

From Genes to Genomes: Universal Scale-invariant Properties of Microbial Chromosome Organisation

Benjamin Audit and Christos A. Ouzounis*

Wellcome Trust Genome
Campus, Computational
Genomics Group, The European
Bioinformatics Institute
EMBL Cambridge Outstation
Cambridge CB10 1SD, UK

The availability of complete genome sequences for a large variety of organisms is a major advance in understanding genome structure and function. One attribute of genome structure is chromosome organisation in terms of gene localisation and orientation. For example, bacterial operons, i.e. clusters of co-oriented genes that form transcription units, enable functionally related genes to be expressed simultaneously. The description of genome organisation was pioneered with the study of the distribution of genes of the *Escherichia coli* partial genetic map before the full genome sequence was known. Deploying powerful techniques from circular statistics and signal processing, we revisit the issue of gene localisation and orientation using 89 complete microbial chromosomes from the eubacterial and archaeal domains. We demonstrate that there is no characteristic size pertinent to the description of chromosome structure, e.g. there does not exist any single length appropriate to describe gene clustering. Our results show that, for all 89 chromosomes, gene positions and gene orientations share a common form of scale-invariant correlations known as “long-range correlations” that we can reveal for distances from the gene length, up to the chromosome size. This observation indicates that genes tend to assemble and to co-orient over any scale of observation greater than a few kilobases. This unexpected property of chromosome structure can be portrayed as an operon-like organisation at all scales and implies that a complete scale range extending over more than three orders of magnitudes of chromosome segment lengths is necessary to properly describe prokaryotic genome organisation. We propose that this pattern results from the effects of the superhelical context on gene expression coupled with the structure and dynamics of the nucleoid, possibly accommodating the diverse gene expression profiles needed during the different stages of cellular life.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: gene position and orientation; transcriptional context; nucleoid structure; circular statistics; long-range correlations

*Corresponding author

Introduction

Recent advances in sequencing technology have generated a multitude of entire genome sequences from a variety of species. This genomic information has inspired novel approaches in computational research by enabling the exploitation of genomic “context” for the prediction of gene function.^{1–3} Contextual approaches take advantage of the inherent properties of genome structure. For example, having

complete genome sequences enables us to delimit the full repertoire of genes available to living organisms, a fact allowing the detection of patterns of presence/absence of proteins in a particular species. Such patterns are very informative and have been exploited in relation to phylogeny^{4,5} or metabolic and regulatory pathways^{6,7} to predict functional associations between proteins. Other important contextual clues are gene location and orientation. For instance, comparative methods have been developed to predict protein functions based on the conservation of clusters of localised genes across large evolutionary distances.^{8–11} Furthermore, these methods have been used to unravel the nature of the dynamical events responsible for

Supplementary data associated with this paper can be found at doi: 10.1016/S0022-2836(03)00811-8

E-mail address of the corresponding author:
ouzounis@ebi.ac.uk

genome shuffling based on the location of orthologous genes identified in pairs of closely related genomes.¹² Information on gene location and orientation can also be used within a single genome to characterize and predict operons.^{13,14}

The above examples illustrate that gene context, in general, can convey meaningful patterns about the function and evolution of genes. Surprisingly, the huge increase in genomic data of the past decade did not raise similar interest for the characterization of global patterns in chromosome organisation. However, such characterization could ultimately lead to a better understanding of the rules governing genome evolution and cellular processes. For example, analyses of gene location and orientation for a number of microbial species show an asymmetry between the leading and the lagging strands, the former being more gene-rich than the latter.^{15–19} It has been argued that this strand asymmetry minimises frontal collisions between the DNA–polymerase and the RNA–polymerase complexes during replication.¹⁵ More recently, it has been shown that the genomes with a *polC* orthologue display stronger strand asymmetries than the other genomes, therefore suggesting a direct relationship between the molecular details of replication and global strand asymmetry.¹⁹ Also, fluorescence microscopy experiments of *Bacillus subtilis* cells suggest that strand orientation asymmetries could help to drive chromosome segregation.²⁰

The lack of large regions of colinearity between distant genomes suggests that the overall gene order is not conserved during bacterial evolution.^{8–10,21} Yet, early work on the *Escherichia coli* genetic map suggested that gene location is not random, e.g. some gene clusters and axes of symmetry have been delineated and interpreted in terms of nucleoid structure and functional domains.^{15,22–24} Some of these findings were criticized²⁵ because these symmetries and the clustering of genes could be fully explained by the log-normal distribution of gene density, determined at a 0.5 minute resolution, i.e. when dividing the *E. coli* genetic map into 200 intervals.

The large number of completely sequenced prokaryotes enables us to revisit the issue of gene distribution with much more data and make the analysis statistically meaningful and with a much broader phylogenetic perspective. We present a global characterization of gene location and orientation for 89 complete microbial chromosomes from 71 Bacteria and 15 Archaea. To perform this analysis we introduce two well-established techniques from statistics and signal processing for the analysis of genome organisation. As the vast majority of the prokaryote chromosomes have a circular structure (86/89), we analysed gene locations as points on a circular map using circular statistics.^{26–28} In a second step, we get a complementary point of view on genome organisation by using wavelet analysis²⁹ which was successfully applied to uncover long-range correlations in DNA sequences^{30,31} in relation to the nucleosomal

structure.^{32,33} Our goal is to describe the salient statistically significant patterns of genome organisation derived by these methods. In principle, such patterns can span different scales, ranging from the gene length (e.g. operon) up to the chromosome scale (e.g. strand asymmetry). In this regard, it is important to note that the two techniques mentioned above enable the dissociation of the different scales involved.

Herein, using a spectral interpretation of circular statistics, we are able to show that no preferred size appears while analysing gene location, thus suggesting that gene clusters in prokaryotic chromosomes are not an important organisational pattern at any particular scale. Note that this result does not necessarily imply that functional domains are absent. Instead, the full spectral analysis demonstrates that gene locations are correlated from the gene length to the chromosome length, in a scale-independent manner demonstrating an unexpected level of gene clustering at any scale in this complete scale range. We then perform wavelet analysis of gene density and of gene orientation asymmetry distributions. We thus confirm the existence of long-range correlations of the gene density distribution and detect the same correlation structure for the gene orientation asymmetry distribution showing that there also exists an unexpectedly high level of co-orientation at any scale from a few kilobases to chromosome length. Together these results suggest that an operon-like organisation is necessary to describe genome organisation at all scales. Since a similar type of scale-invariant organisation is observed for gene positions and gene orientations, it is likely that the forces responsible for this pattern of organisation acts on gene location and gene orientation simultaneously and that some global functional constraints could be at play, for example transcription. Indeed, transcription has a strong effect on the superhelical state of DNA and in return the superhelical context of a gene is very likely to have a major effect on gene expression.³⁴ Therefore, gene location and orientation could be important attributes for genome function and evolution. Finally, it is noteworthy that these results appear to be universal among prokaryotes, as they are detectable across various species from different phylogenetic domains or taxonomic groups.

Methodological Approach

In the work presented here, we study the distribution of gene position and orientation along complete prokaryotic chromosomes. Each gene is associated with two attributes only: its middle position x that is used as position marker and its orientation. Other attributes such as functional annotations are not taken into account.

Complete genome sequences

The present work was performed using 89

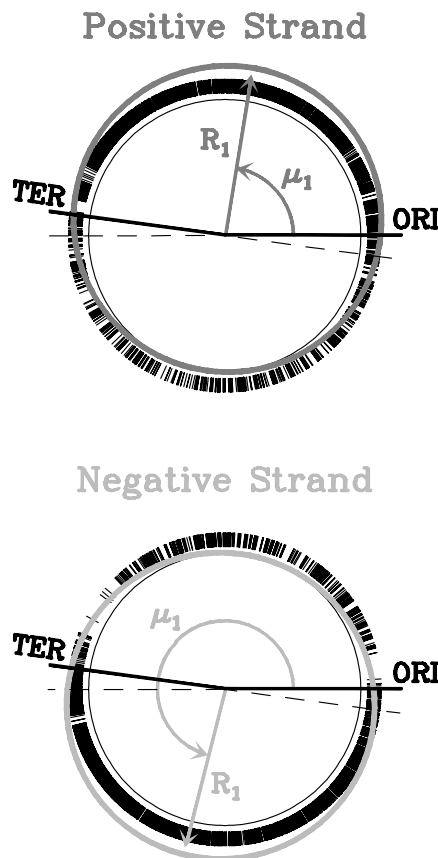


Figure 1. Application of circular statistics to the analysis of the gene distribution asymmetry between the leading and the lagging strands in *Bacillus subtilis*. Genes on the positive strand (top) and genes on the negative strand (bottom) are analysed separately. Each gene is represented by a black bar above its position. The sense of transcription is clockwise for negative strand genes and anti-clockwise for positive strand genes. For the two gene position datasets, we represent the trigonometric means \bar{m}_1 as arrows of arbitrary length pointing in the direction given by the angle $\bar{\mu}_1$. The thick lines in gray represent the description of gene position data captured by \bar{m}_1 (see Supplementary Material). The location of the origin (ORI) and termination (TER) of replication are also indicated. Note that in both cases, $\bar{\mu}_1$ correctly points to the middle of the leading strand without using any *a priori* knowledge of the location of the ORI or TER. Using equation (2) and the values for \bar{R}_1 , we confirm that this asymmetry pattern that can be detected by eye, is highly meaningful: the significance level are $P_- = 1.9 \times 10^{-89}$ and $P_+ = 1.6 \times 10^{-77}$ for the negative and positive strand genes, respectively.

completely sequenced chromosomes from the eubacterial and archaeal domains. Chromosomes have been classified into five categories according to their topology (circular or linear) and their taxonomic classifications: class 1 contains 15 circular archaeal chromosomes, class 2 contains 25 circular firmicutes chromosomes, class 3 contains 31 circular proteobacterial chromosomes, class 4 contains 15 other circular eubacterial chromosomes and

class 5 contains three linear eubacterial chromosomes. A complete list of the species analysed can be found in the Supplementary Material.

Spectral analysis using circular statistics

Gene position data for circular chromosomes have a directional nature so that standard statistical tools are not well suited for their analysis. Instead, we use circular statistics, which is a framework that fully takes into account the periodic nature of circular data^{26–28} (for more details, see Supplementary Material).

