BMC Bioinformatics

BioMed Central

Research article

Open Access

Genomes are covered with ubiquitous 11 bp periodic patterns, the "class A flexible patterns"

Etienne Larsabal and Antoine Danchin*

Address: Unité de Génétique des Génomes Bactériens, Institut Pasteur, URA CNRS 2171, 28, rue du Docteur Roux, 75724 Paris Cedex 15, France

Email: Etienne Larsabal - etienne.larsabal@normalesup.org; Antoine Danchin* - adanchin@pasteur.fr

* Corresponding author

Published: 24 August 2005

BMC Bioinformatics 2005, **6**:206 doi:10.1186/1471-2105-6-206

Received: 22 June 2005

Accepted: 24 August 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/206

Abstract

Background: The genomes of prokaryotes and lower eukaryotes display a very strong 11 bp periodic bias in the distribution of their nucleotides. This bias is present throughout a given genome, both in coding and non-coding sequences. Until now this bias remained of unknown origin.

Results: Using a technique for analysis of auto-correlations based on linear projection, we identified the sequences responsible for the bias. Prokaryotic and lower eukaryotic genomes are covered with ubiquitous patterns that we termed "class A flexible patterns". Each pattern is composed of up to ten conserved nucleotides or dinucleotides distributed into a discontinuous motif. Each occurrence spans a region up to 50 bp in length. They belong to what we named the "flexible pattern" type, in that there is some limited fluctuation in the distances between the nucleotides composing each occurrence of a given pattern. When taken together, these patterns cover up to half of the genome in the majority of prokaryotes. They generate the previously recognized 11 bp periodic bias.

Conclusion: Judging from the structure of the patterns, we suggest that they may define a dense network of protein interaction sites in chromosomes.

Background

The distribution of nucleotides in genomes is not random, various biases are affecting the genome sequences from organisms spanning the three domains of life. For example, the G+C content affects the genome as a whole.

To visualize the biases in the nucleotides distribution in genomes, investigators have performed a variety of statistical analyses; these operations basically consisted in counting the nucleotides in a variety of subtle ways, while attempting to identify how the counting observed in real examples differed from a random distribution. Relevant statistical methods developed so far include the following: computation of correlations [1], power spectrum analysis [2,3], DNA walking analysis [4], computation of entropy [5,6], Hurst index estimation [7], detrended fluctuation analysis [8], wavelet analysis [9], mutual information function analysis [10], computational linguistics analysis [11].

Among the different biases observed in the nucleotides distribution in genomes, two stood out prominently. Both are short-range biases, i.e. correlating nucleotides over a short distance only, inferior to one thousand base pairs (bp), and both are affecting the genome as a whole. Both are present in many different organisms. This prevalent intensity and ubiquity is a hint that these biases are very likely to be the result of some strong physical con-

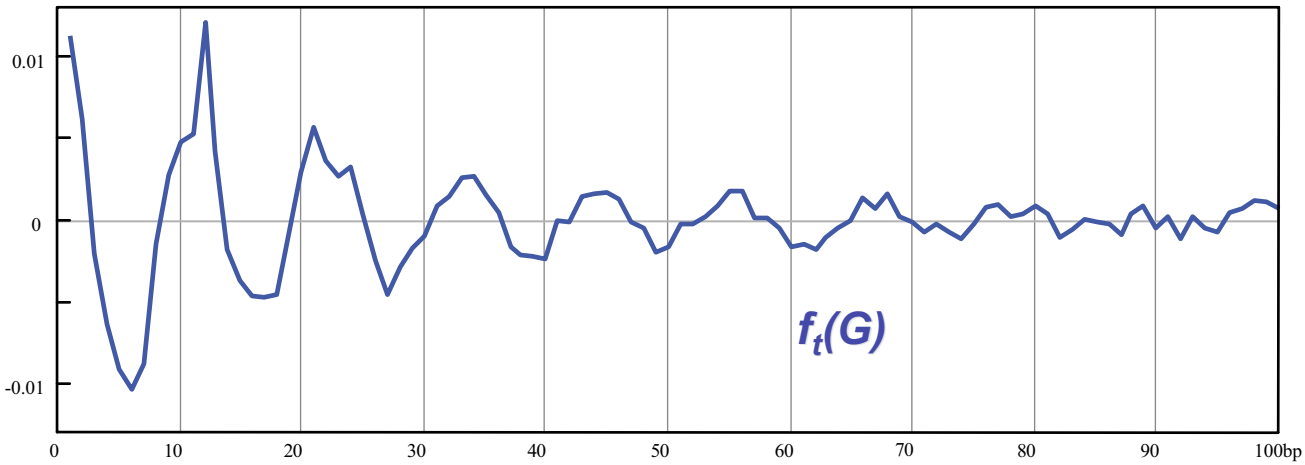

Figure 1
Deconvoluted correlation function of A following A in the genome of *H. pylori*. The correlation function has been treated so as to hide the most intense component of period 3 bp due to the presence of genes in the genome of *H. pylori*. After treatment, the function reveals a prevalent short-range component of period 11 bp. This component represents the prevalent short-range bias of period 11 bp in the distribution of nucleotides in the genome of *H. pylori*.

straints and/or biological functions acting on the affected genomes.

The first prevalent bias, the most intense one, is easily visualized in the genomes of all prokaryotes, as well as of lower eukaryotes. It also appears, though very dimly, in the genomes of higher eukaryotes. This bias is periodic with a periodicity of 3 bp (locally, the probability of presence of a given nucleotide depends on its position modulo three). This ubiquitous bias is effectively uncovered by power spectrum analysis [12-17]. Its presence has never been a mystery: it is due to the presence of protein coding genes in genomes. Indeed, the structure of the genetic code strongly affects the distribution of nucleotides within protein coding sequences, biasing the distribution of nucleotide triplets. As the gene density of higher eukaryotes is very small, this bias cannot easily be detected in these organisms. In contrast, for prokaryotes and for lower eukaryotes, in which the gene density is high, this bias is very easily detected. Its association to protein coding proved to be useful to locate exons in higher eukaryotic genomes [18]. This first bias is therefore generated by genomic sequences that are of strong biological significance.

Likewise, the second prevalent bias, also very intense, is visualized in the genomes of most prokaryotes and lower eukaryotes. For a given genome, the bias is encountered throughout the genome. In contrast with the previous 3 bp periodic bias, which spans large distances (typically several hundreds nucleotides) this bias does not involve nucleotides over a distance longer than about one hundred base-pairs: it is a short-range bias. It is also periodic, but this time with a fuzzy periodicity of mean value 11 bp. This signal has been visualized with the straightforward computation of correlations [1,19] or its equivalent, the power spectrum method [17]. The mean value of the periodicity of this bias varies from organism to organism. In the two articles just mentioned, the authors discuss the relation between phylogeny and the distribution of these periods. It turns out that it is generally of 10 bp for Archaea or hyperthermophilic Bacteria and 11 bp or more for the non-hyperthermophilic Bacteria, though there are many exceptions to this rule [19]. In the case of lower eukaryotes, a period of 10 bp for *C. elegans* and of 11 bp for *S. cerevisiae* has been observed. In the case of higher eukaryotes, a weak bias of period 10 bp is observed once the many repeated sequences present in these genomes have been removed from the analysis [19]. Moreover, in prokaryotes and lower eukaryotes, the bias is affecting coding sequences as well as non-coding sequences. This general observation is illustrated in Figure 1 with a graphic representation of the correlation function of nucleotide A following itself in the genome of *Helicobacter pylori*.

This function measures the probability to get a nucleotide A following another nucleotide A as their distance increases. The correlation function has first been treated by deconvolution so as to hide the overwhelming component of period 3 bp that results from the presence of genes in the genome (see above). The corresponding statistical treatment is described in the Methods section. In the graphic representation of the correlation function shown in Figure 1, there is a prominent component of period 11

bp. It appears as a short-range component as it completely vanishes for nucleotides located more than 70 bp apart. The periodic peaks do not occur every 11 bp exactly but every 10 bp to 12 bp. The strength of the periodic bias is illustrated by their large amplitude.

Although this bias is half as high in intensity as the one created by the presence of genes, and although it is ubiquitous in prokaryotes and lower eukaryotes, the nucleotide sequences generating this bias have not been determined so far. Nonetheless, the biological function that might be at the root of this bias has been proposed. In the case of Archaea, it has been suggested that the positioning of nucleosomes is controlled by some specific sequences, whose nature could however not be identified [1,19].

In the present article, we describe the program we designed, meant to discover the sequences that are generating every short-range bias (excluding the trivial one of period 3 bp generated by the genes) in genomes. Making use of this program, we discovered explicitly the sequences responsible for the bias of period 10–11 bp in the prokaryotic and lower eukaryotic genomes. These sequences, that we named "class A flexible patterns" for reasons that will be clarified in the course of this article, display a new type of organization. We show that the class A flexible patterns are ubiquitous in prokaryotes.

Results

Our aim was to identify the sequences that generate the 11 bp periodic short-range bias. To address this question, we designed a generic program to determine the sequences that generate any short-range bias in genomes nucleotides distribution (see the Methods section): the sequences responsible for the 11 bp periodic bias should belong to the sequences identified by the program.

For each genome of interest, the output of the program is given as a family of patterns. By pattern, we mean any succession of nucleotides with gaps in between (see the Methods section). The family of patterns returned by the program has the following property: the occurrences in the genome of all the patterns belonging to the pattern family match the sequences of the genome supposed to generate its short-range biases (see Methods section). Because of computation time limitations, our program gives an approximate result only: the patterns shape is restricted and the matching may not be exact (see the Methods and Discussion sections).

The program was run with 49 prokaryotic genomes, with four lower eukaryotic genomes and three viruses sequences. We collected the patterns of all the resulting family of patterns and saw that we could class them into two category of patterns. Naming them after their particular structural features, we called them the "rigid patterns" and the "flexible patterns". The rigid patterns are described first, but not discussed in details because they overlap with previously identified repeated sequences. Then we describe the more frequent but elusive flexible patterns. Among those, a great number belongs to a class that we called the "class A flexible patterns", for reasons explained below. The latter patterns are discussed extensively. Finally, we show that the occurrences of the class A flexible patterns define the sequences generating the bias of period 11 bp in genomes.

Rigid patterns

A rigid pattern is a pattern verifying the two following properties: first, the distance between the nucleotides making the pattern is the same for every occurrence of the pattern in the genome. Second, some variability in the nature of the nucleotides composing the pattern is allowed from one occurrence to another one. Most patterns described so far in the literature are rigid patterns. For rigid patterns, the exact distances between the nucleotides and the frequency of occurrence of the nucleotides A,T,G,C composing the pattern account for what is usually termed a "consensus sequence".

As a proof of concept, the program uncovered families of rigid patterns in a few selected genomes. Each family was made of short highly repeated motifs. As could be expected, when present in a genome, highly repeated sequences generate a short-range statistical bias. For example, we found the following rigid pattern in the genome of *Escherichia coli* (an x represents any nucleotide):

5**GC**xxx**AT**xxx**GC**xxxxxx**GC**xxx**AT**xxx**GC**-3'

One can recognize in this pattern a consensus for the repeated Bacterial Interspersed Mosaic Elements (BIMEs) sequences of *E. coli* [20]. It is important to note here that, although these sequences are recognized by our program because they create small but significant biases in the nucleotides distribution of *E. coli*, they do not contribute to the generation of the bias of period 11 bp. However, the very fact that we uncovered them is an independent validation of our approach.

