Genomics 91 (2008) 538–543

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Method

Ecophysiological significance of scale-dependent patterns in prokaryotic genomes unveiled by a combination of statistic and genometric analyses

Juan A.L. Garcia [a], Frederic Bartumeus [a], David Roche [b], Jesús Giraldo [b], H. Eugene Stanley [c], Emilio O. Casamayor [a,*]

[a] Department of Continental Ecology–Limnology, Centre d'Estudis Avançats de Blanes–CSIC, E-17300 Blanes, Girona, Spain
[b] Grup Biomatemàtic de Recerca, Institut de Neurociències, Unitat de Bioestadística, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain
[c] Center for Polymer Studies, Department of Physics, Boston University, Boston, MA 02215, USA

ARTICLE INFO

Article history:
Received 28 August 2007
Accepted 1 March 2008
Available online 16 April 2008

Keywords:
Prokaryote genome
Microbial biodiversity, DNA sequence analysis
DNA walk
Detrended fluctuation analysis
Microbial ecology
Genometrics
Long-range correlation
Comparative genome analyses

ABSTRACT

We combined genometric (DNA walks) and statistical (detrended fluctuation analysis) methods on 456 prokaryotic chromosomes from 309 different bacterial and archaeal species to look for specific patterns and long-range correlations along the genome and relate them to ecological lifestyles. The position of each nucleotide along the complete genome sequence was plotted on an orthogonal plane (DNA landscape), and fluctuation analysis applied to the DNA walk series showed a long-range correlation in contrast to the lack of correlation for artificially generated genomes. Different features in the DNA landscapes among genomes from different ecological and metabolic groups of prokaryotes appeared with the combined analysis. Transition from hyperthermophilic to psychrophilic environments could have been related to more complex structural adaptations in microbial genomes, whereas for other environmental factors such as pH and salinity this effect would have been smaller. Prokaryotes with domain-specific metabolisms, such as photoautotrophy in Bacteria and methanogenesis in Archaea, showed consistent differences in genome correlation structure. Overall, we show that, beyond the relative proportion of nucleotides, correlation properties derived from their sequential position within the genome hide relevant phylogenetic and ecological information. This can be studied by combining genometric and statistical physics methods, leading to a reduction of genome complexity to a few useful descriptors.

Prokaryotes constitute, by far, the largest reservoir of life and encompass the major part of physiological and phylogenetic diversity. A large number of studies have been devoted to exploring microbial biodiversity by 16S rRNA analyses (e.g., [1] and references therein) and, recently, with genomic tools (e.g., [2]). The present capacity to produce genomic information from both laboratory cultures and complex microbial field assemblages widely surpasses the available technical and intellectual skills to analyze and interpret such huge amounts of data into an ecological and evolutionary context. Due to the present size and constantly increasing rate of new raw data, microbiologists and microbial ecologists need new and integrative ways of thinking about microbial genomes to check quickly for similarities and differences among them and to explore and track interactions among genotypes, phenotypes, and the environment. Several authors have recently highlighted the need for new computational tools to analyze and interpret the large amount of nucleotide sequences available in databases [2–4]. Genes contained in genomes provide essential information for understanding evolutionary relationships and ecological adaptations in microorganisms and, although there is a wide repertoire of bioinformatics tools, both further manual checking and lack of close relatives in databases are the main limitations. Conversely, genome size and GC content are two integrative parameters that have been explored by comparative analyses offering interesting information [5–9]. However, DNA is predicted to contain more structural information than would be expected from base composition alone [10].

One of the main features of a DNA sequence related to the whole genome structural composition is the long-range correlation, a scale-invariant property of DNA. In a correlated sequence, occurrence of a nucleotide in a specific position depends on the previous nucleotides (memory). The long-range correlation is related directly to the fractal structure of the DNA sequence or self-similarity. A sequence is defined as self-similar if its fragments can be rescaled to resemble the original sequence itself. Thus, a long-range correlated sequence suggests the existence of repetitive patterns inside it. The search for intrinsic patterns, correlations, and parameters measuring self-similarity by scaling exponents has been carried out in past years by statistical methods [11–16]. Peng et al. [13] studied correlation properties in DNA sequences using a fractal landscape or DNA walk model. DNA walking is a genometric method based on a derivative function of the sequential position for each nucleotide along a DNA sequence. The resulting "walk" can be projected on a two-dimensional plot representative of