To apply this framework to gene location data, the position marker x of each gene is mapped to a polar angle $\theta = 2\pi(x/s)$, where s is the genome size. Such a mapping is fully meaningful for circular chromosomes. In the case of linear topology, this amounts to assuming periodicity for the positional data. A set of genes belonging to a given chromosome is then described by their trigonometric moments \bar{m}_p that are indexed by the positive integer p called the order. Note that the trigonometric moments can be viewed as the equivalent of the moments used in standard statistics. For example, the trigonometric mean \bar{m}_1 corresponds to the commonly used arithmetic mean. For each chromosome, we analysed the position distribution of the complete set of genes G and of the sets of genes on the positive strand G_+ or the negative strand G_- .

In this work, we mainly use circular statistics as a spectral analysis technique. Indeed, the trigonometric moment of order p is an estimate of the p th coefficient of the Fourier decomposition of the gene position distribution (see Supplementary Material). In other words, the order p can be interpreted as a frequency parameter corresponding to a period s/p in DNA physical length unit and \bar{m}_p describes features of size s/p for the localisation data analysed. For example, an asymmetry in gene distribution between the leading and the lagging strands results in one-half of the chromosome being more gene-dense than the second-half for the gene sets G_+ and G_- (Figure 1). This pattern has a size $s = s/1$ (its period is the chromosome size) and will therefore be captured by the trigonometric moment of order $p = 1$ when analysing G_+ or G_- . More precisely, the trigonometric moments are complex numbers and can be written as $\bar{m}_p = \bar{R}_p e^{ip\bar{\mu}_p}$, where \bar{R}_p is the p th resultant length and $\bar{\mu}_p$ the p th direction. The angle $\bar{\mu}_p$ gives a positional information, whereas the real value \bar{R}_p ($\in [0,1]$) quantifies the strength of the patterns of size s/p to the description of the localisation data under study. In the case of gene asymmetry between the leading and lagging strands, the analysis of G_+ and G_- leads to $\bar{\mu}_1$ pointing to the leading strand (the half of the genome the more gene-dense) and \bar{R}_1 measuring the level of asymmetry (Figure 1).

Here, we will only use the resultant lengths to assess which trigonometric moments are relevant

to the description of gene position data, i.e. which pattern sizes are meaningful to chromosome structure. As it is common practice in spectral analysis, we analyse an entire set of possible pattern sizes s/p at the same time using the spectrum of the data in which \bar{R}_p^2 is plotted as a function of p or $\nu = p/s$, where ν is the frequency expressed in units of inverse DNA physical length. In this context, the properties of the trigonometric moments of the uniform distribution (random uncorrelated distribution) are of particular importance, as they provide us with a natural point of comparison. For n data points uniformly distributed, it can be shown (see Supplementary Material) that for n large enough $-n > 30$ is sufficient for most applications,²⁶ $n\bar{R}_p^2$ has an exponential distribution of parameter 1. As a consequence, the spectrum of the uniform distribution is flat (does not depend on p):

$$E(n\bar{R}_p^2) = 1 \quad \text{for all } p \quad (1)$$

where E stands for the mathematical expectation. Then, it is easy to build a significance test for \bar{R}_p . The probability of rejecting the null hypothesis that the n data points were drawn from a uniform distribution is given by:

$$P_u \approx 1 - \exp(-n\bar{R}_p^2) \quad (2)$$

In other words, we can say that the p th trigonometric moment is meaningful, i.e. the pattern size s/p is relevant to chromosome structure, at a confidence level $\exp(-n\bar{R}_p^2)$.

The spectral analysis based on circular statistics we develop here for gene position data consists of computing, for each chromosome, \bar{R}_p for the three position datasets G , G_+ and G_- for $p = 1$ to $p = s/100$ bp. By plotting the corresponding spectra we can then assess which scales from 100 bp up to the chromosome length are relevant for genome structure organisation.

Gene orientation asymmetry and gene density

Circular statistics is a very powerful framework for detecting hidden regularities in circular datasets.²⁸ For a given chromosome, it enables the analysis of the complete set of genes G , as well as the sets of genes on the positive strand G_+ or the negative strand G_- . But, as G_+ and G_- are subsets of G , they strongly depend on G . Therefore, the analysis of G_+ and G_- does not only depend on gene orientation but also on gene location. For example, if the genes have a strong preference to locate close to the origin of replication, and if the gene orientations are equiprobable along the chromosome, then genes on the positive or negative strand will also present a preferred positioning close to the origin of replication. Thus, the spectral analysis of the two gene sets G_+ and G_- will reveal this genome asymmetry, which is a property of the gene location and not of the gene orientation.

To overcome this limitation and further analyse

gene orientation independently of gene location, it is useful to introduce the gene orientation asymmetry a_w alongside gene density d_w . In a window of size w , these quantities are defined as:

$$a_w = \frac{n_{w+} - n_{w-}}{n_w} \quad \text{and} \quad d_w = \frac{n_w}{w} \quad (3)$$

where n_w , n_{w+} and n_{w-} are the total number of genes, the number of genes on the positive strand and the number of genes on the negative strand within the window of size w , respectively. Note that a_w is set to 0 when $n_w = 0$. The asymmetry index a_w is related to the probability of a gene within a given window to have a specific orientation, e.g. the probability p_+ to be on the positive strand is $p_+ = (a_w + 1)/2$. This description of gene location and orientation has the desired property that in any window it is possible to define a_w and d_w independently. In the case of a preferred localisation of genes close to the origin of replication and of equiprobable orientations, gene density will be higher for windows close to the origin of replication but gene orientation asymmetry will not present any such trend. Thus, whereas gene position data can be analysed through gene density or the circular statistics framework, there is no straightforward way to analyse gene orientation data independently from position using spectral analysis and we have to rely on the characterisation of gene orientation asymmetry.

In order to set a comparison point when analysing gene orientation asymmetry and gene density, we first present expected properties of these quantities in the model case of a genome lacking any organisational pattern, i.e. in the case of n genes uniformly distributed on a chromosome of size s , n_+ (respectively $n_- = n - n_+$) genes having a positive (respectively negative) orientation. In other words, we model a uniform and uncorrelated chromosome organisation with an average gene density $d = n/s$ and an average gene orientation asymmetry $a = (n_+ - n_-)/n$. It is straightforward that each gene has a probability $p = w/s$ to belong to a given window of size w . Therefore; the counts n_w , n_{w+} and n_{w-} of genes in windows of size w exhibit binomial distribution $B(n, p)$, $B(n_+, p)$ and $B(n_-, p)$, respectively. We can then calculate the mean m_u and standard deviation σ_u for a_w and d_w in this uniform model:

$$m_u(a_w) \approx a \quad \text{and} \quad \sigma_u^2(a_w) \approx (1 - w/s) \frac{1 - a^2}{wd} \quad (4)$$

$$m_u(d_w) = d \quad \text{and} \quad \sigma_u^2(d_w) = (1 - w/s) \frac{d}{w} \quad (5)$$

The formulae (4) are approximations when the average count number in a window of size w is large compared to 1 ($wd \gg 1$). We observe that a_w and d_w are unbiased estimates of a and d . Importantly, we also notice that both standard deviations are affected by a finite size effect (they are proportional to $1 - w/s$), which is a consequence of

the fixed average parameter values, namely a ($= a_s$) and d ($= d_s$). This is exactly the situation encountered in the analysis of a given chromosome. Accordingly, we always limit the window size w such that $w/s \leq 6$, which corresponds to a weakening of the standard deviation of at most 10%. Neglecting the finite size effect, equations (4) and (5) indicate that $\sigma_u(a_w)$ and $\sigma_u(d_w)$ share a common power law behaviour:

$$\sigma_u(w) = \sigma_u(1)/\sqrt{w} = \sigma_u(1)w^{1/2-1} \quad (6)$$

Equation (6) is a result of the absence of correlation in the position/orientation distribution we have modelled here. It can be used to characterize the lack of genome organization by simply plotting the standard deviation as a function of the window size w and checking if equation (6) holds. This approach actually conveys some specific information that will be discussed below.

Characterising scale-invariant genome organisation

Standard deviation analysis

The power law behaviour of the standard deviation with the scale of analysis w is the hallmark of scale-invariant models in general. An important distinctive attribute within this class of models is the Hurst exponent H defined as follows:^{35–37}

$$\sigma_H(w) = \sigma_H(1)w^{H-1} \quad (7)$$

A complete description of scale-invariance properties and their relation to long-range correlations is beyond the scope of this work, we refer the reader to Audit *et al.*³² where these concepts are discussed in the context of DNA sequence analysis. Here, we will just emphasize two important implications of equation (7). First, the correlation function C_H between gene densities or gene orientation asymmetries measured in windows of size w is constant between all values of w , implying that patterns of genome organisation observed at one scale are statistically indistinguishable from genome organisation patterns at any other scale, i.e. the genome organisation is scale-invariant. Second, C_H decreases as a power-law of the distance l between the windows:^{35–37}

$$C_H(l) \underset{l \rightarrow +\infty}{\propto} l^{2H-2} \quad (H \neq 1/2) \quad (8)$$

This behaviour is in sharp contrast to the much faster exponential decrease observed in many situations, such as Markov models, and justifies its naming as long-range correlations. The larger the value of H (< 1), the slower the decrease in the correlation function and, in that sense, the stronger the long-range correlations. Finally, comparing equations (6) and (7), we see that the absence of genome organisation in which there are no correlations at any scale, corresponds to an exponent $H = 1/2$.

Power-law behaviours are easy to diagnose, as

they appear as straight lines when plotting the standard deviation as a function of the window size w in log–log scale. In order to make any deviation from the uniform model presented previously more apparent, we will analyse the standard deviations of the gene density $\sigma(d_w)$ and of the gene orientation asymmetry $\sigma(a_w)$ normalised by the expected values in the uniform model (equation (6)). In the case of long-range correlations with a Hurst exponent H we thus expect:

$$\log_{10}\left(\frac{\sigma_H(w)}{\sigma_u(w)}\right) = \left(H - \frac{1}{2}\right)\log_{10} w + \log_{10}\left(\frac{\sigma_H(1)}{\sigma_u(1)}\right) \quad (9)$$

So, when plotting $\log_{10}(\sigma_H(w)/\sigma_u(w))$ as a function of $\log_{10} w$, the fact that data points fall on a linear curve enables us to diagnose scale-invariant properties. Measuring the slope of this straight line enables us to estimate $H - 1/2$. Notice, that by normalising the standard deviations, we have selected the uncorrelated state ($H = 1/2$) for the horizontal behaviour.