Flexible patterns

To extend the rigid patterns description, we defined the "flexible patterns". A flexible pattern satisfies the two following properties: first, the nature of the nucleotides composing the pattern is the same for all the occurrences of the pattern in a given genome. Second, the distance between the nucleotides composing the pattern varies in a narrow range between occurrences of the pattern. Hence, a flexible pattern differs from a rigid pattern in that it

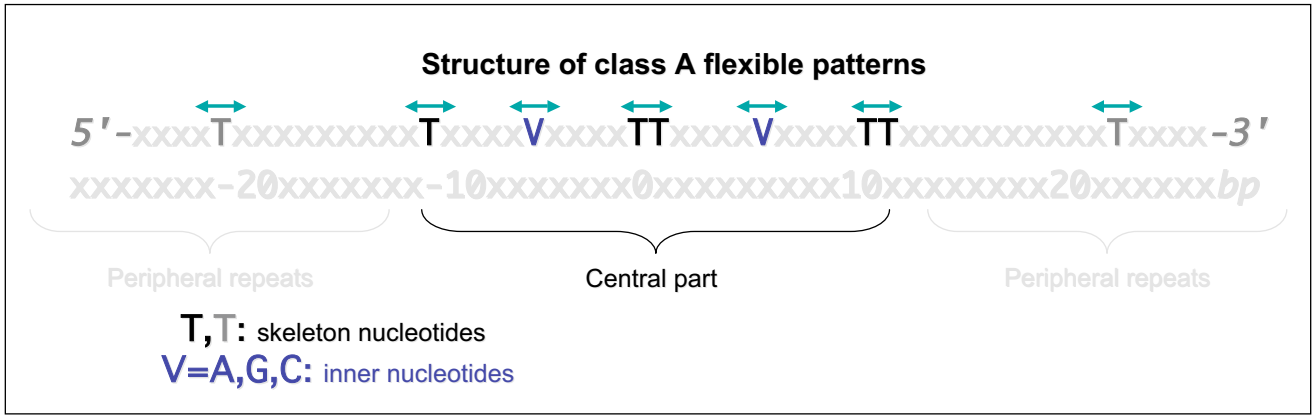

Figure 2
Diagrammatic structure of class A flexible patterns. Class A flexible patterns belong to the category of flexible patterns. Here "flexible" means that there is limited variation in the exact position of their conserved nucleotides. This is shown in the figure by the green arrows, indicating that the position of the conserved nucleotides may vary from one occurrence of the pattern to the next. The particular class of flexible patterns depicted here in the standard 5'-3' orientation is composed of two sets of conserved nucleotides. First, the patterns are shaped by a skeleton of regularly repeated Ts or TTs every 10 bp to 11.5 bp, spanning a maximum of 50 bp. These are called "skeleton nucleotides" and are symbolized by the black and dark grey Ts. The peripheral repeats of the skeleton, in dark grey, are sometimes absent from a given occurrence. The Ts of the central part, spanning 20 bp on average, are always present. Furthermore, class A flexible patterns are composed of a set of "inner nucleotides". These conserved nucleotides are represented here in dark blue. They can be any nucleotide but never Ts. They are located between the Ts of the skeleton and in the central part only.

could not generate a "consensus" by aligning sequences without introducing gaps. As an example, here are different occurrences of a flexible pattern found in the genome of

Pyrococcus furiosus
Gxx**A**xxx**TT**xxx**G**xxx**T**

Gxx**A**xxx**TT**xxx**G**xxx**T**

GAxxx**TT**xxxxx**G**xxx**T**

Gxx**A**xxx**TT**xxx**G**xxxxxx**T**

Gxxx**A**xxx**TT**xxx**G**xxxxxx**T**

Gx**A**xxx**TT**xxxx**G**xxxxxx**T**

5'-xxxxxx-20xxxxxxx--3'

From now on, we will represent a given flexible pattern not by its various spellings but by an average representative, in which the distance between the nucleotides is the mean distance of all the distance observed in all the various spellings. For example, we represent the previous flexible pattern by this average representative:

5'-**G**xx**A**xxx**TT**xxx**G**xxxxx**T**-3'

Conversely, in the following, a flexible pattern mentioned by an average representative is defined by the list of similar patterns which are deviating from the average representative by distances varying withing a narrow range between its conserved nucleotides.

The great majority of the patterns that we found by running our program in various genomes turned out to be of the flexible patterns category. We found on average approximately twenty flexible patterns in each genome, be it of a prokaryotic organism or of a lower eukaryotic organism. We observed that the distances between nucleotides composing the flexible patterns we identified vary generally from one to two base pairs. These patterns are composed of five to ten nucleotides spanning a distance of 10 bp to 60 bp. The nucleotides composing these patterns are most of the time either isolated or grouped as dinucleotides.

The description of patterns is limited by our program due to computing time limitations (see the Methods section), for example they cannot be composed of more than six nucleotides. The patterns that we get often seem to be subsets of longer patterns. In the following we mention the longest pattern that can be inferred, but it should be kept in mind that each of its detected variations are composed of only six nucleotides. For example, the following flexible pattern found in *H. pylori*:

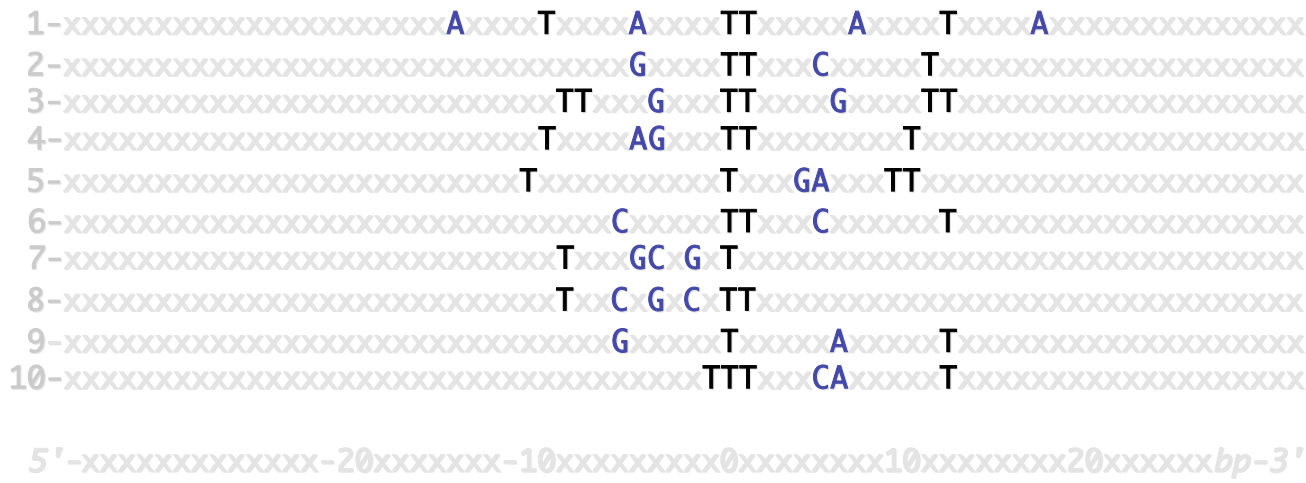

Figure 3
A few identified class A flexible patterns. Ten related yet distinct class A flexible patterns common to different genomes have been identified so far. Their structures share common features, which are characteristic of class A flexible patterns. Peripheral repeats of the skeleton nucleotides of the patterns have not been represented here. Skeleton nucleotides are shown in black. Inner nucleotides are shown in dark blue.

5'-**T**xx**A**x**GC**x**TTT**-3'

is defined by the following variations:

Txxx**GC**xx**TT**

Txxxx**GC**xx**TT**x**T**

Txxxx**GC**xx**T**x**TT**

Txxxxx**GC**xx**TTT**

Txxxxx**A**x**GC**x**TT**

Axx**GC**x**TTT**

Ax**GC**x**TT**x**T**

Axx**GC**x**TT**x**T**

Class A flexible patterns

Among flexible patterns, we observed that a great majority shared a similar structure and were thus easily identifiable. We named "class A flexible patterns" this subset of flexible patterns. We will restrict our study to these patterns, as they account for most, if not all, of the 11 bp period found in the genomes we analyzed.

All class A flexible patterns, though different in spelling, share the same structure, as depicted in Figure 2. The structural features illustrated in this figure are formally defining the class A flexible patterns. The patterns are described here in the standard 5'-3' orientation.

Class A flexible patterns are in total composed of five to ten conserved nucleotides spanning a length of approximately 11 bp to 50 bp. The conserved nucleotides are either isolated or grouped as dinucleotides.

That these patterns belong to the category of flexible patterns is illustrated in Figure 2 by the green arrows above the nucleotides composing the patterns (always isolated nucleotides or dinucleotides). The distance between any of the isolated nucleotides or dinucleotides varies by 1 bp to 2 bp from one occurrence of the pattern to the next in a given genome. Class A flexible patterns are composed of two subsets of conserved nucleotides: the skeleton nucleotides and the inner nucleotides.

The skeleton nucleotides consist of two to five repeats of the single nucleotide T or of the dinucleotide TT, regularly spaced every 10 bp to 11 bp on average. The central part (nucleotides represented in black in Figure 2) is made of two to three repeats. These repeated nucleotides appear at every occurrence of a given pattern in a given genome. Outlying repeats (nucleotides in dark grey in Figure 2) may extend the skeleton outside the central part. Those are involving single nucleotides Ts exclusively and are not always present: they do not appear in every occurrence of a given pattern. Typically, one or two such peripheral repeats of the single nucleotide T on each side of the central part of the skeleton exist in a given occurrence of a pat-

Table 1: Distribution of class A flexible patterns in genomes.

