures nearby transcriptionally attenuated genes (*ilvleu*, *pyr*, *trp* operons) already described in literature [35]. The inverted repeats characterizing 26 S-box related-ORFs were not detected, while only two palindromes were found next to the 41 tRNA ORFs described in SubtiList (for *lysS* and *hisS*) [5,6]. This apparent failure of the computer program to detect secondary structures acting as terminators or antiterminators in the *ilv-leu*, *pyr*, *trp* operons is mainly due to its capacity to detect only DNA hairpins having no gaps. Although StemLoop is capable of computing mismatches present in DNA bubbles, it cannot deal with the lack of one or more nucleotides in the stem structure. Therefore, since the secondary structures of *B. subtilis* genome that function as transcriptional attenuators are gap-containing hairpins, they escape detection.

While we were able to recognize potential rho-independent terminators among the majority of the 'intergenic' hairpins, we were surprised by the presence of palindromic structures within some ORFs, and we analyzed the functional meaning of their encoded peptides. By using different StemLoop stringency parameters, we found a total of 90 proteins that contained DNA secondary structures in their ORF. Therefore, we analyzed the sequences of these proteins to verify if there was any correlation between the hairpin sequence present in their coding DNA and amino acid motifs (ATP–GTP-binding or phosphorylation sites, lipoprotein lipid attachment sits, protein kinase signature, transmembrane domains, etc.). We found that the DNA sequences that show potential to produce secondary structures were clearly different from the DNA sequences encoding the various peptide motifs considered, thereby indicating no correlation between them. Moreover, we categorized the stem-loop-rich proteins by their functional meaning, following the official SubtiList protein classification. The results are shown in Table 2. Among the 90 proteins harboring DNA hairpins in their genes, 24 (26.7%) belong to the cellular process, 20 (22.2%) to the intermediary metabolism, 3 to the information pathways, 9 (10%) to the miscellaneous class (most are correlated to antibiotic production) and 34 (37.8%) proteins are unknown. Thus, it seems that the genes containing DNA stem-loops encode proteins with no apparent functional correlation. In fact, the percent distribution of these proteins reflects the distribution of the total proteins in the different six classes described in SubtiList, with the only exception the information pathway peptides (i.e. helicases, transcriptional regulators and initiation factors) which have fewer hairpins in their genes (Table 2). In this class, named in SubtiList number 3, there are many proteins involved in protein synthesis, such as the ribosomal proteins (3.7.1) or the tRNA synthetases (3.7.2). Given the positive role of hairpins in DNA recombination and their negative role in DNA replication, this could indicate that functional evolution of these genes did not occur through hairpin-mediated DNA rearrangement. The genes involved in protein synthesis are usually highly expressed during the exponential growth phase and show a highly biased codon usage [36]. This may be due, in part, to the fast-growing nature of *B. subtilis* compared to other prokaryotes. To investigate if the low hairpin content of genes related to peptides involved in protein synthesis processes might be consistent with their unusual codon usage and high expression, we checked the 90 genes and compared them to the list of highly expressed genes in *B. subtilis* defined by their codon bias value [36]. None of our 90 ORFs harboring putative hairpins are included in this list. Just two of these genes (*prsA* and *sacB*) have a strong hairpin downstream of their stop codon. Therefore, it would appear that the presence of palindromic sequences

inside the ORFs of a protein negatively affects its level of expression. Furthermore, among the 90 genes encoding these proteins, we observed that in 35 cases (38.9%) the hairpin was located near the 5′ end. These genes code for proteins that belong to different functional categories. In 55 cases (61.1%) the hairpins are localized in the second half of the gene, towards the carboxy terminus of the encoded protein, and in 45 cases (50%) they were clearly located within the last third portion of it. Five genes (*srfAB* in contig II, *ppsC* in contig X, *yunH* in contig XVII, *atpA* in contig XIX and *sacX* in contig XX) contain two long hairpins, while an identical hairpin is present in the ORF of three different proteins (PpsC, PpsD, PpsA), at the same position with respect to their peptide domains [37]. This could indicate that there is a correlation between the DNA secondary structure and the peptide function. Contrary to that view, the hairpin structures do not correspond to the two known peptide motifs present in all three proteins (neither the phosphopantheteine site nor the AMP-binding sites). On the other hand, the chance of finding long hairpins in Pps genes could not be that peculiar since these genes are among the longest genes present in *B. subtilis*. It seems instead that there is no correlation between the number of hairpins and the length of the genes since *yunH*, *atpA* and *sacX* have the average length of bacterial genes (1.3–1.5 kb), while *srfAB* and *ppsC* are unusually several kb long. Within the *pps* genes, the two hairpins could have some correlation with the amino acid-binding site since they surround the GLX domain in *ppsA*, *ppsC* and *ppsD*. By DNA-adapted mFold analysis of the potential hairpins stability, we calculated that for structures characterized by at least 42 bonds per stem, the average free-energy value for optimal structure is about -16.1 kcal mol^{-1}. Hairpins with at least 75 bonds per stem show very different energy values each from the others, but several of them are around -20 kcal mol^{-1}, predicting, at least in vitro, a more stable DNA secondary structure. However, it is known that hairpins containing very large loops are not very stable [30, 38], and therefore hairpins with large loops were discarded.