Wavelet analysis

It is fundamental to stress that the standard-deviation analysis we have just presented can be severely biased by large-scale trends in the data.³⁸ For example, the asymmetry between the leading and the lagging strands clearly constitutes such a trend for the gene orientation asymmetry. A way to overcome this difficulty is to filter out these low-frequency fluctuations by means of oscillating windows, i.e. wavelets, instead of simple windows.^{30,31}

Given the measures of gene orientation asymmetry $a_w(i)$ and gene density $d_w(i)$ in non-overlapping windows of size w indexed by their sequential order i (equation (3)), we now consider:

$$a_w^{(1)}(i) = (a_w(i) - a_w(i+1))/\sqrt{2}$$

and (10)

$$d_w^{(1)}(i) = (d_w(i) - d_w(i+1))/\sqrt{2}$$

This transformation amounts to using the so-called Haar wavelet²⁹ and to characterise the fluctuations of the data around the local mean. In other words, this wavelet removes piecewise constant trends in the data. The normalisation factor $1/\sqrt{2}$ ensures that in the absence of correlation the standard deviation remains unchanged by this transformation. Note that this construct can be generalised to remove higher-order polynomial trends, e.g. $a_w^{(2)}(i) = (2a_w(i) - a_w(i-1) - a_w(i+1))/\sqrt{6}$ corresponds to the french hat wavelet and removes piecewise linear trends in the data. Once the low-frequency trends have been removed, the standard deviation analysis can be performed safely by simply replacing a_w (respectively d_w) by $a_w^{(m)}$ (respectively $d_w^{(m)}$) in the methodology presented previously.

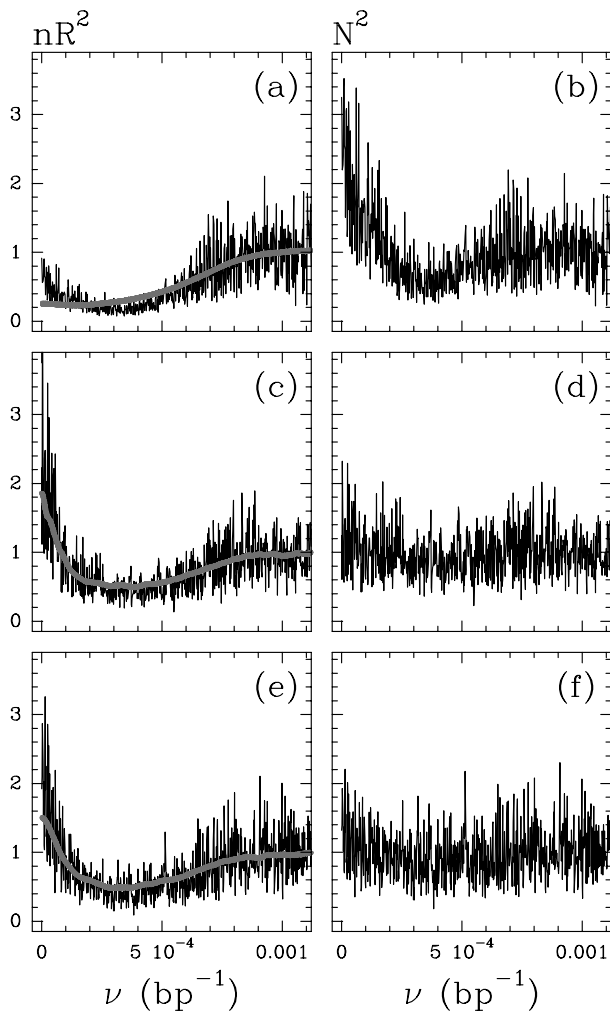


Figure 2. Spectral analysis of gene localisation for *Escherichia coli*. (a) and (b) Complete genome ($n = 4289$ genes). (c) and (d) Genes on the positive strand ($n = 2095$ genes). (e) and (f) Genes on the negative strand ($n = 2194$ genes). In (a), (c) and (e), (—) is the spectrum ($n\bar{R}_p^2$) of the corresponding gene locations and (—) is the expected spectrum given the distribution of inter-gene distances only (nE_p^2). In (b), (d) and (f), (—) is the normalised spectrum (N_p^2) for inter-gene distance distribution. For readability, the curves have been smoothed over a window of eight data points and the order p of the spectral decomposition was converted to frequency: $\nu(\text{bp}^{-1}) = p/\text{genome size}(\text{bp})$. Low frequency or order values are informative of large-scale properties of gene localisation, e.g. $p = 1$ correspond to the entire chromosome length, whereas high frequency or order values indicate segments of small size, e.g. $\nu = 0.001 \text{ bp}^{-1}$ correspond to a size of 1 kbp.

Power-law spectra

Up to now we have characterised scale-invariant properties with standard deviation analysis. But, these distributions also possess a distinctive spectrum characterised by a power-law behaviour of R_p^2 .^{35–37,39}

$$R_p^2 \propto p^{1-2H} \quad (11)$$

This means that scale-invariance can also be diagnosed by a linear behaviour of the spectrum in log–log scale and that H can be estimated by measuring the slope of this line. Note that we recover that in the absence of correlations i.e. when $H = 1/2$, R_p^2 does not depend on p (equation (1)).

Results

We report the results of gene location data in completely sequenced microbial organisms belonging to the eubacterial and archaeal domains. Our goal is to discover and describe the most significant patterns of chromosome organisation with respect to gene position and orientation. In the global approach adopted in this work, we look for such patterns according to the two methodologies presented in Material and Methods. First, we perform a spectral analysis of gene location data using the framework of circular statistics, which are especially well suited to discover hidden patterns with a characteristic size in directional data. Second, we carry out a standard deviation analysis of gene density and gene orientation asymmetry distributions, which is a powerful signal processing technique to investigate correlations over large distances.

Spectral analysis of gene location data

Escherichia coli chromosome

To set up the general framework of our study, we first investigate the genome of *E. coli*. In Figure 2(a), we present the spectrum obtained for the entire *E. coli* genome (4289 genes). Above frequency $\nu \approx 7 \times 10^{-4} \text{ bp}^{-1}$ the spectrum fluctuates around 1, the hallmark of the uncorrelated uniform distribution (equation (1)). That is, no pattern emerges when scanning the gene distribution for lengths smaller than 1500 bp ($\approx 1/7 \times 10^{-4}$). For smaller frequencies, we observe that the amplitude of the spectrum is below the uniform model limit ($n\bar{R}^2 = 1$) and we notice that the spectrum tends to increase again for the lower frequencies. In other words, when we examine the *E. coli* genome over scale lengths greater than 1500 bp, a global ordering emerges. We performed the same analysis focusing only on genes on the positive (2095 genes) or negative (2194 genes) strands (Figure 2(c) and (e)). The two corresponding spectra are indistinguishable and their shape is very similar to the spectrum obtained for the complete gene set. The main differences are a smaller ordering for scale length greater than 1500 bp and especially a much stronger increase in the spectra for the lower frequencies. For these frequencies which correspond to scale lengths comparable to the genome size, both spectra are above the uniform model limit.

From these three spectra it is apparent that the

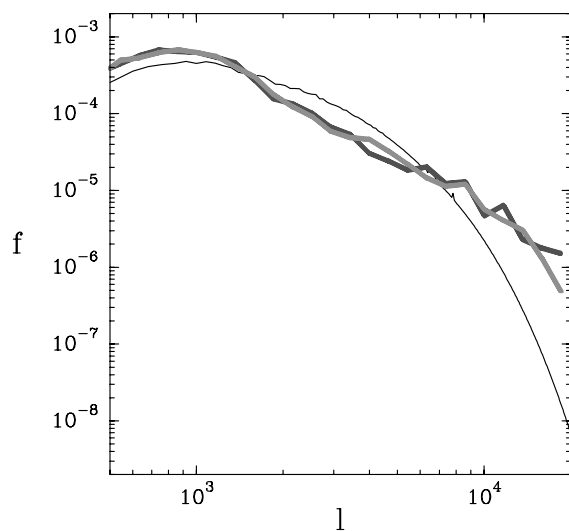


Figure 3. Probability distribution function f of inter-gene distances l for *Escherichia coli*. The observed inter-gene distance distributions for genes on the positive strand (—, 2095 genes) or genes on the negative strand (---, 2194 genes) are compared to the expected distribution of inter-gene distances on one strand given the inter-gene distance distribution for the complete set of genes and equal probability for each of the two gene orientations (· · ·).

global gene localisation as well as gene localisation on each strand do not follow simple uniform distributions. To properly interpret this deviation from the uniform model, it is important to note that genomes can be viewed as the concatenation of objects (gene + intergenic region). From this perspective, the spectra depend on two fundamentally different properties: (i) the length distribution of the objects, i.e. the inter-gene distance distribution and (ii) the correlations between these objects. The inter-gene distance distribution is an isolated object attribute, whereas the correlations describe the actual chromosome structure, i.e. the relationships between the objects. Thus, it is fundamental to assess whether the deviations from the uniform model are a simple consequence of the inter-gene distance distributions or result from genuine genome organisation patterns. If gene positions were uniformly distributed, the inter-gene distance would follow an exponential distribution of parameter $s/n \approx 1$ kbp and would therefore have mean and standard deviation equal to 1 kbp. The actual standard deviation for the *E. coli* inter-gene distances is 0.5 kbp. Therefore, the inter-gene distances are more homogeneous than expected from a uniform model. This could explain that, for frequencies lower than 7×10^{-4} , i.e. when scanning the genome over distances larger than the inter-gene distance, the spectrum drops below the uniform model limit (Figure 2(a)). Along the same lines, we notice that inter-gene distance distributions for genes on either one of the two strands are wider ($\sigma_+ \approx 3$ kbp and $\sigma_- \approx 2.5$ kbp) than expected from a uniform model ($\sigma_u \approx 2$ kbp). This

is apparent also in Figure 3, where the linear decrease in the inter-gene distance distributions on a log-log scale is the signature of a slow power law decrease, compared to a fast exponential decrease as expected for an uncorrelated model. Therefore, the strong increase in the spectra in Figure 2(c) and (e) when going to the lowest frequencies could signify the unexpectedly high number of large inter-gene separations on either strand (Figure 3).