	1	2	3	4	5	6	7	8	9	10	
Aeropyrum pernix	X		X	X							Archaea; Crenarchaeota; Thermoprotei; Desulfurococcales
Sulfolobus solfataricus	X		X	X	X						Archaea; Crenarchaeota; Thermoprotei; Sulfolobales
Sulfolobus tokodaii			X	X	X						Archaea; Crenarchaeota; Thermoprotei; Sulfolobales
Pyrobaculum aerophilum		X								X	Archaea; Crenarchaeota; Thermoprotei; Thermoproteales
Archaeoglobus fulgidus			X	X	X						Archaea; Euryarchaeota; Archaeoglobi; Archaeoglobales
M. Acetivorans	X		X	X	X		X	X			Archaea; Euryarchaeota; Methanosarcinales
Halobacterium sp.			X								Archaea; Euryarchaeota; Halobacteriales
M. thermoautotrophicum			X	X							Archaea; Euryarchaeota; Methanobacteriales
Methanococcus jannashii	X			X	X						Archaea; Euryarchaeota; Methanococcales
Pyrococcus abyssi			X	X	X						Archaea; Euryarchaeota; Thermococcales
Pyrococcus furiosus	X		X	X	X						Archaea; Euryarchaeota; Thermococcales
Pyrococcus horikoshii			X	X	X						Archaea; Euryarchaeota; Thermococcales
Thermoplasma acidophilum				X		X					Archaea; Euryarchaeota; Thermoplasmatales
Tropheryma whipplei		X	X						X		Bacteria; Actinobacteria; Actinomycetales
Aquifex aeolicus	X			X					X		Bacteria; Aquificae; Aquificales
Chlorobium tepidum	X						X				Bacteria; Chlorobi; Chlorobiales
Synechocystis sp.											Bacteria; Cyanobacteria; Chroococcales
Deinococcus radiodurans		X						X			Bacteria; Deinococcus-Thermus; Deinococcales
Bacillus subtilis	X							X			Bacteria; Firmicutes; Bacillales
Oceanobacillus iheyensis	X										Bacteria; Firmicutes; Bacillales
Listeria monocytogenes	X					X					Bacteria; Firmicutes; Bacillales
T. Tengcongensis	X		X								Bacteria; Firmicutes; Clostridia; Thermoanaerobacteriales
Streptococcus pneumoniae	X										Bacteria; Firmicutes; Lactobacillales
Pirellula sp.	X										Bacteria; Planctomycetes; Planctomycetales
Magnetactic cocci		X								X	Bacteria; Proteobacteria
Caulobacter vibrioides		X									Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacteriales
Agrobacterium tumefaciens		X						X			Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales
Sinorhizobium meliloti								X			Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales
Rickettsia conorii	X		X				X				Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsialles
Rickettsia prowozekii	X	X	X		X		X				Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsialles
Bordetella pertussis	X						X				Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales
Neisseria meningitidis		X									Bacteria; Proteobacteria; Betaproteobacteria; Neisseriales
Campylobacter jejuni	X	X									Bacteria; Proteobacteria; Epsilonproteobacteria; Campylobacterales
Helicobacter hepaticus	X	X	X	X							Bacteria; Proteobacteria; Epsilonproteobacteria; Campylobacterales
Helicobacter pylori		X	X				X				Bacteria; Proteobacteria; Epsilonproteobacteria; Campylobacterales
Wolinella succinogenes		X	X		X						Bacteria; Proteobacteria; Epsilonproteobacteria; Campylobacterales
P. haloplanktis		X					X				Bacteria; Proteobacteria; Gammaproteobacteria; Alteromonadales
Candidatus bl. floridanus	X										Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales
Buchnera aphidicola	X										Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales
Escherichia coli		X									Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales
Wigglesworthia glossinidia	X										Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales
Coxiella burnetii	X										Bacteria; Proteobacteria; Gammaproteobacteria; Legionellales
Haemophilus influenzae	X	X			X	X	X				Bacteria; Proteobacteria; Gammaproteobacteria; Pasteurellales
Pseudomonas aeruginosa		X						X			Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales
Pseudomonas putida	X	X					X	X			Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales
Vibrio vulnificus		X				X	X				Bacteria; Proteobacteria; Gammaproteobacteria; Vibrionales
Xylella fastidiosa	X	X			X	X		X			Bacteria; Proteobacteria; Gammaproteobacteria; Xanthomonadales
Leptospira interrogans	X										Bacteria; Spirochaetes; Spirochaetales
Thermotoga maritima	X								X		Bacteria; Thermotogae; Thermotogales
Plasmodium falciparum											Eukaryota; Alveolata; Apicomplexa

Table 1: Distribution of class A flexible patterns in genomes. *(Continued)*

Saccharomyces cerevisiae	X		X	Eukaryota; Fungi; Ascomycota
Encephalitozoon cuniculi				Eukaryota; Fungi; Microsporidia
Caenorhabditis elegans	X	X	X	Eukaryota; Metazoa; Nematoda
Enterobacteria phage T4	X			Virus; Enterobacteria phage T4
S. tengcon.. Vvrus STSV1	X			Virus; Fusellovirus
Human herpesvirus 4				Virus; Human herpesvirus 4

1. AxxxxTxxxxAxxxxTTxxxxxAxxxxTxxxxA
2. GxxxxTTxxxCxxxT
3. TTxxxGxxxTTxxxxGxxxxTT
4. TxxxxAGxxxTTxxxxxxxxT
5. TxxxxxxxxxxTxxxGAxxxTT
6. CxxxxxTTxxxCxxxxxxT
7. TxxxGCxGxT
8. TxxCxGxCxTT
9. GxxxxxTxxxxxAxxxxxT
10. TTTxxxCAxxxxxT

tern. Note that for a given pattern, the distance (averaged over all the occurrences of the given pattern in a given genome) between two neighboring isolated conserved nucleotides Ts or dinucleotides TTs of the skeleton ranges from 7 bp to 12 bp. Yet, the average of these distances over the two to five repeats of the skeleton of the given pattern remains inside the interval of 10 bp to 11.5 bp. The skeleton structure, spanning up to 50 bp in total, is basically the same for all class A flexible patterns, for only the distances between the Ts and the choice of single or dinucleotides can fluctuate.

The inner nucleotides consist of one to three conserved nucleotides located exclusively in the central part of the skeleton. Most importantly, these conserved nucleotides are found to be either A, G or C (a particular nucleotide specifying the particular kind of pattern identified, see Figure 3) but never T. They are either isolated or grouped as dinucleotides (isolated conserved nucleotides are more frequent than conserved dinucleotides). There can be only one isolated nucleotide or dinucleotide between two neighboring skeleton nucleotides. The position of the inner nucleotides is usually located exactly in the middle of two neighboring Ts of the skeleton. These inner nucleotides play a discriminating role in class A flexible patterns as they differentiate patterns from one another.

The central part of these patterns is composed of three to six skeleton nucleotides and of two to four inner nucleotides (see Figure 2). Altogether, the central part is composed on average of six conserved nucleotides covering from 10 bp to 33 bp. This part of the patterns is the one that varies from one class A flexible pattern to another, both in the choice of single or dinucleotides in the skeleton and in the nature of the inner nucleotides. Therefore, we choose to subsequently identify the patterns using this central part only.

The program we ran is limited to identification of patterns spanning up to a maximum of 60 bp (see the Methods section). This implies that we may have been missing some peripheral repeats of Ts in some occurrences of the patterns, but we did not miss important nucleotides as the latter are located in the central parts of the patterns only.

Distribution of class A flexible patterns in organisms

As a whole, cumulating all the tested genomes, we could identify twenty different types of class A flexible patterns. Some genomes harbor specific class A flexible patterns that are found in no other genome. In contrast, some types of patterns are found in more than one genome. We could identify ten such conserved types of patterns. In Figure 3, we list these ten types of class A flexible patterns.

Patterns numbered 1 to 5 in Figure 3 are present in many genomes, patterns numbered 6 to 10 are present in less than ten different genomes.

In Table 1, we display the organisms in which these patterns were identified, as well as the phylogenetic family to which the organisms belong. It turned out that every one of the 49 prokaryotic genomes tested, two of the four lower eukaryotic genomes tested (*Saccharomyces cerevisiae* and *Caenorhabditis elegans*) and the two genomes of bacteriophages analyzed were harboring class A flexible patterns.

First, we found out that class A flexible patterns are ubiquitous in prokaryotes. Indeed, each of 49 genomes of prokaryotes tested harbors one or more different types of class A flexible patterns. The genome of *Xylella fastidiosa* harbors for instance five different types of patterns. Usually, each genome harbors two to four different types of class A flexible patterns. Second, each of the patterns numbered 1 through 5 in Figure 3 is present in more than 10 different genomes. This makes it possible to discuss the

nature of the distribution of these five types of patterns in genomes.

Pattern 1 has been detected in more than 50% of the 56 tested genomes, with no relationship to phylogenetic branches as we found it in Archaea, in Bacteria, in lower eukaryotes and in phages (see Table 1). This pattern alone may be ubiquitous as a low content of this pattern in a given genome would fail to be detected by our approach.

Pattern 2 is present in a total of 19 genomes. Out of these 19 genomes, 16 belong to *Proteobacteria*. Three further genomes, that do not belong to the *Proteobacteria* clade, display this type of pattern. Among those, we found first two Bacteria: *Deinococcus radiodurans* and *Tropheryma whipplei*. The former lives under highly desiccated or radiation-exposed conditions, with remarkable features in DNA maintenance [21], while the latter is a highly degenerate parasite [22]. The third organism which is not a *Proteobacteria* and where this type of pattern is present is an Archaeon: *Pyrobaculum aerophilum* [23]. Overall, the distribution of pattern number 2 in genomes is highly correlated with the *Proteobacteria* class of organisms. It is present throughout this class of organisms as it has been detected in some genomes of the alpha, beta, epsilon and gamma groups (the delta group has not yet been analyzed). It is also remarkably present in all tested genomes of the epsilon group.

Pattern 3 is present in 18 genomes in total, in Archaea, in Bacteria and in lower eukaryotes. Pattern 4 is present in 13 genomes in all. It has been identified in 11 of the 13 archaeal genomes analyzed (in *Crenarcheota* as well as in *Euryarchaeota*). It is also present in two Bacteria (*Aquifex aeolicus* and *Helicobacter hepaticus*). Hence, the distribution of this pattern in genomes seems to be somewhat correlated with the archaeal kingdom.

Pattern 5 is present in 14 genomes in total, in Archaea, in Bacteria and in lower eukaryotes. The other identified class A flexible patterns are present in only a few organisms. Moreover, these organisms do not clearly belong to any specific phylogenetic lineage. In Figure 4 are summarized the few parallels that could be drawn between the distribution of class A flexible patterns and phylogeny. Each of these three patterns is present in more than 10 genomes out of the 56 tested.

Distribution of class A flexible patterns in a given genome

The occurrences of class A flexible patterns are equally distributed in the two strands of chromosomes. These occurrences cover a considerable part of each genome. The conserved nucleotides of all occurrences of all class A flexible patterns are involving up to one fourth of the total number of nucleotides of a given genome (24% in the case of *H. pylori*). If we take into consideration the total length that the occurrences of the patterns span in a genome, then it comes up to one half of each genome (51% in the case of *H. pylori*). In the case of *H. pylori*, the span of the patterns ranges from 9 bp to 29 bp (Table 2). We observed that the patterns' occurrences can be overlapping. Interestingly, class A flexible patterns occur indifferently in coding and in non-coding regions of genomes. They are neither correlated with the leading nor with the lagging strand of chromosomes. All things considered, there seems to be no obvious bias in the distribution of the occurrences of the patterns.

Contribution of class A flexible patterns to the 11 bp periodic bias

The structure of class A flexible patterns is highly reminiscent of the 11 bp periodic bias in genomes of prokaryotes and lower eukaryotes. Indeed, the patterns have a core of repeated Ts or TTs every 10 bp-11 bp on average in all occurrences. It can therefore be expected that because these periodic nucleotides are densely spread, a bias of period 10 bp-11 bp will be generated in the corresponding genome sequences. The length of the patterns when the peripheral repeats are considered (up to 60 bp) is on the same order as the span of the 10 bp-11 bp periodic component in the correlation between nucleotides (see Figure 1). Furthermore, we systematically observed that the component of period 11 bp is somewhat fuzzy (see the blunt shaped peaks in Figure 1). This is consistent with the fact that the distance between neighboring skeleton nucleotides ranges from 7 bp to 12 bp. This is also consistent with the involvement of dinucleotides in class A flexible patterns. Finally, the occurrences of class A flexible patterns distribute throughout a given genome, with no apparent preference for coding or non-coding regions, similarly to the bias of period 10–11 bp. Now we want to show that the class A flexible patterns are indeed the source of the 11 bp periodic bias in genomes. We illustrate this with the genome of *H. pylori* as the statistical bias of period 11 bp is particularly prominent there. We got the same results for all other genomes analyzed.