* Corresponding author. Fax: +34 972337806.
E-mail address: casamayor@ceab.csic.es (E.O. Casamayor).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.03.001

the DNA "landscape" and enables the simultaneous comparison among different genome landscapes [17,18]. From a different perspective, spectral and fractal analyses have been used to unveil long-range correlations in DNA sequences. Li and Kaneko [19] found long-range correlation by means of spectral analysis in the DNA sequence. Fractal analysis has proven useful for revealing complex patterns in natural objects [20,21], and genome fragments have been classified according to their fractal properties [22]. Finally, a prokaryotic phylogenetic tree based on fractal analyses has been proposed [23].

One of the most appropriated methods proposed in recent years for the study of long-range correlations in genomes is the detrended fluctuation analysis (DFA) [13,24]. DFA is a scaling analysis method providing a single quantitative parameter—the scaling exponent α—to represent correlation properties of a sequence and the characteristic length scale of repetitive patterns. It is a method specifically adapted to handle problems associated with nonstationary sequences. DFA takes into account differences in local nucleotide content (heterogeneity) and can be applied to the entire sequence. It shows linear behavior in log–log plots for all length scales, and the long-range correlation property is characterized by the scaling exponent (α), i.e., the log–log slope. DFA has two clear advantages over other methods. First, it detects long-range correlations embedded in seemingly nonstationary series (conventional methods such as spectral analysis or root mean square fluctuation can be applied reliably only to stationary sequences). Second, it also avoids the spurious detection of apparent long-range correlations that are an artifact of nonstationary sequences and differentiates local patchiness (excess of one type of nucleotide in a specific region) from long-range correlations. Conventional methods such as Markov models have limitations in coping with dependencies at multiple scales, although they are more appropriate for analyzing short-range nucleotide correlations. The case of the fast Fourier transform (FFT) method is strongly affected at high-frequencies analysis by short-range correlations related to codon structure, whereas at low frequencies the signal is distorted by artifacts of the method. The scaling exponent values performed by FFT at midfrequency, however, are close to the values reported by DFA [25].

DFA may help characterize different complex systems according to its different scaling behavior. One of the already shown potentials of DFA is a change in the quantification of genome complexity with evolution [14]. Thus, an increase in the self-similarity (fractal structure) of DNA sequences with evolution has been reported [26], and links between long-range correlations and higher order structure of the DNA molecule have been suggested [27]. It has been shown that scale-independent correlations offer the best compromise between efficient information transfer and immunity to errors on all scales [26], whereas the information theory suggests that one can package the largest amount of information into characters of constant length when a sequence is self-similar [28].

In this work, we propose a combination of DNA walking and DFA methods to help decipher the biological significance of long-range correlations in microbial genomes and the influence of lifestyle in the DNA structure. First, we computed a DNA walk for 456 prokaryotic genome sequences to translate the DNA base sequence into a numerical sequence of Euclidean distances. Next, we used DFA to represent and characterize the correlation properties of the numerical sequence. The specific patterns and long-range correlations were related to phylogenetic, ecological, and metabolic information, providing a combined window to look into prokaryotic genome complexity and microbial biodiversity.

Results and discussion

Within the 456 microbial strains analyzed we covered a wide range of both genome lengths and GC content from several phylogenetic lineages. The range of lengths was between 0.16 Mb in *Candidatus carsonella ruddii* and 9.97 Mb in *Solibacter usitatus*. The percentage of GC content ranged between 16.56% in *Candidatus carsonella ruddii* and 74.90% in *Anaeromyxobacter dehalogenans*. Genome length and percentage of GC content were also heterogeneous within each phylogenetic group. For example, the 33 strains analyzed for Actinobacteria differed by up to one order of magnitude in length, whereas the largest difference in GC content was found within the Gammaproteobacteria (up to fourfold difference). We also covered microorganisms with different ecophysiological lifestyles related with optimal growth temperature, pH, salinity and metabolism, according to information from the taxonomy database at NCBI (www.ncbi.nlm.nih.gov) and *Bergey's Manual of Systematic Bacteriology* [29]. For more details see Supplementary Tables S1 and S2.