In order to quantify the possible effects on inter-gene distance distributions on the spectra, we computed the expected spectra given the observed distances on the *E. coli* genome only (no correlations). For each spectrum, this was achieved by averaging 1000 spectra obtained after a bootstrap resampling of the corresponding gene distances. In Figure 2(a), (c) and (e), we have superposed these expected spectra on the observed spectra. Two different situations emerge. For the genes on either of the two strands, the shape of the observed and expected spectra follow one another closely. The wide distribution of inter-gene distances on each strand (Figure 3) is sufficient to explain the observed spectra (Figure 2(c) and (e)). On the contrary, the global inter-gene distances do not fully account for the observed spectrum (Figure 2(a)). As we anticipated above, the homogeneity of inter-gene distances induce a weakening of the spectral amplitude when scanning the genome on a length scale greater than the characteristic inter-gene distance (≈ 1 kbp). This weakening should propagate up to the lowest frequencies: once we scan the genome over a length scale much greater than the characteristic inter-gene distance, we should recover a flat spectrum characteristic of the absence of correlations. As a consequence, the low frequency increase in the spectrum observed for the complete gene localisation data cannot be explained by the distribution of inter-gene distances. This is the signature of the presence of correlations between the inter-gene distances and therefore of a genome structure where genes tend to cluster over large distances.

The previous results suggest that an interesting way to investigate genome structure is to use a spectral analysis normalised for inter-gene distance distributions. This simply amounts to using the normalised spectra \tilde{N}_p^2 obtained by dividing the observed spectrum \tilde{R}_p^2 by the expected spectrum given the inter-gene distances only \tilde{E}_p^2 . Importantly, the expected normalised spectra have the same statistical properties as the spectrum of the uniform model (\tilde{E}_p^2 is the exponential of parameter 1). This implies that equation (2) applied to a normalised spectrum gives the probability to reject the null hypothesis, i.e. the absence of correlations in the data. The normalised spectra for the *E. coli* genome are displayed in Figure 2(b), (d), and (f). They enable a quick reading of the results obtained previously. Figure 2(d) and (f) reveal two flat spectra: there are no correlations between inter-gene distances on either of the two strands and hence

no pattern in genome structure is uncovered. Figure 2(b) exhibits a spectrum that can be divided into two parts. For frequency $\nu \geq 3.5 \times 10^{-4} \text{ bp}^{-1}$ we identify a flat spectrum but when we go toward the low frequencies we see a spectrum that steadily increases. This means that over scale lengths larger than $\sim 3 \text{ kbp}$ ($\approx 1/3.5 \times 10^{-4}$) the inter-gene distances are (positively) correlated and consequently that the *E. coli* genome is structured. It is important to understand that the continuous increase in the spectrum down to the lowest frequencies is a clear indication that these correlations extend to distances of the order of the genome size and therefore that the structural patterns we are unravelling effectively refer to a range of lengths from the inter-genic distance up to the complete genome scale.

Finally, it is important to stress that the spectral approach associated with the perspective that genomes are the concatenation of elementary patterns (gene + inter-genic region) enables a real deconvolution of the element characteristics (length) from the collective properties (correlations). This distinction cannot be achieved by simply testing for uniformity. For example, if we use the Kuiper's test²⁶ to assess the random nature of gene localisation data for the *E. coli* genome we get the following probabilities of uniformity: $P = 0.58$, $P_+ = 1.2 \times 10^{-4}$ and $P_- = 4.1 \times 10^{-6}$ for the complete *E. coli* gene set, the positive strand genes and the negative strand genes, respectively. From this analysis, we would conclude there clearly exists some kind of unexpected arrangement for gene localisation on either of the two strands but that the overall gene distribution is uniform. Therefore, we would completely miss the correlations that signify a high-order organisation for the *E. coli* chromosome, such as those detected in Figure 2(b).

Microbial chromosomes

We now extend the spectral analysis of gene localisation, as we performed for the *E. coli* chromosome, to 86 circular chromosomes from the eubacterial and archaeal domains. For each of the 86 chromosomes, we computed the spectra $n\bar{R}_p^2$ and the normalised spectra \bar{N}_p^2 in the same manner as for the *E. coli* chromosome (Figure 2). The inspection of these $3 \times 86 = 258$ spectra and 258 normalised spectra shows that the shape of the spectra for all the bacterial as well as all the archaeal chromosomes analysed here are qualitatively similar to the corresponding *E. coli* spectra (data not shown). To illustrate this universality of the spectral content across the bacterial and the archaeal domains, we present the low-frequency section ($p \leq 15$, i.e. genome segments corresponding to 1/15th of the chromosomal length and above) of the normalised spectra \bar{N}_p^2 for the 86 chromosomes (Figure 4). The images in Figure 4 are colour-coded according to the significance level of the normalised spectra (equation (2)).

The uniformity within each column depicts the universality of spectral content. As discussed for Figure 2, the deviations from the uniform model for the normalised spectra strongly indicate the presence of correlations in localisation data, and not an unexpected inter-gene distance distribution. It is clear that complete gene sets (Figure 4, first column) systematically exhibit strong correlations at all low frequencies, compared to positive and negative strand genes (Figure 4, two last columns).

Following Figure 4, it is possible to look for patterns with a characteristic size in the positional arrangement of genes. Such patterns should appear as very significant modes that strongly outweigh the neighbouring spectral background. A close examination of Figure 4 reveals that no such motif appears from the complete gene sets analysis (first column). In comparison, order $p = 1$ and to a lesser extent its harmonic $p = 3$ are clearly above the spectral background for the positive and negative strand gene spectra (Figure 4, two last columns). It is remarkable that in most cases, the two strands of a given chromosome behave qualitatively the same. This pattern of organisation clearly displays phylogenetic specificity (B.A., unpublished results) and corresponds to the asymmetry in gene density between the leading and the lagging strand.¹⁹

To get a view on a wider frequency range, we present the mean normalised spectra obtained as the arithmetic mean of the corresponding complete gene set, positive strand gene or negative strand gene spectra (Figure 5). This approach can help to uncover characteristic modes that are not strong enough to be significant within the noise of single chromosomes. On the one hand, we have seen previously for *E. coli* (Figure 2) that the average inter-gene distance plays an important role as a high-frequency cutoff for the spectra. This suggests that we should average spectra over equal frequency $\nu = p/s$, i.e. over equal physical size. On the other hand, the only significant mode we delineated so far corresponds to a specific order ($p = 1$), so that it may be more appropriate to perform the average over equal order p , i.e. over equal fraction of chromosome size. The two complementary analyses are presented in Figure 5. Note that the data are restricted to the 69 chromosomes with a length s greater than 1.5 Mbp. This ensures a resolution of 35 data points linearly distributed in logarithmic scale in the frequency range $5 \times 10^{-6} \leq \nu \leq 10^{-3} \text{ bp}^{-1}$ for the equal frequency average and guarantees for the equal order average that for $p \leq 150$ we scan chromosomes for patterns of period $T = s/p$ greater than 10 kbp.

Figure 5(a) exhibits the three mean normalised spectra for the equal frequency average. It is remarkable that no characteristic frequency emerges from the spectra except for a significant weakening below the uncorrelated limit $\bar{N}_\nu^2 = 1$ of the spectral power for the frequency $\nu = 3.5 \times 10^{-4} \text{ bp}^{-1}$ which signifies some repulsive

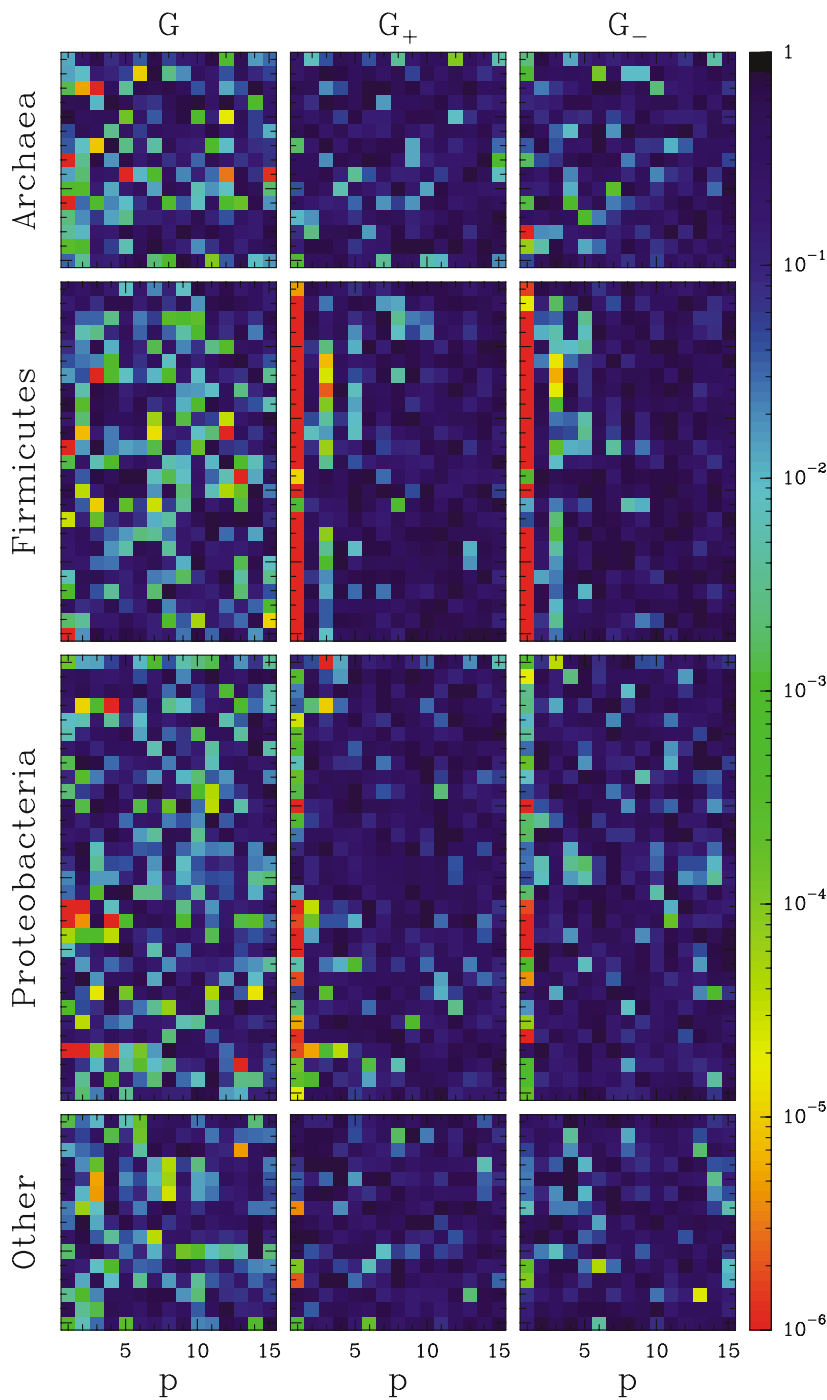


Figure 4. Low-frequency section of the normalised spectra of gene localisation for 86 circular chromosomes. Chromosomes are along the ordinate and spectral orders $p \leq 15$ are along the abscissa. From top to bottom, lines of images correspond to 15 archaeal chromosomes, 25 firmicutes chromosomes, 31 proteobacterial chromosomes and 15 other bacterial chromosomes and within each image genomes are ordered by taxonomic classification, e.g. crenarchaeota on top of euryarchaeota. From left to right, columns of images are the normalised spectra for the complete gene sets G , positive strand gene set G_+ and negative strand gene set G_- , respectively. Images are colour-coded according to the significance level of the normalised spectra ($1 - P_u = \exp(-\bar{N}_p^2)$, equation (2)) from dark blue, non-significant modes, to red, most significant modes.