The class A flexible patterns discovered in the *H. pylori* genome are the following:

1-5'-**T**xxxxxxxxxx**T**xxxx**G**xxx**TT**x**T**-3'

2-5'-**GG**xx**TTT**xxxxxxxxxx**T**xxxxxxxxx**T**-3'

3-5'-**T**xxxxxxxxx**TTT**xx**AA**x**C**xx**T**-3'

4-5'-**GG**xx**TTT**xxxxx**C**-3'

5-5'-**T**xx**A**x**GC**x**TTT**-3'

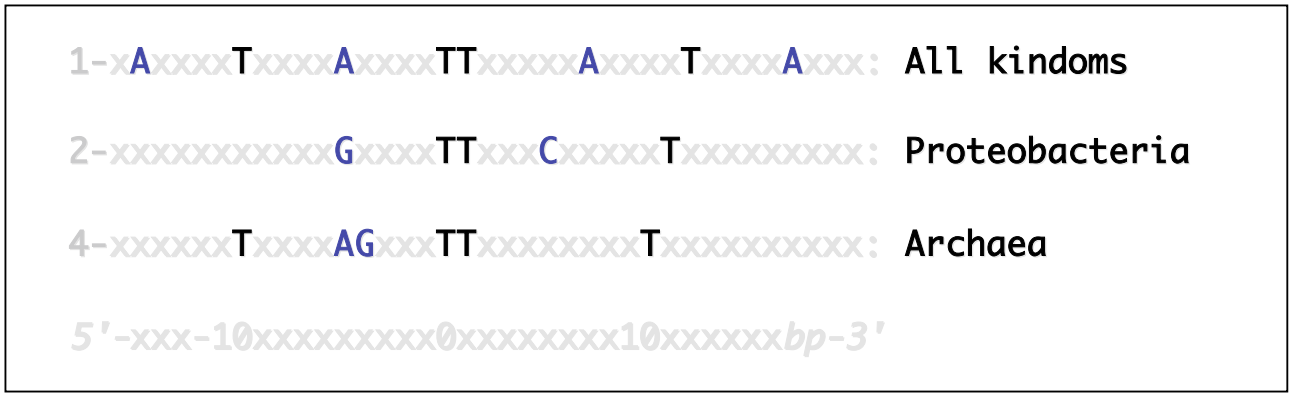

Figure 4
The distribution of three types of class A flexible patterns is correlated to specific phylogenetic groups of organisms. We identified five class A flexible patterns distributed in many different organisms. Three of them, displayed here, show a distribution which can be related to the phylogeny.

Patterns numbered from 1 to 3 are also found in genomes of other organisms, while patterns 4 and 5 are found only in this genome. *Helicobacter pylori* is remarkable as the skeleton nucleotides are composed of the trinucleotide TTT. For each of those flexible patterns, Table 2 illustrates the list of their variations. No peripheral repeats are displayed, as we failed to determine any in this particular genome. It is interesting to note that all the variations of these five patterns are indeed over-represented in the genome of *H. pylori*. We compared the number of occurrences of the patterns in the authentic genome to the number of occurrences in a model genome that keeps only the crude statistical features of the nucleotide distribution in the *H. pylori* genome (see the Method section). We found that the variations of pattern 1 occur approximately 30% more often in the authentic genome than in the model genome, the variations of pattern 2 approximately 40%, the variations of pattern 3 approximately 30%, the variations of pattern 4 approximately 40%, the variations of pattern 5 approximately 30%. All the nucleotides involved in the occurrences of patterns 1 to 5 and of their reverse complements amount to 24% of the total number of nucleotides contained in the whole genome. To explore whether the bias of period 11 bp in the distribution of the nucleotides is due to these 24% of the genome of *H. pylori*, we constructed two reference genomes for comparison.

We constructed a first "deconvoluted" genome $G_{mo}(G^-)$ in the following way (see the Methods section): starting from the authentic genome of *H. pylori*, every nucleotide which belongs to any occurrence of any of the five class A flexible patterns or of their reverse complements is replaced by the nucleotide of a model genome preserving the local composition in hexanucleotides of the authentic genome but not their order (see the Methods section) while every other nucleotide is kept unaltered. We plotted the treated correlation function of $G_{mo}(G^-)$ for the nucleotide A following A (see the Methods section) in Figure 5. The 11 bp periodic bias is now absent from this plot. This means that the 76% of the genome of *H. pylori* which is not covered by class A flexible patterns does not have any significant 11 bp periodic statistical bias. Hence, we concluded that class A flexible patterns are generating the 11 bp bias in genomes.

Interestingly, the 11 bp periodic bias disappeared even at correlations over 30 bp, despite the fact that our patterns are never longer that 30 bp for this genome (we have deconvoluted the central parts of the patterns but not the hypothetical peripheral repeats). Deprived of the core sequences of the patterns, the peripheral repeats, even if they exist, can no longer generate much bias. In Figure 5, one can notice a small peak pointing downwards at 11 bp. This probably reflects the fact that we failed to describe accurately the patterns and therefore removed too many sequences, some of which artefactually taken as genuine patterns. Second, we plotted the treated correlation function (see the Methods section) of a complementary model: $G_{mo}(G^+)$, the "convoluted" genome (Figure 6). As in the preceding model, $G_{mo}(G^+)$ is built starting from the authentic genome of *H. pylori*: all the nucleotides not belonging to occurrences of class A flexible patterns and of their reverse complements are replaced by the nucleotides of a model genome (see the Methods section). The 11 bp statistical bias from the original genome is now visible again (the treated correlation function of the original genome is shown in Figure 1). The correlations over 30 bp are hardly visible, which is consistent with the fact that no peripheral repeats were introduced in the convolution

Table 2: The variations defining the five class A flexible patterns found in the genome of *H. pylori*

1-5'-xxxxxxxxxxxxxxx**T**xxxxxxxxxx**T**xxxx**G**xxx**TT**x**T**xxxxxxxxxxxxxxxxxxxxxx-3'
xxxxxxxxxxxxxxxxxxx**T**xxxxxxxxxx**T**xxxx**G**xxx**TT**x**T**xxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxx**T**xxxxxxxxxx**T**xxxx**G**xxx**TT**x**T**xxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxx-20xxxxxxx-10xxxxxxxxx0xxxxxxxxx10xxxxxxx20xxxxxxx30
2-5'-xxxxxxxxxxxxxx**GG**xx**TTT**xxxxxxxxxx**T**xxxxxxxxx**T**xxxxxxxxxxxxxxxxxx-3'
xxxxxxxxxxxxxxxxxx**G**xxx**T**x**TT**xxxxxxxxx**T**xxxxxxxxxx**T**xxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxx**G**xx**TTT**xxxxxxxxx**T**xxxxxxxxxx**T**xxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxx**G**xx**TTT**xxxxxxxxxx**T**xxxxxxxxxx**T**xxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxx**G**xx**TTT**xxxxxxxxxx**T**xxxxxxxxx**T**xxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxx**GG**xx**TTT**xxxxxxxxxx**T**xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxx**GG**xx**TT**x**T**xxxxxxxxx**T**xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxx**G**xx**TTT**xxxxxxxxx**TT**xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxx**G**x**G**xx**TTT**xxxxxxxxxx**T**xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxx-20xxxxxxx-10xxxxxxxxx0xxxxxxxxx10xxxxxxx20xxxxxxx30
3-5'-xxxxxxxxxxxxxxxx**T**xxxxxxxxx**TTT**xx**AA**x**C**xx**T**xxxxxxxxxxxxxxxxxxxxxx-3'
xxxxxxxxxxxxxxxxxxx**T**xxxxxxxxx**TT**xx**A**xx**C**xx**T**xxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxx**T**xxxxxxxxx**T**xx**AA**x**C**xx**T**xxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxx**T**xxxxxxxxxx**T**xx**AA**x**C**xx**T**xxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxx**T**xxxxxxxxx**TT**xx**AA**xx**C**xxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**T**x**AA**x**CC**x**T**xxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**T**xx**AA**x**CC**xx**T**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**TT**xx**A**x**CC**xx**T**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**TTT**xx**A**xx**C**xx**T**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**TT**x**T**xxx**A**xx**C**xx**T**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**T**x**TT**xx**A**xx**C**xx**T**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxx-20xxxxxxx-10xxxxxxxxx0xxxxxxxxx10xxxxxxx20xxxxxxx30
4-5'-xxxxxxxxxxxxxxxxxxxxxxxxxxx**GG**xx**TTT**xxxxx**C**xxxxxxxxxxxxxxxxxxxxxx-3'
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**GG**x**TTT**xxxx**C**xxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**GG**xx**TT**x**T**xxxx**C**xxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**GG**xx**T**x**TT**xxxx**C**xxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**G**x**G**x**T**x**TT**xxxx**C**xxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**G**x**G**xx**TTT**xxxx**C**xxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxx-20xxxxxxx-10xxxxxxxxx0xxxxxxxxx10xxxxxxx20xxxxxxx30
5-5'-xxxxxxxxxxxxxxxxxxxxxxxxxxx**T**xx**A**x**GC**x**TTT**xxxxxxxxxxxxxxxxxxxxxxx-3'
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**T**xxx**GC**xx**TT**xxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**T**xxxx**GC**xx**TT**x**T**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**T**xxxx**GC**xx**T**x**TT**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**T**xxxxx**GC**xx**TTT**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**T**xxxxx**A**x**GC**x**TT**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**A**xx**GC**x**TTT**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**A**x**GC**x**TT**x**T**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**A**xx**GC**x**TT**x**T**xxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxx-20xxxxxxx-10xxxxxxxxx0xxxxxxxxx10xxxxxxx20xxxxxxx30

|p463pt|The variations defining the five class A flexible patterns found in the genome of*H. pylori*

process. In this "realistic" imitation of the *H. pylori* genome, the correlations below 30 bp are somewhat too intense when compared to the real ones, displayed in Figure 1. This shows again that we removed too many core sequences, as they were not described with enough accuracy. The sum of the treated correlation function of the deconvoluted genome and of the treated correlation function of the convoluted genome fails to be exactly equal to the treated correlation function of the authentic genome. This shows that there exist correlations between occurrences of class A flexible patterns and of neighboring sequences. It can be expected that these correlations involve the undetected peripheral repeats.

Finally we must note that we chose to illustrate the relationship between class A flexible patterns and the 11 bp bias with the correlation function calculated for an A following an A, as the correlations are specially strong for

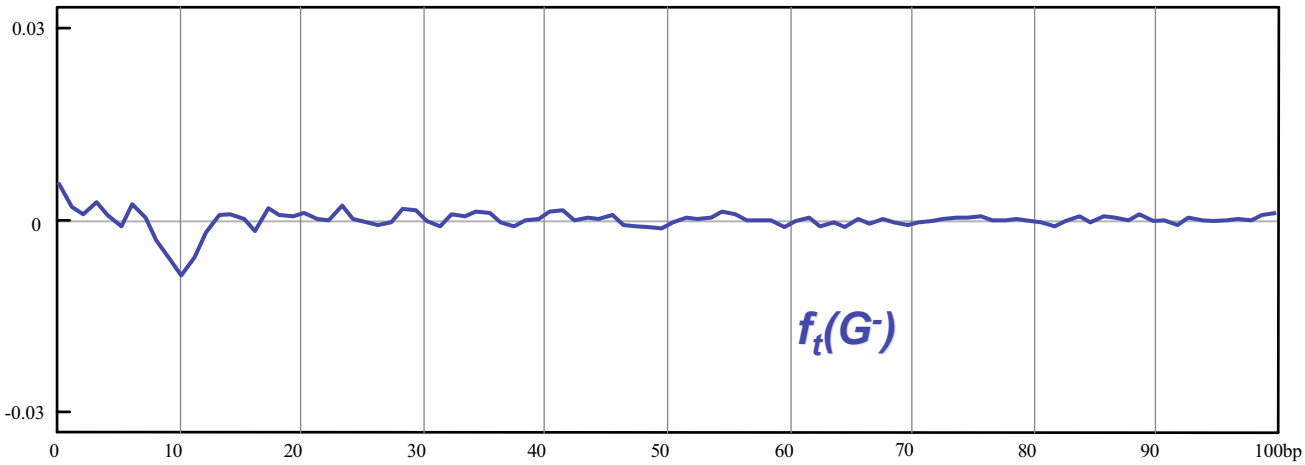

Figure 5
The treated correlation function of $G_{mo}(G^-)$. This correlation function of nucleotide A following A reveals biases generated by the part of the genome of *H. pylori* that do not contain occurrences of class A flexible patterns.

those two nucleotides. However the results reported are still valid for any combination of two nucleotides.