DNA walk architecture

For each genome we run an SW (strong–weak pairing) DNA walk and a 2D (two-dimensional) DNA walk (see Methods) as reported in previous works [11,18]. Because a direct relationship exists between %GC and slope in the SW plot (correlation coefficient 0.998), SW slopes were used as the equivalent variable for the percentage of G+C bases: positive slopes indicated dominance of GC, whereas negative slopes reflected the opposite trend. The complete set of genomes fit the previously reported assumption that large genomes have a tendency to be richer in GC [30–32] and therefore they showed higher SW slopes (Supplementary Tables S3 and S4). This has been related to the fact that random mutations are mainly from C to T and from G to A and to the lack of repair mechanisms in reduced genomes that would lead to a TA enrichment [30,31].

The 2D DNA walk for the complete set of genomes was also within the expected results [11,33]. These plots are characterized by the so-called mutational strand bias [18]. Many microorganisms show a preference for G over C, and T over A, in the leading strand and C over G in the lagging strand because of several factors including proofreading efficiencies for the different types of DNA polymerases ([34,35] and references therein). A simple model for explanation is based on the spontaneous deamination of cytosine that induces mutations from C to T. The rate of this deamination is highly increased in single-stranded DNA, such as the leading strand during DNA replication. This causes prevalence of G over C in the leading strand relative to the lagging strand [18]. Most of the chromosomes analyzed (~80% of total) showed strong strand bias that resulted in a symmetric chromosomal inversion in the 2D DNA walks, in which one-half on the genomic sequence was persistently enriched in two of the bases and the other half was enriched in the complementary ones. Both halves commonly split after an inversion point at which the walk changed direction to return back to the run origin (see an example in Supplementary Fig. S2A). The remaining chromosomes (~20%) showed weak strand asymmetry (Supplementary Fig. S2B). Artificial controls run for the different genomes lost the observed architecture and fit a single linear path (see inner plots in Supplementary Fig. S2).

DFA and biological significance of long-range correlations

The 2D DNA landscapes were translated to a numerical series of Euclidean distances (see Methods) for running the DFA. The resulting curves showed scaling exponents within $\alpha = 0.5417$ (*Brucella melitensis*) the lowest and $\alpha = 0.7714$ (*Methanococcus jannaschii*) the highest (Fig. 1). We found for each prokaryotic genome a specific scaling exponent with small variations among them. In all the cases, DFA scaling exponents were higher than 0.5, indicating persistent long-range correlations (see Methods). DFA run for artificial control genomes always had scaling exponents up to 0.50 as expected for uncorrelated sequences (Fig. 1). Therefore, long-range correlations in the genome landscape indicate the existence of selective pressures modeling the architecture along the whole prokaryotic genomes ([16,23] and references therein).

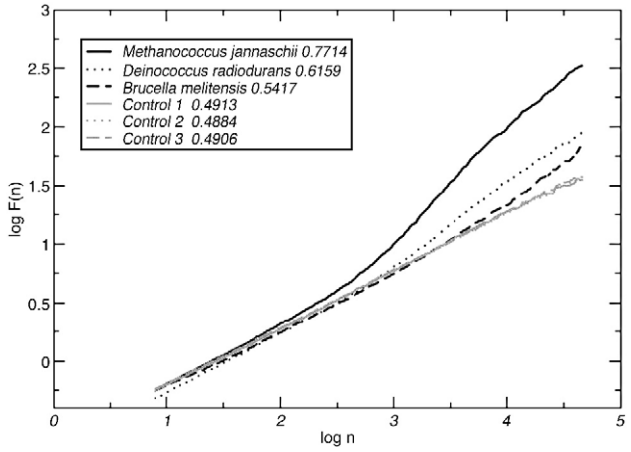

Fig. 1. DFA values calculated on two-dimensional DNA walks for three selected chromosomes and controls. The linear log–log plots of the integrated and detrended time series versus "box size" ($F(n)$ vs n) yielded scaling exponents (value of the slope α) ranging from 0.5417 of *Brucella melitensis* to 0.7714 of *Methanococcus jannaschi.* The three controls (artificial genomes) were close to 0.5 as expected for random sequences. Remaining graphs are available at http//nodens.ceab.csic.es/ecogenomics.

The observed long-range correlations in all the DNA sequences may be due to two factors. On one hand is the elongation of the molecule by repetitive structures added inside the genomes [19]. The fact that long-range correlations were persistent (independent of the scale) means that repetitive structures with different lengths along the genome were present. These repetitive structures may be generated by two possible biological mechanisms important for evolution: first, elongation of sequences by gene duplication [19] and second, elongation and repetition in the genomes by massive lateral transfer of genes from other genomes. On the other hand, long-range correlation can also be related to asymmetric DNA replication along the whole microbial genome, as discussed earlier [36,37].