4. Conclusions

In comparison to extensive DNA sequence analysis studies, relatively little is known on the information encoded in the DNA secondary structure. There is ample evidence that the activity of DNA-binding proteins is dependent on the 3D structure of DNA, but general rules are not available. As structural elements such as curves, hairpins, etc. can be predicted with a reasonable accuracy, one can expect that a systematic analysis of DNA secondary structures can reveal substantial novel information. The analysis of the *B. subtilis* genome is a step in this direction.

References

[1] Yarnell, W.S. and Roberts, J.W. (1999) Mechanism of intrinsic transcription termination and antitermination. Science 284, 611–615.

[2] Lu, Y., Turner, R.J. and Switzer, R.L. (1996) Function of RNA secondary structures in transcriptional attenuation of the *Bacillus subtilis pyr* operon. Proc. Natl. Acad. Sci. USA 93, 14462–14467.

[3] Grandoni, J.A., Fulmer, S.B., Brizzio, V., Zahler, S.A. and Calvo, J.M. (1993) Regions of the *Bacillus subtilis ilv-leu* operon involved in regulation by leucine. J. Bacteriol. 175, 7581–7593.

[4] Babitzke, P. and Yanofsky, C. (1993) Reconstitution of *Bacillus subtilis* trp attenuation in vitro with TRAP, the trp RNA-binding attenuation protein. Proc. Natl. Acad. Sci. USA 90, 133–137.

[5] Grundy, F.J., Moir, T.R., Haldeman, M.T. and Henkin, T.M. (2002) Sequence requirements for terminators and antiterminators in the T box transcription antitermination system: disparity between conservation and functional requirements. Nucleic Acids Res. 30, 1646–1655.

[6] Grundy, F.J. and Henkin, T.M. (1998) The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in Gram-positive bacteria. Mol. Microbiol. 30, 737–749.

[7] Bhasin, A., Goryshin, I.Y. and Reznikoff, W.S. (1999) Hairpin formation in Tn5 transposition. J. Biol. Chem. 274, 37021–37029.

[8] Kennedy, A.K., Guhathakurta, A., Kleckner, N. and Haniford, D.B. (1998) Tn10 transposition via a DNA hairpin intermediate. Cell 95, 125–134.

[9] Shockett, P.E. and Schatz, D.G. (1999) DNA hairpin opening mediated by the RAG1 and RAG2 proteins Mol. Cell Biol. 19, 4159–4166.

[10] Ma, Y., Pannicke, U., Schwarz, U.U. and Lieber, M.R. (2002) Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. Cell 108, 781–794.

[11] McBride, M.S. and Panganiban, A.T.J. (1996) The human immunodeficiency virus type 1 encapsidation site is a multipartite RNA element composed of functional hairpin structures. Virology 70, 2963–2973.

[12] Quiao, X., Quiao, J. and Mindich, L.J. (1995) Interference with bacteriophage phi 6 genomic RNA packaging by hairpin structures. Virology 69, 5502–5505.

[13] Bolshoy, A. and Nevo, E. (2000) Ecological genomics of DNA: upstream bending in prokaryotic promoters. Genome Res. 10, 1185–1193.

[14] Horton, N.C. and Perona, J.J. (1998) Recognition of flanking DNA sequences by *Eco*RV endonuclease involves alternative patterns of water-mediated contacts. J. Biol. Chem. 21, 21721–21729.

[15] Carafa, Y., Brody, E. and Thermes, C.J. (1990) Prediction of rho-independent *Escherichia coli* transcription terminators: a statistical analysis of their RNA stem-loop structures. Mol. Biology 216, 835–858.