(anti-correlated) organisation for distances in the order of 3 kbp. The most striking feature in [Figure 5\(a\)](#) is the linear behaviour of the three mean spectra over a wide frequency range ($5 \times 10^{-6} \leq \nu \leq 10^{-4} \text{bp}^{-1}$) in the log-log representation we adopted. It demonstrates that the three mean normalised spectra are power laws of frequency ν , reflecting a scale-invariant genome organisation and long-range correlations between gene locations from 10 kbp to 0.2 Mbp at least (see Material and Methods and equation (11)). [Figure 5\(b\)](#) presents the normalised spectra average over equal order p . Note that for the full range of order presented ($1 \leq$

$p \leq 150$) we scan genomes for distances greater than 10 kbp, as identified in [Figure 5\(a\)](#). Again, except for modes $p = 1$ and its harmonics $p = 3$ and 5 for positive or negative-strand gene spectra, no other characteristic order is revealed after averaging. Interestingly, we observe the same linear behaviour of the spectra as described previously from $p \sim 150$ down to the lowest order. This demonstrates that the scale-invariant organisation uncovered by this spectral analysis propagates up to the full-chromosome sizes.

Finally, by measuring the slope sl of the linear relationship between the normalised spectra and

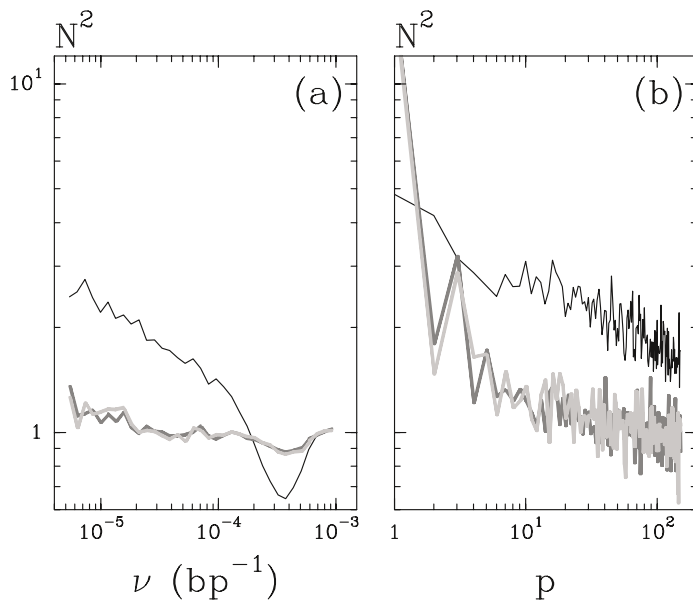


Figure 5. Spectral analysis of 69 bacterial and archaeal circular chromosomes of length s greater than 1.5 Mbp. In (a) the normalised spectra (\bar{N}_p^2) are average over equal frequency $\nu = p/s$. Note that in order to perform this average all spectra have to be smoothed in order to obtain a common sampling in the frequency domain. Here, in the range $5 \times 10^{-6} \leq \nu \leq 10^{-3} \text{bp}^{-1}$, we have 35 frequency samples linearly distributed in log scale. In (b) the normalised spectra are average over equal order p . In (a) and (b) (—) corresponds to complete gene sets, (---) to positive strand genes and (· · ·) to negative strand genes.

the frequency ν or the order p in log–log scale, we can quantify the strength of the long-range correlations, i.e. the Hurst exponents $H = (1 - sl)/2$ (equation (11)). The two types of spectra give the same estimation for the Hurst exponents. For the complete gene sets spectra, we get $H_G = 0.60 \pm 0.01$ for $5 \times 10^{-6} \leq \nu \leq 10^{-4} \text{bp}^{-1}$ or $1 \leq p \leq 150$. For the positive-strand gene spectra and the negative-strand gene spectra we estimate $H_{G_+} = H_{G_-} = 0.54 \pm 0.02$ for $5 \times 10^{-6} \leq \nu \leq 10^{-4} \text{bp}^{-1}$ or $6 \leq p \leq 150$. For the complete gene set, the estimated Hurst exponent value $H_G = 0.60 \pm 0.01$ is significantly greater than the value $H = 1/2$ which corresponds to the absence of long-range correlations. So, this measurement confirms the existence of a scale-invariant organisation for prokaryotic chromosomes that act on gene location from 10 kbp up to the chromosome sizes. The situation is somewhat different for the positive G_+ and the negative G_- strand genes: (i) the strength of the correlations is smaller $H_{G_+} = H_{G_-} = 0.54 \pm 0.02$ and (ii) as we explained in Material and Methods, the spectra for G_+ and G_- depend on the spectra for the complete gene set G . Therefore, we cannot conclude whether these correlations are significant or if they simply result from the correlations observed for G . In order to assess if gene orientation contributes to the scale-invariant organisation we need to complete the study with the analysis of gene orientation asymmetry.

Gene orientation asymmetry and gene density

We now describe the investigation of gene orientation asymmetry and gene density using the standard deviation technique outlined in Material and Methods. The main idea is to describe a window of size w along a chromosome by its gene density d_w , i.e. its gene content expressed as a num-

ber of genes per base-pair, and its gene orientation asymmetry a_w , i.e. the propensity of genes to be oriented in the positive ($a_w > 0$) or the negative ($a_w < 0$) sense (equation (3)). The most important property of this description of gene content and orientation is that d_w and a_w are defined independently. For a given scale of analysis w , we characterise d_w and a_w by the amplitude of the observed fluctuations when sliding the window along the chromosome, i.e. we measure the standard deviations $\sigma(a_w)$ and $\sigma(d_w)$. Then, we analyse how the standard deviations behave as a function of w . Indeed, in the same manner an unexpected value for a spectral component \bar{R}_p can result from the distribution of inter-gene distances or from correlations between gene location, the standard deviation value at a given scale cannot be readily interpreted as it also depends on both of these attributes (equation (7)). Thus, to effectively assess the existence of correlations in the orientation or location, we use the rate of change of the standard deviations when the scale of analysis is modified, i.e. we estimate the Hurst exponent H (see Material and Methods, and equation (9)).

Fixed-scale analysis

Figure 6 presents the standard deviations of gene density and gene orientation asymmetry for 89 prokaryotic chromosomes at a window size $w = 51.2$ kbp. In Figure 6(a), $\sigma(d_w)$ is plotted as a function of the mean gene density d . The data points cluster around the curve corresponding to the uncorrelated model (equation (5)) and $d \sim$ one gene/kbp. So, at scale 51.2 kbp the gene density conforms to the expectation given by the uniform model and we confirm that prokaryotic genomes are mostly protein-coding with an average gene size of the order of 1 kbp. A closer examination of the scattering of the data along the

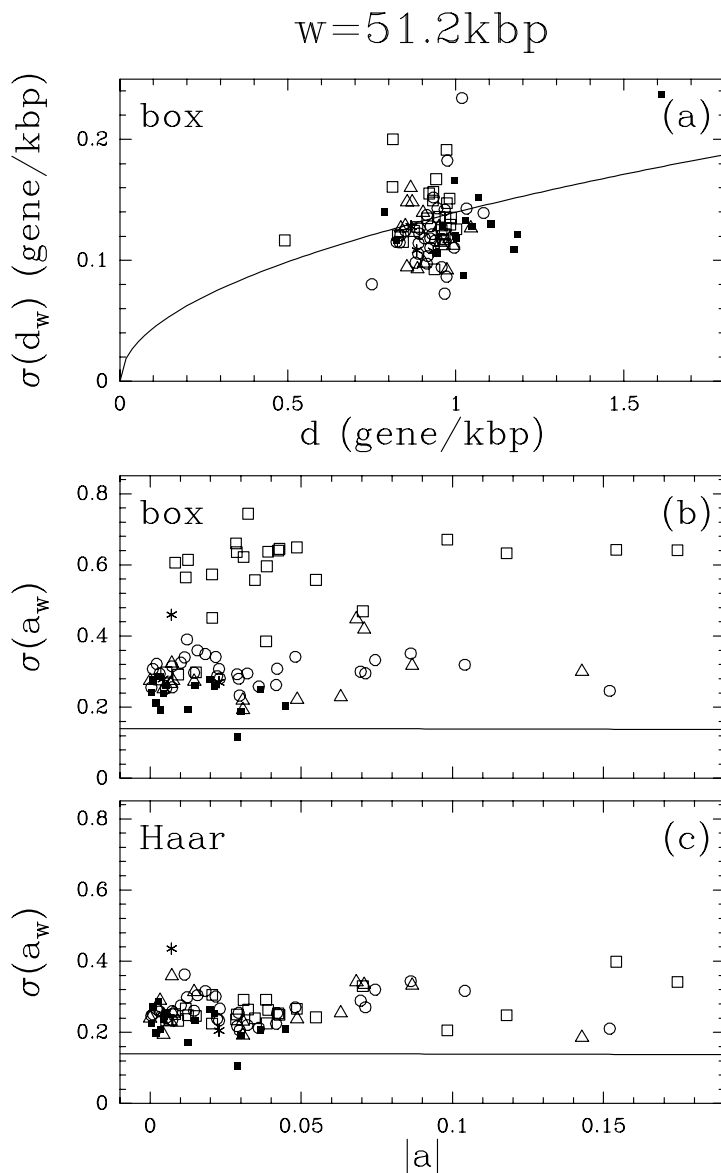


Figure 6. Gene density and gene orientation asymmetry analysis at a fixed window size $w = 51.2 \text{ kbp}$ for 89 prokaryotic chromosomes. (a) Standard deviation $\sigma(d_w)$ of the gene density measured in non-overlapping box windows (equation (3)) as a function of the mean gene density $d = n/s$, where n is the total number of genes and s the chromosome physical length. (b) Standard deviation $\sigma(a_w)$ of the gene orientation asymmetry measured in non-overlapping box windows (equation (3)) as a function of the absolute value of the mean gene orientation asymmetry $a = (n_+ - n_-)/n$ where n_+ (respectively n_-) is the total number of genes on the positive (respectively negative) strand and $n = n_+ + n_-$ is as above. (c) Standard deviation $\sigma(a_w^{(1)})$ of the gene orientation asymmetry measured with the Haar wavelet (equation (10)) as a function of $|a|$. In (a)–(c), (—) follows the expected standard deviations for the uncorrelated model (equations (4) and (5)) when neglecting the finite size effects ($1 - w/s \approx 1$) and assuming $d =$ one gene/kbp in equation (4). The different symbols follow the phylogenetic classification and topology of chromosomes: (■) for the 15 circular archaeal chromosomes, (□) for the 25 circular firmicutes chromosomes, (○) for the 31 circular proteobacterial chromosomes, (△) for the 15 other circular bacterial chromosomes and (*) for the three linear bacterial chromosomes.