Discussion

In the present work we focused on class A flexible patterns as they are the source of the 11 bp periodic bias long known to exist in genomes. Because of the technical limitations of our approach we expect that there may still be other classes of flexible patterns in DNA sequences. They must be however relatively less important as genome sequences do not display prominent short-range biases other than the 3 bp and the 11 bp periodic long identified, while deconvolution of authentic genome sequences from the patterns we identified yielded sequences which no longer displayed any outstanding periodicity.

Limitations in the description of class A flexible patterns

As explained in the Methods section, our approach suffers some limitations, mainly due to computational time limitations. First, simply for stochastic reasons (the signal must be significantly higher than the noise), we would not find sequences that are generating weak biases or that are present in a too limited amount in genomes (with a frequency below $\frac{1}{3000} bp^{-1}$). Hence we probably missed the presence of some class A flexible patterns in some genomes. Second, the output of our program may have been somewhat inaccurate. Namely, because of the limitation we had to impose on the correlations order (see the Methods section), we may have identifed some patterns as genuine while they would represent a mix of different patterns present at distinct locations in the genomes. Third, we are bound to miss completely any pattern in which the shorter distance between conserved nucleotides is longer than 14 bp (see the Methods section). Fourth, the patterns spellings are but an approximation. Our program has restrictions in the maximum length and number of conserved nucleotides of patterns it is able to determine. As a consequence, we may have missed peripheral parts of the patterns we identified. Still, these restrictions probably did not affect much our spelling of class A flexible patterns, as these patterns are short enough: the central parts span only 20 bp on average. In contrast, in the identification of rigid patterns, typically made of continuous sequences of conserved nucleotides ("words" or "motifs"), we could not retrieve all conserved nucleotides. This was not, however, the main goal of this work.

Connection to optimal growth temperature

As phylogeny cannot account for the distribution of patterns numbered 3 and 5 in Figure 3, we may wonder whether the distribution of these two class A flexible patterns could be related to physical or biological parameters of the organisms in which they have been identified. We took into account the Gram staining, the cell shape, oxygen dependency, sporulation ability, encapsulation ability, optimal pH and maximum growth temperature, GC content and GC skew. Among those features, the optimal growth temperature somewhat correlates with the distribution of these class A flexible patterns. Indeed, both patterns are present mostly in thermophilic organisms. Still, it remains difficult to draw any firm conclusion in this matter as all tested Archaea but one (*Methanosarcina ace-*

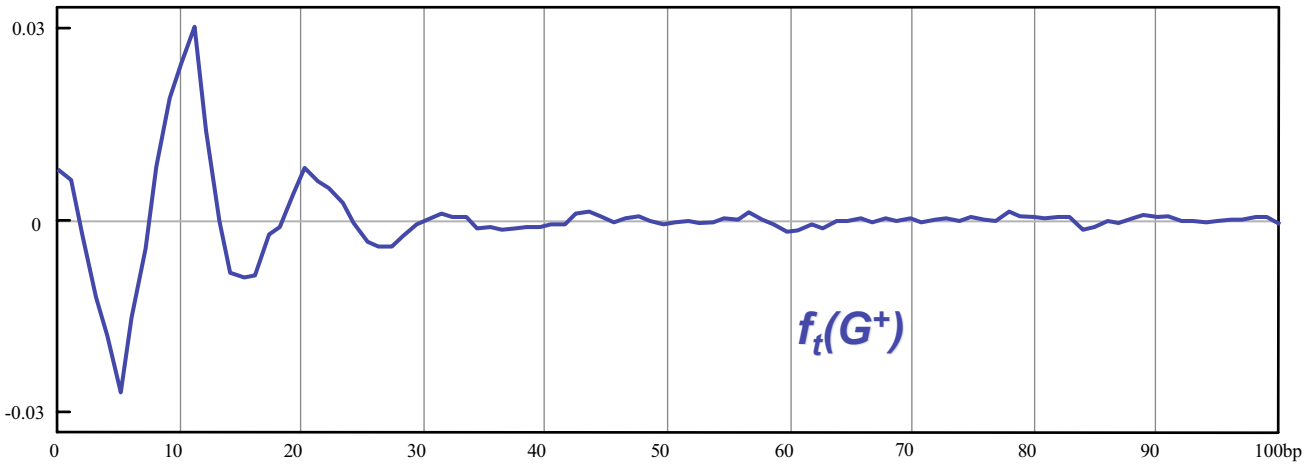

Figure 6
The treated correlation function of $G_{mo}(G^+)$. This correlation function of nucleotide A following A reveals biases in the genome of *H. pylori* which are generated by the occurrences of class A flexible patterns in its genome.

tivorans) are thermophilic and as these patterns are found mostly in Archaea. The question thus arises to determine whether these patterns are present in archaeal organisms or in thermophilic organisms. It is not yet possible to draw a clear rule from the presently tested genomes.

Class A flexible patterns may define protein interaction sites on the DNA molecule

The very structure of class A flexible patterns offers precious hints to conjecture their biological function. The hypothesis we propose is that the patterns are the signatures of DNA-protein interaction sites. Five arguments tend to support this idea. These are only theoretical arguments and our hypothesis needs to be substantiated by further experiments. First argument: to our knowledge, the length of class A flexible patterns is in a range appropriate for DNA-protein interactions. The total length of the patterns ranges from 11 bp to 60 bp while the length of the central part ranges from 10 bp to 33 bp (see Figure 2). The size of the DNA-protein binding sites usually ranges from 10 bp to 40 bp [24,25]. Hence the central part of the patterns, which is specific and conserved, may be the interacting protein-DNA interface.

Second argument: the number of conserved nucleotides composing the central parts of class A flexible patterns (six on average, see Figure 2) is compatible with the hypothesis. Indeed, if more nucleotides were conserved in the sequence, it is likely that the interaction would be very strong and would therefore have been already identified. Furthermore it would correspond to a stable interaction that would presumably preclude any function of the DNA molecule requiring its opening. In contrast, if there were fewer conserved nucleotides, the interaction would be too weak to create a specific interaction with proteins. Previous studies have established that the average number of conserved nucleotides in DNA-protein interaction sites ranges from five to ten conserved nucleotides [25].

Third argument: the position of the conserved nucleotides of class A flexible patterns is remarkably consistent with the hypothesis of a DNA-protein interaction site. Class A flexible patterns are composed of a skeleton made of regularly repeated Ts or TTs every 10 bp-11.5 bp on average. As the shape of the DNA molecule is helical, with a pitch of average 10.5 bp, varying from 10 bp to 12 bp [26], when unbound, repeated conserved nucleotides of the skeleton always appear at the same side of the helix, in the major groove and in the minor groove respectively (see Figure 7). Inner nucleotides of the patterns, which are always A, G or C depending on the particular pattern considered, are set between the repeated Ts of the skeleton, most often in the middle of two neighboring repeats. Hence, the inner nucleotides also appear on the same two sides of the DNA molecule, through grooves that are opposite to those of the skeleton nucleotides. Note that interactions between proteins and DNA minor grooves are well documented [27,28].

The spatial structure of the DNA molecule of class A flexible patterns is illustrated in Figure 7. The nucleotides composing the example pattern of the figure are accessible from the upper side, with the skeleton nucleotides visible through major grooves and the inner nucleotides visible through minor grooves, or from the lower side, with the skeleton nucleotides visible through minor grooves and the inner nucleotides visible through major grooves.

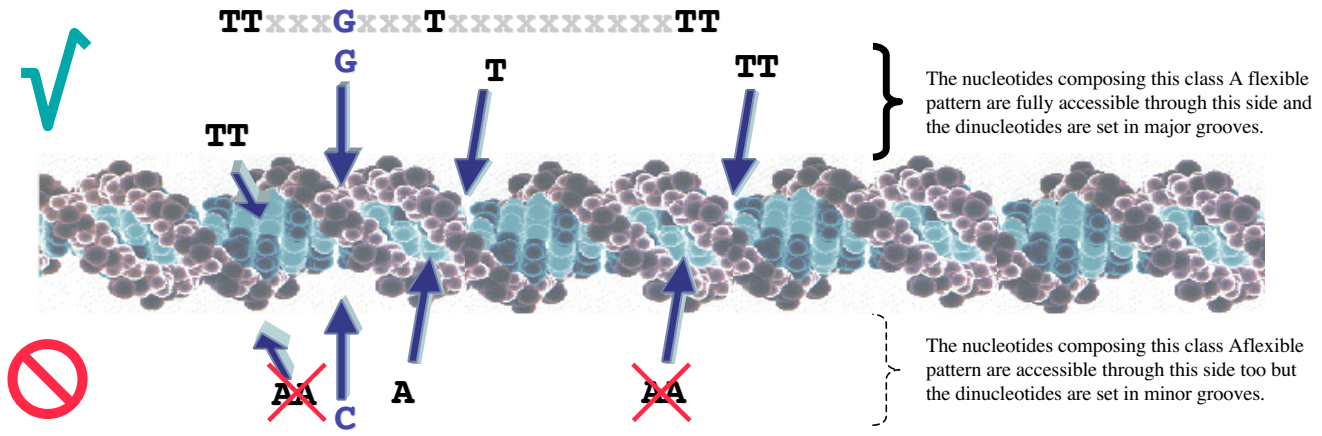

Figure 7
Accessibility of class A flexible patterns. There are two opposed sides from which nucleotides composing this occurrence of this given class A flexible pattern are accessible. The dinucleotides are visible through major grooves only from the upper side and hence fully accessible from this side only. Hence a given occurrence of a given class A flexible pattern in a genome is only accessible from one side of the DNA molecule.

The skeleton of the patterns is half composed of repeated dinucleotides TTs. In contrast, inner nucleotides are mostly isolated conserved nucleotides. A dinucleotide may be less easily accessed through a minor groove because this groove is too narrow. Conversely, it may be easily accessed through a major groove as the latter is wider. Hence, class A flexible patterns may be actually accessible by only one of the two opposed sides of the DNA double helix, the one where skeleton nucleotides are seen through major grooves, as shown in Figure 7. This gives a very specific argument to think that the function of these patterns may be to define interaction sites with some proteins. Indeed, a protein interacting with the DNA molecule usually comes along one defined side of the molecule and at any rate is never covering the molecule on all sides [29]. The position of the nucleotides composing the patterns is fully consistent with this requirement.

Fourth argument: class A flexible patterns belong to the group of flexible patterns. This means that the exact position of conserved nucleotides of the patterns varies from one occurrence of the patterns in genomes to the next one. This property is fully consistent with the hypothesis that the patterns are signatures of motifs allowing interaction with a geometrically rigid protein, as explained below.