[16] Danchin, A., Guerdoux-Jamet, P., Moszer, I. and Nitschke, P. (2000) Mapping the bacterial cell architecture into the chromosome. Philos. Trans. R. Soc. London Ser. B 355, 179–190.

[17] Kunst, F. et al. (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 390, 249–256.

[18] Washio, T., Sasayama, J. and Tomita, M. (1998) Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. Nucleic Acids Res. 26, 5456–5463.

[19] Rocha, E.P., Danchin, A. and Viari, A. (1999) Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis. Nucleic Acids Res. 27, 3567–3576.

[20] Moszer, I., Rocha, E.P. and Danchin, A. (1999) Codon usage and lateral gene transfer in *Bacillus subtilis*. Curr. Opin. Microbiol. 2, 524–528.

[21] Rocha, E.P., Danchin, A. and Viari, A. (1999) Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes Mol. Biol. Evol. 16, 1219–1230.
[22] Brukner, I., Sanchez, R., Suck, D. and Pongor, S. (1996) Sequence-dependent bending propensity of DNA as revealed by Dnase I: parameters for trinucleotides. EMBO J. 14, 1812–1818.
[23] Gabrielian, A. and Pongor, S. (1996) Correlation of intrinsic DNA curvature with DNA property periodicity. FEBS Lett. 393, 65–68.
[24] Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule. Science 244, 48–52.
[25] SantaLucia Jr., J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc. Natl. Acad. Sci. USA 95, 1460–1465.
[26] Munteanu, M.G., Vlahovicek, K., Parthasaraty, S., Simon, I. and Pongor, S. (1998) Rod models of DNA: sequence-dependent anisotropic elastic modeling of local bending phenomena. Trends Biochem. Sci. 23, 341–347.
[27] Perez-Martin, J. and de Lorenzo, V. (1997) Clues and consequences of DNA bending in transcription. Annu. Rev. Microbiol. 51, 593–628.
[28] McDonagh, P.D., Myler, P.J. and Stuart, K. (2000) The unusual gene organization of *Leishmania major* chromosome 1 may reflect novel transcription processes. Nucleic Acids Res. 28, 2800–2803.
[29] Tosato, V., Ciarloni, L., Ivens, A.C., Rajandream, M.A., Barrell, B.G. and Bruschi, C.V. (2001) Secondary DNA structure analysis of the coding strand switch regions of five *Leishmania major* Friedlin chromosomes. Curr. Genet. 40, 186–194.
[30] Teleman, A.A., Graumann, P.L., Lin, D.C., Grossman, A.D. and Losick, R. (1998) Chromosome arrangement within a bacterium. Curr Biol. 8, 1102–1109.
[31] Chedin, F., Noirot, P., Biaudet, V. and Ehrlich, S.D. (1998) A five-nucleotide sequence protects DNA from exonucleolytic degradation by AddAB, the RecBCD analogue of *Bacillus subtilis*. Mol. Microbiol. 29, 1369–1377.
[32] Gordenin, D.A. and Resnick, M.A. (1998) Yeast ARMs (DNA at-risk motifs) can reveal sources of genome instability. Mutat. Res. 400, 45–58.
[33] ElKaroui, M., Biaudet, V., Schbath, S. and Gruss, A. (1999) Characteristics of Chi distribution on different bacterial genomes. Res. Microbiol. 150, 579–587.
[34] Lobachev, K.S., Shor, B.M., Tran, H.T., Taylor, W., Keen, J.D., Resnick, M.A. and Gordenin, D.A. (1998) Factors affecting inverted repeat stimulation of recombination and deletion in *Saccharomyces cerevisiae*. Genetics 148, 1507–1524.
[35] Switzer, R.L., Turner, R.J. and Lu, Y. (1999) Regulation of the *Bacillus subtilis* pyrimidine biosynthetic operon by transcriptional attenuation: control of gene expression by an mRNA-binding protein. Nucleic Acids Res. 26, 824–830.
[36] Karlin, S., Campbell, A.M. and Mrázek, J. (1998) Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. 32, 185–225.
[37] Tosato, V., Albertini, A.M., Zotti, M., Sonda, S. and Bruschi, C.V. (1997) Sequence completion, identification and definition of the fengycin operon in *Bacillus subtilis* 168. Microbiology 143, 3443–3450.
[38] Nag, D.K. and Kurst, A. (1997) A 140-bp-long palindromic sequence induces double-strand breaks during meiosis in the yeast *Saccharomyces cerevisiae*. Genetics 146, 835–847.