ordinate reveals that the fluctuations are roughly twice as large as those expected from the uniform model and that no chromosome appears as an outlier in the distribution of $\sigma(d_w)$ (data not shown). Along the same lines, two chromosomes (*Mycobacterium leprae*: $d \approx 0.49 \text{ gene/kbp}$ and *Aeropyrum pernix*: $d \approx 1.61 \text{ gene/kbp}$) clearly display unexpected mean gene density values (abscissa). For *M. leprae*, the low gene density corresponds to the massive gene decay observed in this genome where 41% of the predicted genes are pseudogenes.⁴⁰ The very high gene density measured for *A. pernix* correlates with a much larger proportion of the annotated proteins having a length between 100 and 150 amino acid residues compared to the same proportion calculated over the other 88 chromosomes (data not shown), possibly representing overpredictions as previously noted.⁴¹

We present the corresponding fixed-scale analysis for the gene orientation asymmetry in Figure

6(b). We observe that the dispersion of the data points for the mean gene orientation asymmetry (abscissa) is again twice as large as the dispersion expected if genes had either of the two possible orientations with equal probability (data not shown). But the most striking feature in Figure 6(b) is the phylogenetic stratification of the standard deviation values $\sigma(a_w)$. The firmicute chromosome $\sigma(a_w)$ values are greater with a mean value $\sigma(a_w) \approx 0.56$ followed by the proteobacterial chromosomes $\sigma(a_w) \approx 0.30$, the other circular bacterial chromosomes $\sigma(a_w) \approx 0.28$ and the archaeal chromosomes $\sigma(a_w) \approx 0.23$. This ordering of the chromosomes closely matches the classification resulting from the intensity of the positive and negative-strand gene normalised spectra for order $p = 1$ (Figure 4). This behaviour suggests a varying distribution of positive and negative strand genes around the chromosomes which induce a comparable varying (non-stationary) distribution of gene orientation asymmetry: orientation

asymmetry fluctuates around a positive value for one-half of the chromosomes and around a negative value for the second-half. Note that this pattern corresponds to the asymmetry of gene density between the leading and lagging strands. As we have stressed in Material and Methods, standard deviation measurement is severely biased by such low-frequency trends and the stratification observed in Figure 6(b) is a consequence of this bias. In order to overcome this technical pitfall, the standard deviations of gene orientation asymmetry can be evaluated using the Haar wavelet (see Material and Methods). In Figure 6(c), we clearly see that the stratification is thus eliminated and that the obtained standard deviation values for 88 out of 89 prokaryotic chromosomes are above the uniform model limit. Moreover, the scattering of the standard deviation values along the ordinate is almost three times as large as expected from the uniform model and the chromosome of *A. pernix* is a clear outlier with a lower level of gene orientation asymmetry fluctuation ($\sigma(a_w^{(1)}) \approx 0.11$), compared to the other 88 chromosomes (data not shown).

The study of gene orientation asymmetry and gene density and their fluctuations at scale $w = 51.2$ kbp demonstrates that there exists a greater variability between chromosomes than expected from the uniform model. At this scale, gene orientation asymmetry within most chromosomes (88/89) seems quite inhomogeneous, although the gene density fluctuations within chromosomes do not exhibit unexpected variations. But as we underlined previously, a single scale of analysis is not sufficient to demonstrate the presence or absence of genome organisation patterns.

Analysis across scales

In Figure 7 we present the standard deviations of gene density and gene orientation asymmetry as a function of the window size w . We have plotted the normalised version σ_n of the standard deviations in log-log scale according to equation (9) so that linear behaviour corresponds to scale-invariant properties, horizontal curves signify a Hurst exponent value $H = 1/2$, i.e. the absence of correlations and, the zero level, i.e. $\sigma_n = 1$ is the expected value for the uniform model (see Material and Methods). In Figure 7(a), the results for gene density d_w calculated in box windows are shown (equation (3)). We see that the curves for the five groupings of chromosomes according to their phylogenetic classification and their topology almost superimpose and are linear with a common slope significantly greater than 0, corresponding to a Hurst exponent $H_d = 0.61 \pm 0.02$ for $3 \text{ kbp} \leq w \leq 400 \text{ kbp}$. As a consequence of the homogeneous inter-gene distance distribution (see Spectral analysis of gene location data), the gene density standard deviation curves are below the uniform model value 0 over most of the accessible scale range ($3 \text{ kbp} \leq w \leq 100 \text{ kbp}$). Similar results

are obtained in Figure 7(b) where gene density standard deviations were estimated using the Haar wavelet method (equation (10)), showing that there are no significant trends underlying the gene location data. This study corroborates the existence of a scale-invariant distribution of gene position and the existence of long-range correlations already demonstrated with the spectral analysis of complete gene sets G (Figure 5) and we observe that the estimates of the Hurst exponent obtained by the two methods are in very good agreement: $H_G = 0.60 \pm 0.01$ and $H_d = 0.61 \pm 0.02$. Moreover, by combining the different scale ranges in which each of these two complementary analyses were possible, we can conclude that the long-range correlations between gene positions extend over three orders of magnitude from 3 kbp to the chromosome sizes.

In Figure 7(c) and (d), we display the scale dependence of the normalised standard deviation of gene orientation asymmetry. In the same manner as for the analysis at a fixed window size $w = 51.2$ kbp, we see that when standard deviations are calculated using box counting (equation (3)), the estimates are biased by the non-stationary gene orientation asymmetry distributions: the stronger the spectral modes for $p = 1$ (Figure 4), the steeper the curves in Figure 7(c). When these biases are removed by calculating the standard deviations with the Haar wavelet method (equation (10)), we observe that the curves corresponding to four out of the five chromosome groups behave identically (Figure 7(d)). The curve for the firmicutes still displays a divergent behaviour when the window size w increases. Indeed, the Haar wavelet enables us to remove the local mean value for the gene orientation asymmetry but at the frontier between two portions of a chromosome that have different mean values, a local peak remains in the data. When these local peaks are too strong, as is the case for the firmicutes due to the severe change of gene orientation asymmetry observed at the origin and the termination of replication,¹⁹ these estimates can be affected. For the other four chromosome groups, we notice that from a window size $w = 20$ kbp, which is of the order of the largest inter-gene distances on one strand (Figure 3), up to the maximum scale $w = 400$ kbp reached in this analysis, the curves are linear with a common positive slope corresponding to a Hurst exponent $H_a = 0.59 \pm 0.03$. Below $w = 20$ kbp, it is not possible to claim for linearity with only three data points. Nevertheless, we see that the curves have larger slopes than for $w > 20$ kbp and suggest a Hurst exponent $H_a \sim 0.70$. This confirms that the length distribution of the tracks of co-oriented genes correspond to strong correlations between the orientation of neighbouring genes (Figure 3). This study demonstrated that there exist long-range correlations in the distribution of gene orientation asymmetry and, as a consequence, that gene orientation does participate in the scale-invariant