The DNA molecule is a flexible molecule that can be elastically bent, elongated and supercoiled negatively or positively. As a matter of fact, in living cells, the molecule keeps on being constrained by thermal agitation and even more dramatically by the constant action of various molecules. For example, the action of polymerases will induce strong supercoiling ahead and behind where it acts [30]. Finally, the pitch and bending of the DNA helix keeps on varying locally, depending in particular on the local base composition [31].

Under these conditions, the constraint on the precise position in the genome of the conserved nucleotides of an interaction site is low. Indeed, when one conserved nucleotide of a given pattern is shifted from one base pair in the genome, chances are high that one of the probable conformations of the DNA molecule will place this nucleotide at the same spatial position compared to when it is not shifted in the genome and with another conformation of the DNA molecule. This is obviously true only if the shifts are not too important. This tends to confirm that class A flexible patterns define protein interaction sites. Indeed, we observed that from one occurrence to the next, the relative position of nucleotides composing them can vary from one to two base pairs. This is small enough so that there exists a likely conformation of the DNA molecule suitable to make it interact with its associated rigid protein. Alternatively, locally constrained DNA segments (for example through preexisting interaction with particular factors) might interact with proteins with flexible segments. Note that the absence of strong constraints on the position of the conserved nucleotides in class A flexible patterns is not easily compatible with other biological functions.

Fifth argument: the presence of optional peripheral repeats of Ts extending the skeleton at its two sides in class A flexible patterns (see Figure 2), can easily be accounted for under this DNA-protein interaction hypothesis. There are at least two ways to interpret the presence of the

peripheral repeats. A first idea is to suppose that they could be used by the DNA molecule to stabilize an interacting protein, as they appear on the same side of the DNA molecule as the rest of the conserved nucleotides of the pattern. These peripheral repeats would not be essential in the interaction, which would be possible only when the central part of class A flexible patterns is involved. A second idea is that the peripheral repeats of Ts in class A flexible patterns may help proteins slide along the DNA molecule in order to reach rapidly the central part of the patterns.

Now we may wonder which interacting proteins could be involved. Here is a few requirements that must be fulfilled by proteins to be good candidates according to the features of class A flexible patterns. First requirement: proteins have to be present in large enough amount in cells in order to be good candidates. Indeed, there are many interaction sites defined by the occurrences of class A flexible patterns in genomes. Alternatively, they may be involved in a dynamic process progressively threading the whole DNA molecule through a ratchet-like mechanism (for example forcing DNA segregation into daughter cells). Second requirement: proteins must not play a role exclusively in the transcription process as the pattern occurrences can be found inside coding regions as well as outside. Third requirement: the interaction sites of proteins with the DNA molecule must not be rigidly defined, as the sites we have uncovered in the present study have never been found previously. The fourth requirement that these proteins must fulfill is related to their presence in the organisms of interest. For each candidate protein, we checked whether its distribution in organisms matched the distribution of class A flexible patterns presented in Table 1. Here are some example of plausible candidates: archaeal histones [32,33], histone-like proteins H-NS and IHF [34-40], two topoisomerases (the reverse gyrase and the topoisomerase IIB-VI) [41-44] and the SMC family of proteins [45-49].

Since the patterns are ideally shaped to display specific but labile interaction with proteins, and since they are densely present in genomes with no relationship to the position of genes, we propose that they may be involved in some biological function such as the shaping of the prokaryotic nucleoid or its segregation before cell division.

Class A flexible patterns could be recognized during homologous recombination

The widespread distribution of flexible patterns of class A along genomes is consistent with selection of the motifs through processes that are fairly ubiquitous and happen sufficiently often in the life of an organism to provide some selective advantage. Until now we have mostly considered structural or regulatory processes involving the DNA molecule as a whole. In the course of evolution the process of recombination plays an essential role as it both permits proof-reading and insertion or deletion of DNA segments. In prokaryotes, recombination involves the formation of long helical filaments of the RecA protein double-stranded DNA [50] and homologs exist in eukaryotes [51]. During the process of recombination, the DNA double helix is distorted, asking for a nucleation process of the first RecA proteins binding, making use of the flexibility of the DNA molecule. The class A flexible patterns, distributed throughout genomes, and insensitive to the origin of the DNA (regions of the genome which are from horizontal gene transfer descent are as likely to harbour the patterns as are the core regions), might play such role. Exchange of base pairs between segments undergoing recombination is essential for recognition of homology, and physical evidence indicates that such an exchange occurs early enough to mediate recognition at A:T base pairs [52]. The conserved skeleton of the class A flexible patterns would provide the required biochemical basis for the process.

Conclusion

In this article, the source of the ubiquitous bias of period 10–11 bp in genomes has been identified. It is generated by specific and ubiquitous sequences that we named "class A flexible patterns". These patterns are flexible patterns whose main property is to display 10 bp-11 bp periodic repeats of Ts. As the patterns are densely spread in genomes, their occurrences naturally generate the bias.

The patterns account for the second largest bias in the nucleotides distribution of prokaryotic genomes, second to the one generated by the use of genetic code in genes, hence their biological function has to be of an essential nature. We discussed what this function could be and suggested that class A flexible patterns could be defining a new category of protein-DNA interaction sites in genomes.

Methods

First we introduce the definition of a correlation function which is used throughout this article. Then we explain the theoretical basis of the program we designed to find the sequences responsible for short-range biases, its actual implementation and its controls.

The correlation function

Definition – a genome G

A genome G of length L_G is written $G = (x_i)_{i\in[1..L_G]}$ with $\forall i \in [1..L_G]$, $x_i \in \{A, T, G, C\}$. It is taken in the standard 5'-3' orientation.

Definition – a sub-genome S *extracted from a genome* G

Let $G = (x_i)_{i\in[1..L_G]}$ be a genome. A sub-genome S of length L_S extracted from G is a sub-series of G. We call $E_{sg}(G)$ the set of all the sub-genomes of G. Then, for $S \forall E_{sg}(G)$ composed of N_S nucleotides, $\exists\sigma$: $[1..N_S] \to [1..L_G]$ a strictly increasing function so as $S = (x_{\sigma(i)})_{i\in[1..N_s]}$.

Definition – a pattern m

A pattern m composed of N_m nucleotides and of length L_m is written $m = (x_i, p_i)_{i\in[1..N_m]}$; $N_m \geq 1$, $p_1 = 1$, $p_{N_m} = L_m$ with p a strictly increasing series and $\forall i \in [1..N_m]$, $x_i \in \{A, T, G, C\}$. We call $E_m(N, L)$ the set of patterns composed of exactly N nucleotides and with a length shorter or equal to L. We call $E_m = \bigcup_{N\in[1..+\infty]} E_m(N, +\infty)$.

Definition – an occurrence of a pattern m *in a genome* G

Let $m = (x_i, p_i)_{i\in[1..N_m]} \in E_m$ be a pattern composed of N_m nucleotides and of length L_m.

Given the sub-genome $S = (y_{\sigma(i)})_{i\in[1..N_s]} \in E_{sg}(G)$ composed of $N_s = N_m$ nucleotides, S is an occurrence of m in G if and only if $\forall i \in [1..N_m]$, $x_i = y_{\sigma(i)}$ and $p_i = \sigma(i) - \sigma(1) + p_1$. We call $E_{oc}(m, G)$ the set of the occurrences of m in G. $\#E_{oc}(m, G)$ is the number of occurrences of m in G and $\#E_{oc}(x, G)$ is the number of occurrences of the single nucleotide x in G.

Definition – the correlation function f *(*G*)*

Given a genome G, an order of correlation O_{cor} and a length for the computation of the correlation L_{ana}, we define the correlation function $f(G)$ on the space $E_m(O_{cor}, L_{ana})$: for $m = (x_i, p_i)_{i\in[1..O_{cor}]} \in E_m(O_{cor}, L_{ana})$,

$$f(G)(m) = \frac{\#E_{oc}(m, G)}{\#E_{oc}(x_1, G)}.$$

Our practical calculation of correlation functions is performed as follows: the function is represented by an array of size $4^{O_{cor}} \cdot \begin{pmatrix} L_{ana} \\ O_{cor} - 1 \end{pmatrix}$. For each nucleotide of G, the array cells of all the patterns composed of O_{cor} nucleotides included in the next L_{ana} bp are increased by one. The number of steps is then proportional to $L_G \cdot \begin{pmatrix} L_{ana} \\ O_{cor} - 1 \end{pmatrix}$.

The correlation functions of all prokaryotic and lower eukaryotic genomes reveal a strong statistical bias of period 3 bp due to the dense presence of genes in genomes [1]. This bias is of little interest as its source is known. In order to study the other biases in the present work, we always pre-treated the correlation functions so as to hide this trivial bias. This deconvolution step was performed by subtracting the correlation function of a model genome constructed so as to contain only the trivial bias. The concept of model genome has been developed in [53,54]. This is performed here as follows:

Definition – the model genome G_{mo}*(*G*)*

Let us write the genome G as a series of dihexanucleotides: $G = (H_{i,1}H_{i,2})_{i\in\left[1..\frac{L_G}{12}\right]}$ with $H = (x_1x_2x_3x_4x_5x_6)$ representing an hexanucleotide.

The model genome $G_{mo}(G)$ is a random genome built from G by following these probability rules:

$$G_{mo}(G) = (H_{i,\sigma(i)}H_{i,\overline{\sigma(i)}})_{i\in\left[1..\frac{L_G}{12}\right]} \text{ with } \begin{cases} P(\sigma(i) = 1, \overline{\sigma(i)} = 2) = \frac{1}{2} \\ P(\sigma(i) = 2, \overline{\sigma(i)} = 1) = \frac{1}{2} \end{cases}$$

Definition – the treated correlation function f_t *(*G*)*

$$f_t(G) = \overline{f(G) - f(G_{mo}(G))}^{G_{mo}(G)}$$

The upper line means that $f_t(G)$ is the average of correlation functions of several model genomes derived from the same genome G. The treated correlation function is an average of probabilistic functions. Practically, for genomes long enough, after averaging over a few model genomes (usually three) one gets a function that almost completely lost the effects of biases with very short ranges (inferior to 6 bp) and hence lost the effect of the 3 bp periodic bias due to the presence of the genes, but saved most of the effects of other kind of information included in genomes. In the Background section, on Figure 1, we plotted $f_t(G)$ restricted on the following set of patterns: $(A, A, 1, l)_{l\in[1..100]}$.

Definition – the complementary sub-genome $\bar{S}^G$ *of the sub-genome* S

Given a genome *G* and a sub-genome *S*, we define naturally $\bar{S}^G$ as the sub-genome of *G* which includes in the right order all the nucleotides of *G* which are not in *S*.

Definition – the model genome $G_{mo}(S)$ *for a sub-genome* S

Let $G = (x_i)_{i\in[1..L_G]}$ be a genome, $S = (x_{\sigma(i)})_{i\in[1..N_s]}$ be a sub-genome of *G*, $\bar{S} = (x_{\bar{\sigma}(i)})_{i\in[1..N_{\bar{S}}]}$ be its complementary sub-genome and $G_{mo}(G) = (\gamma_i)_{i\in[1..L_G]}$ be a model genome derived from *G*.

$G_{mo}(S) = (z_i)_{i\in[1..L_G]}$ is defined as : $\begin{matrix}\forall i \in [1..N_S], z_{\sigma(i)} = x_{\sigma(i)} \\ \forall i \in [1..N_{\bar{S}}], z_{\bar{\sigma}(i)} = \gamma_{\bar{\sigma}(i)}\end{matrix}$.