We found significant differences in scaling exponents (α) between prokaryotes with weak and strong strand bias (t-Student, $p<0.0001$) for the complete set of genomes analyzed (i.e., genomes with weak strand bias consistently had a higher scaling exponent than strong strand-biased genomes). We also found a significant negative correlation between α value and GC content ($R=-0.474$, $p<0.005$) (Supplementary Fig. S3). Weak strand asymmetry has been related to the presence of multiple origins of replication [38] in both Archaea and Bacteria [35,39]. However, along the complete set of genomes we found weak strand asymmetry in archaeal species with single (e.g., *Methanobacterium thermoautotrophicum* and *Archaeoglobus fulgidus*) as well as multiple origins of replication (e.g., *Methanocaldococcus jannaschii* and *Sulfolobus solfataricus*). Conversely, strong strand biases were observed in Archaea with single (*Methanosarcina mazei*) and multiple origins of replication (*Halobacterium* NCR-1). This suggests that processes acting in genomes with weak strand asymmetry are somehow different from those that occur in the other genomes. Weak mutational bias appeared mainly in the genomes from hyperthermophiles and acidophilic microorganisms. It is possible that adaptations to environmental stresses in extremophiles may minimize strand asymmetries. The rates of spontaneous mutation (hydrolytic depurination or hydrolytic deamination) are greatly accelerated at extremely high temperatures [40]. In consequence, hyperthermophiles should have very efficient molecular strategies for repairing DNA under these conditions of chemical instability, because mutation rates in hyperthermophiles are not significantly different from those observed in mesophiles [41].

Grouping genomes by phylogeny and lifestyle

The fact that the raw genome sequence harbors a phylogenetic signal is known [23]. On one hand, over- and underrepresentation of oligonucleotide frequencies have been used by Pride et al. [42] and Teeling et al. [43], and more recently by McHardy et al. [44], for whole-genome phylogeny and classification of genomic fragments. On the other hand, the genomic GC content may change faster than previously thought and seems to be globally and actively influenced by environmental conditions ([8] and references therein). Therefore, the combination of DFA and SW slopes should capture these phylogenetic, ecological, and metabolic signals.

First, we looked for differences at the phylogenetic level. We plotted the combined graph between the DFA scaling exponent and the SW DNA walk slope (Fig. 2B) against the single percentage of each of the four bases (A, T, C, and G) obtained by a PCA (principal components analysis) using the covariance matrix (Fig. 2A). The combination of DFA values (a quantification of the self-similarity or presence of repetitive patterns over all the length scales contained in the genomes) and SW slopes (directly proportional to the GC content) clearly split prokaryotic chromosomes and controls into two different clusters and showed differences between bacterial and archaeal genomes (Fig. 2B). Controls clearly were on the left part with the lowest slopes, close to 0.5, as expected for randomly ordered sequences (the position of one nucleotide was completely uncorrelated with any previous nucleotide),