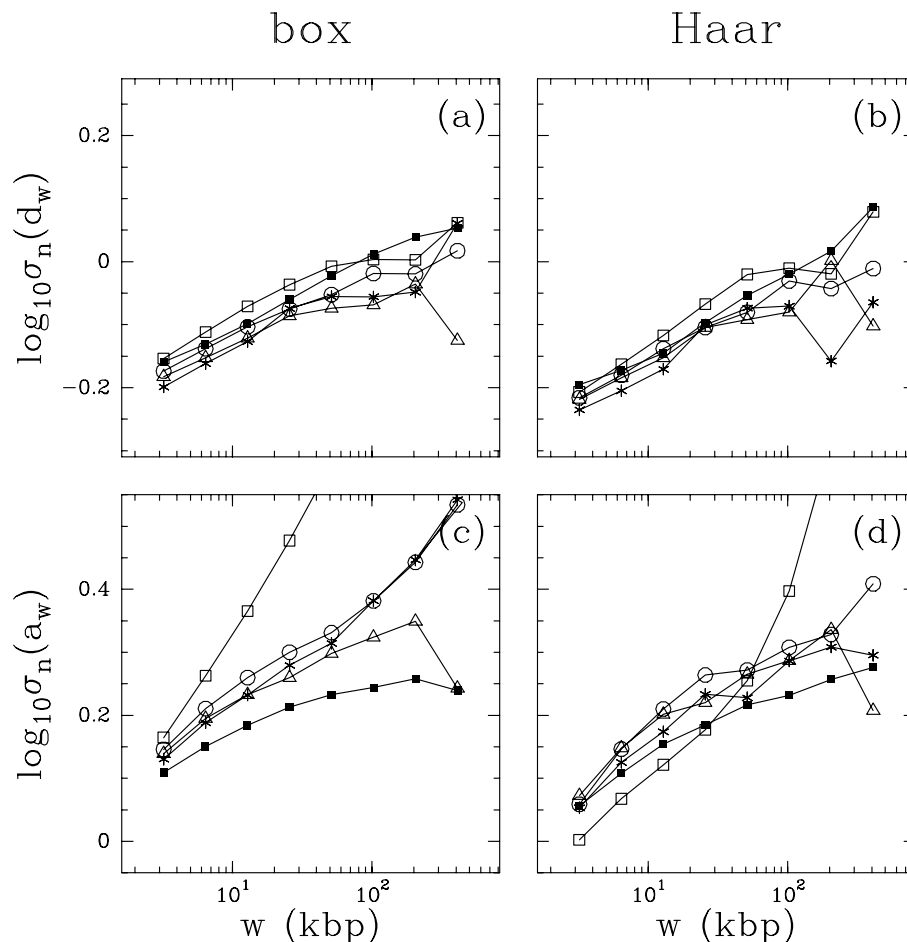


Figure 7. Standard deviations of gene density $\sigma(d_w)$ and gene orientation asymmetry $\sigma(a_w)$ as a function of the window size w for 89 prokaryotic chromosomes. In the four graphs, each curve corresponds to the mean of the standard deviations over a set of chromosomes with homogeneous phylogenetic classification and topology: (■) for the 15 circular archaeal chromosomes, (□) for the 25 circular firmicutes chromosomes, (○) for the 31 circular proteobacterial chromosomes, (△) for the 15 other circular bacterial chromosomes and (*) for the three linear bacterial chromosomes. We plot the normalised mean standard deviation σ_n in log–log scale following equation (9) where we have neglected the finite size effect ($1 - w/s \approx 1$) to calculate the uniform model standard deviations $\sigma_u(w)$ (equations (4) and (5)). We present in (a) and (b) the normalised standard deviations of the gene density $\sigma_n(d_w)$ and, in (c) and (d) the normalised standard deviations of the gene orientation asymmetry $\sigma_n(a_w)$. In (a) and (c), gene density and gene orientation have been calculated in non-overlapping box windows (equation (3)) and in (b) and (d) we used the Haar wavelet approach (equation (10)). In these graphs, a linear behaviour signs scale-invariance, a horizontal line is the hallmark of the absence of correlations ($H = 1/2$), whereas a strictly positive slope signifies the existence of long-range correlations ($H > 1/2$) (equation (9)).

organisation of prokaryotic genomes at least up to a scale of 400 kbp. It is also interesting to notice that over the complete range of window sizes, the curves are over the uniform model and that this gap steadily increases with the window size w . So, when we observe gene orientation asymmetry along a chromosome at progressively larger scales, it appears increasingly heterogeneous (compare to the uniform model expectation).

Discussion

There is no characteristic size in chromosome organisation

Here, using circular statistics,^{26–28} we have per-

formed a spectral analysis of 86 circular chromosomes from the archaeal and the bacterial domains (Figures 2, 4, and 5). This statistical framework enables us to recover previously established characteristics of *E. coli* gene location patterns and extend them to other genomes. Previously, it has been proposed that gene density in *E. coli* is higher at the origin of replication,^{15,22,23} which would correspond to a strong $p = 1$ mode in this analysis. Furthermore, it has been suggested that gene clusters are regularly spaced²⁴ and more specifically with a 4-fold symmetry for the *E. coli* glucose catabolism genes.⁴² Within the context of our analysis, the former should correspond to frequency characteristic of the reported cluster sizes while the latter to spectral modes $p = 4$. In general, any pattern pertinent to gene organisation as a whole

and involving a characteristic size should be detectable by the spectral analysis. This analysis was performed for complete gene sets G and positive G_+ or negative G_- strand genes of individual prokaryotic chromosomes (Figures 2 and 4) as well as the mean spectra obtained when averaging over equal order p or equal frequency ν (expressed as the inverse of DNA physical length) (Figure 5). Our findings underline the significance of the $p = 1$ symmetry for G_+ and G_- (Figure 1) that is dependent on the phylogenetic class of the species under investigation (Figure 4). Surprisingly, this is the only significant mode emerging from the spectral analysis of genome organisation. In other words, except for the chromosome length, there is no single characteristic size or scale that describes the localisation and orientation of genes along chromosomes (Figures 2, 4, and 5). Instead, we observe the spectral content behaves as a power-law of the order p and frequency ν (Figure 5) from a few inter-gene distances ($\nu = 10^{-4}$) up to the chromosome length ($p = 1$). More precisely, we have discovered that there exist long-range correlations with a Hurst exponent $H_G = 0.60 \pm 0.01$ between gene locations, and thus we have unravelled a scale-invariant organisation of chromosomes. This result implies that for all scales of observation from 10 kbp up to the chromosome length, genes display significant clustering and that all these sizes are important and necessary to describe prokaryotic chromosomes organisation.

As mentioned above, previously proposed patterns for the case of *E. coli* have not been confirmed in the present analysis. This apparent contradiction can be explained by the fact that the early work on this subject was performed on partial genetic maps containing only between 600²² and 1400 genes,²⁴ not necessarily representative of the *E. coli* complete genome. Interestingly, the spectral analysis of sets of *E. coli* genes found to be conserved across large evolutionary distances, e.g. genes with an ortholog in *B. subtilis*, exhibit a higher gene concentration around the origin of replication (data not shown). Note that some of the previous results involved functional classification of genes,^{23,42} which was not taken into account in the present work. The results described here do not rule out the possible existence of functional domains²³ or periodicities specific to the distribution of genes sharing a common functional class.⁴²

Universal hierarchical organisation of prokaryotic chromosomes

To complete our description of gene position and orientation data, we extended the analysis to assess whether gene orientation is involved in the scale-invariant organisation of prokaryotic chromosomes. Because it is fundamental to study gene orientation independently of gene localisation, we analysed gene orientation asymmetry (equation (3)). This quantity does not lend itself to spectral analysis, we thus used an alternative approach

based on standard deviation analysis. The non-stationary nature of gene orientation asymmetry induced by the $p = 1$ symmetry which is present in the G_+ and G_- gene sets (Figure 4) required wavelet filtering prior to any standard deviation estimation.^{30,38} We observed that the standard deviation of gene orientation asymmetry behaved as a power-law of the window size w from $w = 20$ kbp up to at least $w = 400$ kbp (Figure 7(d)). These results clearly demonstrate that gene orientation is long-range correlated with a Hurst exponent $H_a = 0.59 \pm 0.03$ and distributed in a scale-invariant manner, i.e. for all scales of observation in this scale range we discern unexpectedly high gene co-orientation. We also conducted a study of gene density distribution using the same methodology (Figure 7(a) and (b)). This analysis confirms that gene positions are long-range correlated from $w = 3$ kbp up to $w = 400$ kbp with a Hurst exponent $H_d = 0.61 \pm 0.02$.

Finally, we have been able to detect for the first time the following elements: (i) scale-invariant properties of the spatial gene organisation that are universally present across archaeal and bacterial chromosomes including the three linear bacterial chromosomes available to date; (ii) the Hurst exponents describing the scale-invariance distributions of gene positions ($H_G = 0.60 \pm 0.01$ and $H_d = 0.61 \pm 0.02$) and gene orientation ($H_a = 0.59 \pm 0.03$) are virtually identical, strongly suggesting that the forces responsible for the long-range correlations are common between gene position and orientation; and (iii) that scale-invariance propagates up to the chromosome length for gene positions. Together, these results suggest an operon-like organisation of genome structure from a few kilobases up to the megabase scale.

Global superhelical context and nucleoid structure

We believe that the involvement of gene orientation in the universal scale-invariant organisation of prokaryotic genomes is the key starting point towards the understanding of its biological significance. We observe that gene orientation asymmetry is correlated along the chromosomes, i.e. knowing that a gene as a positive (respectively negative) orientation, the neighbouring genes are likely to have that same positive (respectively negative) orientation. This co-orientation constraint may be induced by regulatory mechanisms of gene expression, possibly involving transcription.

In fact, transcription induces superhelicity along the DNA template, positive in front of the advancing RNA polymerase and negative in its trail and in turn, the convergent (respectively divergent) progression of two simultaneously transcribing RNA-polymerases induces the accumulation of positive (respectively negative) superhelicity in the intervening DNA stretch. During this process, the effects of two co-orientated RNA polymerases cancel each other.^{34,43–45} The DNA superhelical

state strongly influences gene expression,^{34,45} by modulating the level of activity of promoters.⁴⁶ Also, positive supercoiling in front of RNA polymerase could reduce the speed of transcription as it is the case for DNA polymerase when replication and transcription occur in opposite directions.⁴⁷ Thus, there exists a possibility for direct feedback of transcription on the control of gene expression through DNA superhelicity. Moreover, the short-range structural order of the DNA in the nucleoid is partly supported by superhelically induced supercoiling⁴⁸ so that transcription can also affect DNA structure. In living cells, the level of superhelicity is regulated to a constant level by the combined action of DNA topoisomerases⁴⁹ although these enzymes cannot accommodate acute superhelical changes.^{34,43,44} Therefore, co-orientation of genes could be favored during prokaryotic genome evolution because it favors gene expression by reducing the structural consequences of transcription and the need for elevated topoisomerase activity.³⁴

The importance of the superhelical context as an evolutionary force can easily be understood at the operon level. In addition, the propagation of transcriptionally induced superhelicity is limited^{44,47} so that the superhelical context alone is unlikely to have consequences up to the genome scale. How can we understand the observed long-range correlation of gene orientation given this limitation? One possibility is the coupling of the superhelical context with the structure and dynamics of the nucleoid. The nucleoid structure has been described mainly for *E. coli* and it involves the folding of DNA into large chromosomal domains that further condense due to supercoiling and other mechanisms.^{48,50} Experimental data suggest that gene expression only takes place on the periphery of the nucleoid, implying that the nucleoid is a highly dynamical structure that has to be remodelled to match transcriptional requirements.^{48,51} Note that *in vitro* experiments also substantiate the possibility of a highly dynamic nucleoid.⁵⁰ To further expand on the above possibility taking into account the scale-invariant organisation of prokaryotic genomes, a putative scenario is as follows. At a given time, all genes that need to be expressed are located on the nucleoid surface and define domains along the chromosomes that are highly expressed and on which the superhelical context constraint is active. Because the surface of the nucleoid is a restricted space, we can also understand that these domains also correspond to gene dense regions of the chromosome. During the cell-cycle or in response to environmental changes the pattern of gene expression is modified and the previous domain structure is reorganised. This may require a complete redefinition of the domains in terms of size and location, which could be facilitated by a scale-invariant organisation of the chromosome into domains whose sizes follow a power-law distribution.