Definition – the treated correlation function of a sub-genome $f_t(S)$

$f_t(S) = f_t(G_{mo}(S))$

Notation – a pattern family M

A pattern family *M* is a finite set of patterns. It is noted $M = (m_i)_{i\in[1..N_m]}$, $\forall i \in [1..N_m]$, $m_i \in E_m$.

In the Results section, $G^+(G,M) = \bigcup_{m\in M, S\in E_{oc}(m,G)} S$ and $G^-(G,M) = \bigcup_{m\in M, S\in E_{oc}(m,G)} S$. On Figure 5, 6, we plotted two correlation functions of those two sub-genomes restricted on the following set of patterns: $(A, A, 1, l)_{l\in[1..100]}$.

The rationale of the program

Our goal was to determine which sequences of a given genome *G* account for the statistical bias of period 11 bp affecting the distribution of its nucleotides. We designed a program meant to find out which sequences were responsible for all short-range non-trivial biases present in a given genome *G*. Here, "non-trivial" means different from the bias of period 3 bp due to the presence of the genes in genomes. Since the bias of period 11 bp is indeed a short-range bias, the sequences of *G* generating the bias should be included in the sequences determined by the program. Assuming that the majority of significant statistical biases present in a genome *G* can be revealed by the correlation function of *G*, our program does not look directly for the sequences generating the short-range biases but, rather, identifies the sequences generating $f_t(G)$ for a given O_{cor} and L_{ana} (practically four nucleotides and thirty base-pairs). The treated correlation function of a genome that would be biased only by the genes structure is the null function. Our program stands on the approximated formula (1) that we are introducing now.

Definition – a special pattern family for the genome G

A pattern family *M* will be called "special pattern family" if $(E_{oc}(m, G))_{m\in M}$ covers exactly, with no overlapping, the sequences of *G* that generates $f_t(G)$ for a given O_{cor} and L_{ana} and if the positions of the occurrences of the different patterns of *M* are not correlated. These conditions are written:

$f_t(G_{mo}((E_{oc}(m,G))_{m\in M})) = f_t(G)$ and $f_t((\overline{E_{oc}(m,G)})_{m\in M}) = \tilde{0}$

$$\sum_{m\in M}\sum_{S\in E_{oc}(m,G)} N_m = N_{(E_{oc}(m,G))_{m\in M}}$$

$$f_t\left(\bigcup_{m\in M, S\in E_{oc}(m,G)} S\right) = \sum_{m\in M} f_t\left(\bigcup_{S\in E_{oc}(m,G)} S\right)$$

We call $E_{spe}(G)$ the set of all special pattern families of *G*.

Assuming that such families containing only short enough patterns (shorter than one hundred base-pairs) exist, the aim of our program was to determine one of them.

Definition – the simulated genome G_{sim} *(G, m, β)*

For a given pattern *m*, let $G_{sim}(G, m, \beta)$ be the simulated genome derived from a genome *G* and constructed by repeatedly overwriting the pattern *m* on the original sequence of *G* (with a frequency β). We call $E_{ocin}(m, G_{sim}(G, m, \beta))$ the set of all the occurrences of *m* artificially introduced in $G_{sim}(G, m, \beta)$.

Property – for $M \in E_{spe}(G)$ **and** $\frac{1}{L_G} << \beta < \frac{1}{\max((N_m)_{m\in M}, L_{ana})}$,

$$f_t(G) \approx \sum_{m\in M} \frac{\#E_{oc}(m,G)}{\beta \cdot L_G} \cdot f_t(G_{sim}(m,G,\beta)). \qquad (1)$$

Indeed, we have $f_t(G) = f_s\left(\bigcup_{m\in M, S\in E_{oc}(m,G)} S\right)$. As $\sum_{m\in M}\sum_{S\in E_{oc}(m,G)} N_m = N_{(E_{oc}(m,G))_{m\in M}}$ $f_t\left(\bigcup_{m\in M, S\in E_{oc}(m,G)} S\right) = \sum_{m\in M} f_t\left(\bigcup_{S\in E_{oc}(m,G)} S\right)$, we have $f_t(G) = \sum_{m\in M} f_t\left(\bigcup_{S\in E_{oc}(m,G)} S\right)$..

Considering the way we derived the simulated genomes, it is obvious that the occurrences of the patterns m introduced in $G_{sim}(m, G, \beta)$ are not correlated to neighboring sequences. We then assume that natural occurrences of m in G are not too much correlated to neighboring sequences. Hence one gets:

$$\forall m \in M,\ f_t\left(\bigcup_{S\in E_{oc}(m,G)} S\right) \approx \frac{\#E_{oc}(m,G)}{\beta \cdot L_G} \cdot f_t\left(\bigcup_{S\in E_{ocin}(m,G_{sim})} S\right)$$

As we introduced the occurrences of the pattern m in a non-correlated manner in G_{sim} (m, G, β), it results that

$$f_t(G_{sim}(m,G,\beta)) \approx f_t\left(\bigcup_{S\in E_{oc}(m,G_{sim})} S\right) + f_t\left(\overline{\bigcup_{S\in E_{oc}(m,G_{sim})} S}^{G_{sim}}\right)$$

.

We have $f_t\left(\overline{\bigcup_{S\in E_{oc}(m,G_{sim})} S}^{G_{sim}}\right) << f_t\left(\bigcup_{S\in E_{oc}(m,G_{sim})} S\right)$ because many occurrences of the pattern m have been introduced in G_{sim} (m, G, β), generating very strong correlations. Hence $f_t(G_{sim}(m,G,\beta)) \approx f_t\left(\bigcup_{S\in E_{oc}(m,G_{sim})} S\right)$. As many more occurrences of the pattern m were introduced in G_{sim} (m, G, β) than there are naturally in G, one has $f_t(G_{sim}(m,G,\beta)) \approx f_t\left(\bigcup_{S\in E_{ocin}(m,G_{sim})} S\right)$. Finally it comes that $f_t(G) \approx \sum_{m\in M_e} \frac{\#E_{oc}(m,G)}{\beta \cdot L_G} \cdot f_t(G_{sim}(m,G,\beta))$.

Hence the treated correlation function of G can be approximated by a linear combination of the correlation functions of the simulated genomes associated to the patterns belonging to a special pattern family. This property gave us a theoretical framework to determine such a special pattern family.

Definition – a positively free family

Let E be a vectorial space and F a family of vectors.

Let us define $Pos(F) = \left\{ \sum_{\vec{h}\in F} \alpha_{\vec{h}} \vec{h} \,/\, \forall \vec{h} \in F, \alpha_{\vec{h}} \geq 0 \right\}$. The family $F = \{\vec{u}_i\}_{i\in[1..n]} \in E^n$ is positively free in E if and only if $\forall \vec{a} \in Pos(F), \exists!(\alpha_i)_{i\in[1..n]} \in \Re^{+^n} / \vec{a} = \sum_{i\in[1..n]} \alpha_i \vec{u}_i$

Our idea was to choose a pattern family M_{input} containing as many patterns as possible that is positively free. If there exists one and only one special pattern family M_{spe} included in M_{input}, then there exists a linear decomposition of $f_t(G)$ on the $(f_t(G_{sim}(m,G,\beta)))_{m\in M_{input}}$ with positive coefficients (for any β so as $\frac{1}{L_G} << \beta < \frac{1}{\max((N_m)_{m\in M_{input}}, L_{ana})}$), i.e. $f_t(G) \approx \sum_{m\in M_{spe}} \frac{\#E_{oc}(m,G)}{\beta \cdot N_G} \cdot f_t(G_{sim}(m,G,\beta))$. As this decomposition is unique, by calculating the decomposition of $f_t(G)$ on the $(f_t(G_{sim}(m,G,\beta)))_{m\in M_{input}}$, one can determine which patterns belong to M_{spe}. Hence basically our program, for an input of a genome G and a pattern family M_{input}, chose a suitable β, calculated the $(f_t(G_{sim}(m,G,\beta)))_{m\in M_{input}}$ and the unique decomposition with positive coefficients of $f_t(G)$ on these functions. It gave as an output a pattern family M_{output} which consisted in the patterns of M_{input} for which the treated correlation functions of the associated simulated genomes are involved.

Practical implementation of the program

First of all, we assumed that, for $O_{cor} = 4$ and $L_{ana} = 30$ *bp*, there exist $N > 0$ and $L > 0$ so that there exists one and only one special pattern family included in E (N, L).

Because of computational time limitation, only input pattern families that are not containing too many patterns (less than one thousand patterns) could be tested. To extend the output possibilities of the program, we ran it in a few steps, at the cost of further approximations. First, we entered $M_0 = E(2,14) \cup E(3,14)$ as an input family (this family is positively free). As we did not expect any special pattern family to belong to M_0, we did not calculate the decomposition of $f_t(G)$ on $(f_t(G_{sim}(m,G,\beta)))_{m\in M_0}$, but

rather a "positive projection" of $f_t(G)$ on $(f_t(G_{sim}(m,G,\beta)))_{m\in M_0}$.

Definition – $\vec{p}^{+}(\vec{a},F)$ *the positive projection of* $\vec{a}\in E$ *a vectorial space of finite dimension in the non-void family* $F=\{\vec{u}_i\}_{i\in[1..n]}\in E^n$

Be < > a scalar product in E and $||\ ||$ the associated norm. It is possible to prove that $\exists!\vec{p}\in Pos(F)$ so as $\forall\vec{s}\in Pos(F)$, $\|\vec{s}-\vec{a}\|\geq\|\vec{p}-\vec{a}\|$. We call this vector $\vec{p}^{+}(\vec{a},F)$, the positive projection of $\vec{a}$ in F.

We calculated $f_1(G)=\vec{p}^{+}(f_t(G),(f_t(G_{sim}(m,G,\beta)))_{m\in M_0})$. The coefficients of this positive projection can be assimilated to a frequency of patterns present in G, expressed in bp^{-1}. Then we constructed M_1 the output pattern family with all the patterns of M_0 for which the coefficient of the treated function of the associated simulated genome is large enough. The selectivity of the program is adjustable at this level. Practically, we kept the patterns for which the coefficients are above $\frac{1}{6000}bp^{-1}$, with an average approximately $\frac{1}{2000}bp^{-1}$, which makes usually approximately twenty patterns. This is a first approximation in our program. As a second step, we used M_2 as an input pattern family. M_2 is containing M_1 plus all the patterns that can be built by extending the patterns of M_1 with one extra nucleotide. The added nucleotide can be placed at any position inside the original patterns or at their sides (as far as 15 bp from the extremities of the original patterns). Again, we calculated a positive projection and got a resulting pattern family M_3.

We repeated this step as long as we got patterns that were strictly included in $\bigcup_{i\in[2..6]}E(i,30)$ (i.e. all the patterns that are composed of up to six nucleotides and span less than 30 bp). We got usually close to one hundred patterns in this pattern family. Let us call M_{final} this resulting pattern family. It is an approximation of M_{spe}. Then, by merging the patterns (composed of six nucleotides) that could be identify as subsets of a same longer pattern (composed of more than six nucleotides), we obtained patterns that belonged to $\bigcup_{i\in[2..10]}E(i,60)$ while becoming closer to M_{spe}. Finally, from the patterns contained in M_{final}, we could define approximately twenty flexible patterns per organisms (see the Results section).

Besides the approximation generated by the division of the program into a few steps, a few more approximations were introduced during that process. First, the calculation of the positive projection was performed approximately so as to save calculation time. Second, the correlation functions were calculated on restricted sets, practically on $E(4,30)$, i.e. O_{cor} = 4 and L_{ana} = 30 *bp*. This made the description of patterns approximate since we aimed at determining patterns containing more than four nucleotides. The correlation order should be longer than the maximum number of nucleotides we want to find in patterns, otherwise the program may find patterns which are actually artefacts (a mix of genuine patterns present at distinct locations in the genome).