Indeed, similar principles have been described for protein interaction networks, exhibiting connectivity distribution that follows a power-law and forming intricate scale-free graphs.⁵² This type of topology does not accommodate the sequential ordering of genes along the chromosomes, e.g. it is not possible to position genes in such an order that each metabolic pathway corresponds to a gene cluster. Therefore, a reasonable hypothesis explaining our observation is as follows: the different expression patterns needed during the various stages of cellular life dynamically define expression and structural domains whose length distribution follows a power-law. The superhelical as well as spatial constraints would then be at work on these domains and result in a common scale-invariant organisation of gene orientation and gene localisation. Note that the forces exerted by RNA polymerases on the chromosomes coupled with the gene orientation asymmetry could participate in the dynamic of these domains.²⁰

To substantiate the above hypothesis, we mention supporting evidence published recently. This hypothesis necessitates the coupling of nucleoid structure to cellular expression patterns. In *E. coli* the FIS architectural protein has been described to achieve this connection⁵³ and it has been shown for *E. coli* that the regions of high density of binding sites for the integration host factor protein correspond to regions of low expression, suggesting possible links between transcription and chromosomal structure.⁵⁴ The binding of chromatin-associated (or other) proteins to DNA is affected by DNA curvature, while universal long-range correlations between DNA bending sites have been described³³ allowing the possibility for a scale-invariant distribution of structural domains. In our model, co-expressed genes need to belong to the nucleoid surface concurrently. As a consequence, functional association should then favor localisation in co-existing domains. Some indirect evidence of this coupling between function and localisation is available: functional domains have been previously described on the *E. coli* genetic map²³ and more recently it was shown that genomic rearrangements maintain genes in a specific functional context called uber-operon.⁵⁵

The work presented here strongly underlines the need of performing this type of analyses of gene position and orientation data in relation to gene function, expression profiles and protein interaction data. Such studies could provide us with valuable insights towards a deeper understanding of the major underlying principles for the encoding of cellular functions in DNA sequences.

Acknowledgements

We thank Dag Ahrén, Ildefonso Cases and Pietro Liò for useful comments on the manuscript. B.A. is

currently supported by a Marie Curie Fellowship of the European Community programme "Improving Human Research Potential and the Socio-economics Knowledge Base" under contract number HPMF-CT-2001-01321. Additional support was provided by the European Molecular Biology Laboratory.

References

- Huynen, M. A., Snel, B., Lathe, W. C., III & Bork, P. (2000). Exploitation of gene context. *Curr. Opin. Struct. Biol.* **10**, 366–370.
- Marcotte, E. M. (2000). Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.* **10**, 359–365.
- Tsoka, S. & Ouzounis, C. A. (2000). Recent developments and future directions in computational genomics. *FEBS Letters*, **480**, 42–48.
- Huynen, M. A. & Bork, P. (1998). Measuring genome evolution. *Proc. Natl Acad. Sci. USA*, **95**, 5849–5856.
- Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
- Ouzounis, C. A. & Kyrpides, N. C. (1996). The emergence of major cellular processes in evolution. *FEBS Letters*, **390**, 119–123.
- Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., *et al.* (1996). Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* **6**, 279–291.
- Tamames, J., Casari, G., Ouzounis, C. A. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66–73.
- Dandekar, T., Snel, B., Huynen, M. A. & Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328.
- Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
- Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucl. Acids Res.* **28**, 3442–3444.
- Tillier, E. R. M. & Collins, R. A. (2000). Genome rearrangement by replication-directed translocation. *Nature Genet.* **26**, 195–197.
- Salgado, H., Moreno-Hagelsieb, G., Smith, T. F. & Collado-Vides, J. (2000). Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
- Moreno-Hagelsieb, G. & Collado-Vides, J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18**, S329–S336.
- Brewer, B. J. (1988). When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell*, **53**, 679–686.
- Mrázek, J. & Karlin, S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl Acad. Sci. USA*, **95**, 3720–3725.
- Rocha, E. P. C., Guerdoux-Jamet, P., Moszer, I., Viari, A. & Danchin, A. (2000). Implication of gene distribution in the bacterial chromosome for the bacterial cell factory. *J. Biotechnol.* **78**, 209–219.
- Lopez, P. & Philippe, H. (2001). Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation. *C. R. Acad. Sci., ser. III*, **324**, 201–208.
- Rocha, E. P. C. (2002). Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes. *Trends Microbiol.* **10**, 393–395.
- Dworkin, J. & Losick, R. (2002). Does RNA polymerase help drive chromosome segregation in bacteria. *Proc. Natl Acad. Sci. USA*, **99**, 14089–14094.
- Mushegian, A. R. & Koonin, E. V. (1996). Gene order is not conserved in bacterial evolution. *Trends Genet.* **12**, 289–290.
- Bachmann, B. J., Low, K. B. & Taylor, A. L. (1976). Recalibrated linkage map of *Escherichia coli* K-12. *Bacteriol. Rev.* **40**, 116–167.
- De Martelaere, D. A. & Van Gool, A. P. (1981). The density distribution of gene loci over the genetic map of *Escherichia coli*: its structural, functional and evolutionary implications. *J. Mol. Evol.* **17**, 354–360.
- Williamson, R. M., Hetherington, J. & Jackson, J. H. (1993). Detection of fundamental principles and a level of order for large-scale gene clustering on the *Escherichia coli* chromosome. *J. Mol. Evol.* **36**, 347–360.
- Jurka, J. & Savageau, M. A. (1985). Gene density over the chromosome of *Escherichia coli*: frequency distribution, spatial clustering, and symmetry. *J. Bacteriol.* **163**, 806–811.
- Fisher, N. I. (1993). *Statistical Analysis of Circular Data*, Cambridge University Press, Cambridge, UK.
- Horimoto, K., Suyama, M., Toh, H., Mori, K. & Otsuka, J. (1998). A method for comparing circular genomes from gene locations: application to mitochondrial genomes. *Bioinformatics*, **14**, 789–802.
- Janssen, P. J., Audit, B. & Ouzounis, C. A. (2001). Strain-specific genes of *Helicobacter pylori*: distribution, function and dynamics. *Nucl. Acids Res.* **29**, 4395–4404.
- Mallat, S. (1998). *A Wavelet Tour of Signal Processing*, Academic Press, New York.
- Arneodo, A., Bacry, E., Graves, P. V. & Muzy, J.-F. (1995). Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Letters*, **74**, 3293–3296.
- Arneodo, A., d'Aubenton-Carafa, Y., Bacry, E., Graves, P. V., Muzy, J.-F. & Thermes, C. (1996). Wavelet based fractal analysis of DNA sequences. *Physica sect. D*, **96**, 291–320.
- Audit, B., Vaillant, C., Arneodo, A., d'Aubenton-Carafa, Y. & Thermes, C. (2002). Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes. *J. Mol. Biol.* **316**, 903–918.
- Audit, B., Thermes, C., Vaillant, C., d'Aubenton-Carafa, Y., Muzy, J.-F. & Arneodo, A. (2001). Long-range correlations in genomic DNA: a signature of the nucleosomal structure. *Phys. Rev. Letters*, **86**, 2471–2474.
- Charlebois, R. L. & St. Jean, A. (1995). Supercoiling and map stability in the bacterial chromosome. *J. Mol. Evol.* **41**, 15–23.
- Mandelbrot, B. B. (1982). *The Fractal Geometry of Nature*, Freeman & Co., San Francisco.
- Stanley, H. E., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Ostadnik, S. M., Peng, C.-K. & Simons,

- M. (1993). Fractal landscapes in biological systems. *Fractals*, **1**, 283–301.
37. +Samorodnitsky, G. & Taqqu, M. S. (1994). *Stable Non-Gaussian Random Processes*, Chapman & Hall, New York.
38. Karlin, S. & Brendel, V. (1993). Patchiness and correlations in DNA sequences. *Science*, **259**, 677–679.
39. Li, W. & Kaneko, K. (1992). Long-range correlation and partial $1/f^\alpha$ spectrum in a noncoding DNA sequence. *Europhys. Letters*, **17**, 655–660.
40. Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., *et al.* (2001). Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–1011.
41. Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D. W. & Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.* **17**, 425–428.
42. Riley, M., Solomon, L. & Zipkas, D. (1978). Relationship between gene function and gene location in *Escherichia coli*. *J. Mol. Evol.* **11**, 47–56.
43. Liu, L. F. & Wang, J. C. (1987). Supercoiling of the DNA template during transcription. *Proc. Natl Acad. Sci. USA*, **84**, 7024–7027.
44. Rahmouni, A. R. & Wells, R. D. (1992). Direct evidence for the effect of transcription on local DNA supercoiling *in vivo*. *J. Mol. Biol.* **223**, 131–144.
45. Dröge, P. (1994). Protein tracking-induced supercoiling of DNA: a tool to regulate DNA transactions *in vivo*? *BioEssays*, **16**, 91–99.
46. Pruss, G. J. & Drlica, K. (1989). DNA supercoiling and prokaryotic transcription. *Cell*, **56**, 521–523.
47. French, S. (1992). Consequences of replication fork movement through transcription units *in vivo*. *Science*, **258**, 1362–1365.
48. Pettijohn, D. E. (1996). The nucleoid. *Escherichia coli and Salmonella* (Neidhardt, F., Curtiss, R., Ingraham, J. L., Lin, E. C. C., Low, K. B., Magasanik, B., *et al.*), pp. 158–166, AMS Press, Washington, DC.
49. Drlica, K., Burger, R. M., Franco, R. J., Hsieh, L. S. & Berger, B. A. (1990). Roles of DNA topoisomerases in bacterial chromosome structure and function. In *The Bacterial Chromosome* (Drlica, K. & Riley, M., eds), pp. 195–203, ASM Press, Washington, DC.
50. Cunha, S., Woldringh, C. L. & Odijk, T. (2001). Polymer-mediated compaction and internal dynamics of isolated *Escherichia coli* nucleoids. *J. Struct. Biol.* **136**, 53–66.
51. Kellenberger, E. (1990). Intracellular organization of the bacterial genome. In *The Bacterial Chromosome* (Drlica, K. & Riley, M., eds), pp. 173–186, ASM Press, Washington, DC.
52. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.
53. Schneider, R., Travers, A., Kutateladze, T. & Muskhelishvili, G. (1999). A DNA architectural protein couples cellular physiology and DNA topology in *Escherichia coli*. *Mol. Microbiol.* **34**, 953–964.
54. Ussery, D. W., Larsen, T. S., Wilkes, K. T., Friis, C., Worning, P., Krogh, A. & Brunak, S. (2001). Genome organisation and chromatin structure in *Escherichia coli*. *Biochimie*, **83**, 201–212.
55. Lathe, W. C., III, Snel, B. & Bork, P. (2000). Gene context conservation of a higher order than operons. *Trends Biochem. Sci.* **25**, 474–479.

Edited by I. B. Holland

(Received 17 March 2003; received in revised form 17 June 2003; accepted 20 June 2003)

SCIENCE @ DIRECT®
www.sciencedirect.com

Supplementary Material for this paper comprising three Figures is available on Science Direct