The program was written in C code. Built and operated in this way, the program was run on a genome of 2 Mbp in 3 weeks with a 1.8 Ghz G5 CPU. The most time-consuming step is the calculation of the correlation functions with O_{cor} = 4 and L_{ana} = 30 *bp*.

Controls of the program

Different controls were performed to test the selectivity of the program. First, when run on completely random genomes, the coefficients of the first positive projection were below the threshold, so that the resulting pattern family was empty. Second, the program was also tested with artificial genomes built from completely random genomes in which we introduced a given pattern at random locations. The program proved able to extract the pattern back provided that the pattern frequency of introduction was above $\frac{1}{3000}bp^{-1}$. Third, the program proved able to identify already known rigid patterns in genomes (see the Results section).

Authors' contributions

EL designed the algorithm and performed the bulk of the outlined study. AD proposed the rationale for the study and outlined its biological implications. Both authors participated in the writing of this article.

Acknowledgements

This work was supported by the BIOSUPPORT program of the Innovation and Technology Fund (ITF) of the Hong Kong's government. We are grateful to Benoît Arcangioli for his suggestion regarding the RecA recognition hypothesis.

References

1. Herzel H, Weiss O, Trifonov EN: **10–11 bp periodicities in complete genomes reflect protein structure and DNA folding.** *Bioinformatics* 1999, **15:**187-193.
2. Fukushima A, Ikemura T, Oshima T, Mori H: **Detection of period in eukaryotic genomes on the basis of power spectrum analysis.** *Genome Inform Ser Workshop Genome Inform* 2002, **13:**21-29.
3. Li W, Stolovitzky G, Bernaola-Galvan P, Olivier JL: **Compositional heterogeneity within, and uniformly between, DNA sequences of yeast chromosomes.** *Genome Res* 1998, **8:**916-918.
4. Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL: **Mosaic organization of DNA nucleotides.** *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 1994, **49:**1685-1689.
5. Herzel H, Ebeling W, Schmitt AO: **Entropies of biosequences: the role of repeats.** *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 1994, **50:**5061-5071.
6. Schmitt AO, Herzel H: **Estimating the entropy of DNA sequences.** *J Theor Biol* 1997, **188:**369-377.
7. Yu ZG, Anh VV, Wang B: **Correlation property of the length sequences based on global structure of the complete genome.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2001, **63:**011903.
8. Buldyrev SV, Goldberder AL, Havlin S, Mantegna RN, Matsa ME, Peng CK, Simons M, Stanley HE: **Long-range correlation properties of coding and non-coding DNA sequences: GenBank analysis.** *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 1995, **51:**5084-5091.
9. Audit B, Vaillant C, Arneodo A, d'Aubenton-carafa Y, Thermes C: **Long-range correlations between sites: relation to the structure and dynamics of nucleosomes.** *J Mol Biol* 2002, **316:**903-918.
10. Grosse I, Herzel H, Buldyrev SV, Stanley HE: **Species independence of mutual information in coding and non-coding DNA.** *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 2000, **61:**5624-5629.
11. Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng CK, Simons M, Stanley HE: **Linguistic features of non-coding DNA sequences.** *Phys Rev Lett* 1994, **73:**3169-3172.
12. Shepherd JC: **Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification.** *Proc Natl Acad Sci USA* 1981, **78:**1596-600.
13. Shepherd JC: **Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code.** *J Mol Evol* 1981, **17:**94-102.
14. Staden R: **Finding protein coding regions in genomic sequences.** *Methods Enzymol* 1990, **183:**163-180.
15. Tsonis AA, Elsner JB, Tsonis PA: **Periodicity in DNA coding sequences: implications in gene evolution.** *J Theor Biol* 1991, **151:**323-331.
16. Gutierrez G, Oliver JL, Marin A: **On the origin of the periodicity of three in protein coding DNA sequences.** *J Theor Biol* 1994, **167:**413-414.
17. Fukushima A, Ikemura T, Kinouchi M, Oshima T, Kudo Y, Mori H, Kanaya S: **Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis.** *Gene* 2002, **300:**203-211.
18. Fickett JW, Tung CS: **Assessment of protein coding measures.** *Nucleic Acids Res* 1992, **20:**6441-6450.
19. Schieg P, Herzel H: **Periodicities of 10–11 bp as indicators of the supercoiled state of Genomic DNA.** *J Mol Biol* 2004, **343:**891-901.
20. Espéli O, Moulin L, Boccard F: **Transcription attenuation associated with bacterial repetitive extragenic BIME elements.** *J Mol Biol* 2001, **314:**375-386.
21. Makarova KS, Aravind L, Wolf YI, Tatusov RL, Minton KW, Koonin EV, Daly MJ: **Genome of the Extremely Radiation-Resistant Bacterium *Deinococcus radiodurans* Viewed from the Perspective of Comparative Genomics.** *Microbiol Mol Biol Rev* 2001, **65:**44-79.
22. Bentley SD, Maiwald M, Murphy LD, Pallen MJ, Yeats CA, Dover LG, Norberczack HT, Besra GS, Quail MA, Harris DE, von Herbay A, Goble A, Rutter S, Squares R, Barell BG, Parkhill J, Relman DA: **Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*.** *The Lancet* 2003, **361:**637-644.
23. Fitz-Gibbon ST, Ladner H, Kim UJ, Stetter KO, Simon MI, Miller JH: **Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*.** *Proc Natl Acad Sci USA* 2002, **99:**984-989.
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242.
25. Mirny LA, Gelfand MS: **Structural analysis of conserved base pairs in protein-DNA complexes.** *Nucleic Acids Research* 2002, **30:**1704-1711.
26. Strick TR, Allemand JF, Bensimon D, Croquette V: **Behavior of supercoiled DNA.** *Biophys J* 1998, **74:**2016-2028.
27. Dervan PB, Burli RW: **Sequence-specific DNA recognition by polyamides.** *Curr Opin Chem Biol* 1999, **3:**688-693.
28. Moravek Z, Neidle S, Schneider B: **Protein and drig interactions in the minor groove of DNA.** *Nucleic Acids Res* 2002, **30:**1182-1191.
29. O'flanagan RA, Paillard G, Lavery R, Sengupta AM: **Non-additivity in protein-DNA binding.** *Bioinformatics* 2005 in press.
30. Travers A, Muskhelishvili G: **DNA supercoiling – a global transcriptional regulator for enterobacterial growth ?** *Nat Rev Microbiol* 2005, **3:**157-169.
31. Stefl R, Wu H, Ravindranathan S, Sklenar V, Feigon J: **DNA A-tract bending in three dimensions: solving the dA4T4 vs. dT4A4 conundrum.** *Proc Natl Acad Sci USA* 2004, **101:**1177-1182.
32. Malik HS, Henikoff S: **Phylogenomics of the nucleosome.** *Nat Struct Biol* 2003, **10:**882-890.
33. Pavlov NA, Cherny DI, Jovin TM, Slesarev AI: **Nucleosome-like complex of the histone from the hyperthermophile Methanopyrus kandleri (MkaH) with linear DNA.** *J Biomol Struct Dyn* 2002, **20:**207-214.
34. Nishino K, Yamaguchi A: **Role of Histone-Like Protein H_NS in Multidrug Resistance of *Escherichia coli*.** *J Bacteriol* 2004, **186:**1423-1429.
35. Rouquette C, Serre MC, Lane D: **Protective role for H_NS protein in IS1 transposition.** *J Bacteriol* 2004, **186:**2091-2098.
36. Rimsky S: **Structure of the histone-like protein H-NS and its role in regulation and genome superstructure.** *Curr Opin Microbiol* 2004, **7:**109-114.
37. Tendeng C, Bertin PN: **H-NS in Gram-negative bacteria: a family of multifaceted proteins.** *Trends Microbiol* 2003, **11:**511-517.
38. Murtin C, Engelhorn M, Geiselmann J, Boccard F: **A quantitative UV laser footprinting analysis of the interaction of IHF with specific binding sites: re-evaluation of the effective concentration of IHF in the cell.** *J Mol Biol* 1998, **284:**949-961.
39. Lynch TW, Read EK, Mattis AN, Gardner JF, Rice PA: **Integration host factor: putting a twist on protein-DNA recognition.** *J Mol Biol* 2003, **330:**493-502.
40. Swinger KK, Rice PA: **IHF and HU: flexible architects of bent DNA.** *Curr Opin Struct Biol* 2004, **14:**28-35.
41. Champoux JJ: **DNA topoisomerases: structure, function, and mechanism.** *Annu Rev Biochem* 2001, **70:**369-413.
42. Massé E, Drolet M: **Relaxation of Transcription-induced Negative Supercoiling Is an Essential Function of *Escherichia coli* DNA Topoisomerase I.** *J Biol Chem* 1999, **274:**16654-16658.
43. Massé E, Drolet M: ***Escherichia coli* DNA Topoisomerase I Inhibits R-loop Formation by Relaxing Transcription-induced Negative Supercoiling.** *J Biol Chem* 1999, **274:**16659-16664.
44. Bouthier de la Tour C, Portemer C, Nadal M, Stetter KO, Forterre P, Duguet M: **Reverse Gyrase, a Hallmark of the Hyperthermophilic Archaebacteria.** *J Bacteriol* 1990, **172:**6803-6808.
45. Cobbe N, Heck MM: **Review: SMCs in the world of chromosome biology- from prokaryotes to higher eukaryotes.** *J Struct Biol* 2000, **129:**123-143.
46. Cobbe N, Heck MM: **The evolution of SMC proteins: phylogenetic analysis and structural implications.** *Mol Biol Evol* 2004, **21:**332-347.
47. Melby TE, Ciampaglio CN, Briscoe G, Erickson HP: **The symmetrical structure of structural maintenance of chromosomes (SMC) and MukB proteins: long, antiparallel coiled coils, folded at a flexible hinge.** *J Cell Biol* 1998, **142:**1595-1604.
48. Löwe J, Cordell SC, Van den Ent F: **Crystal structure of the SMC head domain: An ABC ATPase with 900 residues antiparallel coiledcoil inserted.** *J Mol Biol* 2001, **306:**25-35.

49. Haering CH, Löwe J, Hochwagen A, Nasmyth K: **Molecular architecture of SMC proteins and the yeast cohesin complex.** *Mol Cell* 2002, **9:**773-788.
50. Prevost C, Takahashi M: **Geometry of the DNA strands whithin the RecA nucleofilaments: role in homologous recombination.** *Q Rev Biophys* 2003, **36:**429-453.
51. Krogh BO, Symington LS: **Recombination proteins in yeast.** *Annu Rev Genet* 2004, **38:**233-271.
52. Gupta RC, Folta-Stogniew E, O'Malley S, Takahashi M, Radding CM: **Rapid exchange of A:T base pairs is essential for recognition of DNA homology by human Rad51 recombination protein.** *Mol Cell* 1999, **4:**705-714.
53. Karlin S, Cardon LR: **Computational DNA sequence analysis.** *Annu Rev Microbiol* 1994, **48:**619-654.
54. Hénaut A, Lisacek F, Nitschké P, Moszer I, Danchin A: **Global analysis of genomic texts: the distribution of AGCT tetranucleotides in the Escherichia coli and Bacillus subtilis genomes predicts translational frameshifting and ribosomal hopping in several genes.** *Electrophoresis* 1998, **19:**515-527.