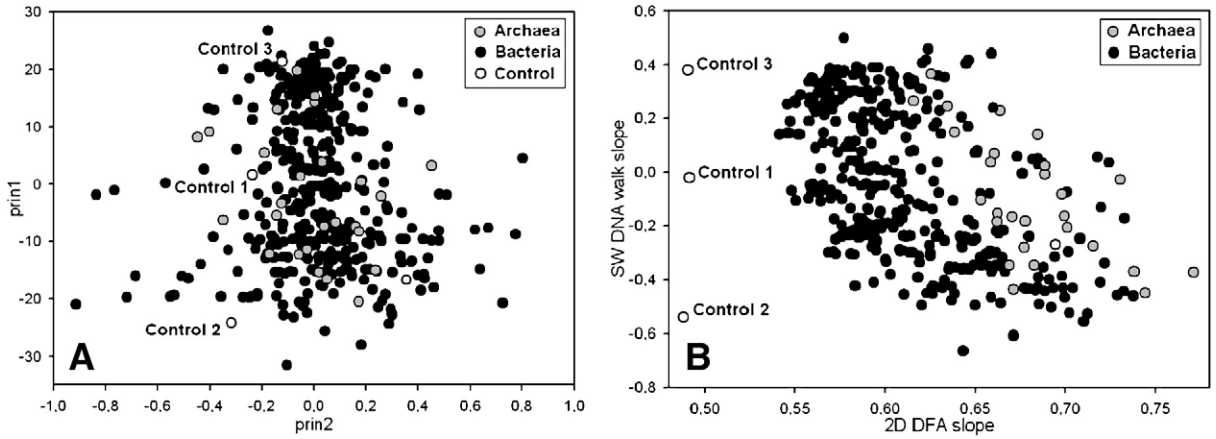

Fig. 2. (A) PCA using percentage of bases for the whole set of genomes analyzed (split into Bacteria and Archaea) and controls (randomly mixed genomes). (B) The same data combination after genometric (SW DNA walk slope on the *y* axis, i.e., a value proportional to the %GC content) and statistical (DFA scaling exponents on the *x* axis, i.e., a measure of the likelihood that one nucleotide will be followed by the same nucleotide) analyses on the entire genomes. Discriminant analysis showed a correct prediction in 96% of Archaea and 85% of Bacteria and 100% in controls (B), whereas they were mixed in the quantitative PCA (A).

and separated along the *y* axis (SW DNA slope) in agreement with their GC content. On average, Archaea had the highest scaling exponents (DFA slopes>0.62) and were located on the right part of the plot. Bacteria appeared mainly in the middle zone of the plot (DFA slopes between 0.54 and 0.74). Discriminant analysis showed a correct prediction in 96% of Archaea and 85% of Bacteria. Conversely, Archaea and Bacteria, as well as the control genomes, were mixed in the quantitative PCA (Fig. 2A).

Second, we focused on ecological lifestyle, and some of the groups clustered separately according to DFA and SW slope values (Fig. 3). For instance, looking at the optimal growth temperature (T_{opt}), hyperthermophiles showed higher scaling exponents than thermophiles and psychrophiles. The three thermophiles placed within the hyperthermophiles were microorganisms with the highest T_{opt} within their group (close to 60 °C). Psychrophiles were discriminated according to GC content in the low scaling exponent values region (Fig. 3A). The discriminant analysis correct prediction was 79% for hyperthermophiles, 80% for thermophiles, and 100% for psychrophiles. Correlation between T_{opt} and GC content in prokaryotes has been the focus of a recent controversy. Musto et al. [9,45] found in a limited number of genomes (ca. 20 genomes) that GC content increased at higher T_{opt}. Conversely, several authors [6,7,46,47] concluded that high GC content is not an adaptation to high temperatures and argued that the correlation between both variables is not robust. The data calculated in our survey (456 microbial genomes) indicate that a tendency to the low GC content exists in hyperthermophiles, but examples of genomes with high GC content are present as well. The decrement of GC content in parallel with T_{opt} is very clear between thermophiles and psychrophiles. Thus, it appears that the transition from a hyperthermophilic to a psychrophilic environment would imply a structural adaptation in microbial genomes both in the GC content and in the sequential position of the nucleotides along the genome.

We also observed various clusters related to salinity and pH (Fig. 3B). Halophiles showed low scaling exponents (<0.65) and high GC content. Conversely, most acidophiles presented high scaling exponents and low GC content, although examples of lower DFA values and higher GC contents were also detected. Alkalophiles showed intermediate values of both DFA slopes and GC contents. Therefore pH itself does not seem to have enough separation power. The true prediction calculated using discriminant analysis was 75% for acidophiles, 83% for alkalophiles, and 87% for halophiles. Most of the acidophiles were hyperthermophilic Archaea and a biased effect with temperature and phylogeny may be present in these cases. In fact, the acidophilic thermophilic bacterium *Acidothermus cellulolyticus* showed low scaling exponent (0.58) and high GC content, in agreement with moderate thermophiles. This example illustrates that temperature is an environmental factor that might have stronger influence in the microbial genomic structure than pH. Another outlier was the genome of the alkalophilic and moderate halophilic bacterium *Natronomonas pharaonis*. This genome shows higher GC content than the remaining alkalophiles and again pH would have smaller influence on the genomic structure than another environmental factor such as salinity. Finally, photoautotrophs and methanogens were classified into two distinct groups with no overlap in their respective DFA slopes (Fig. 3C). Discriminant analysis showed a

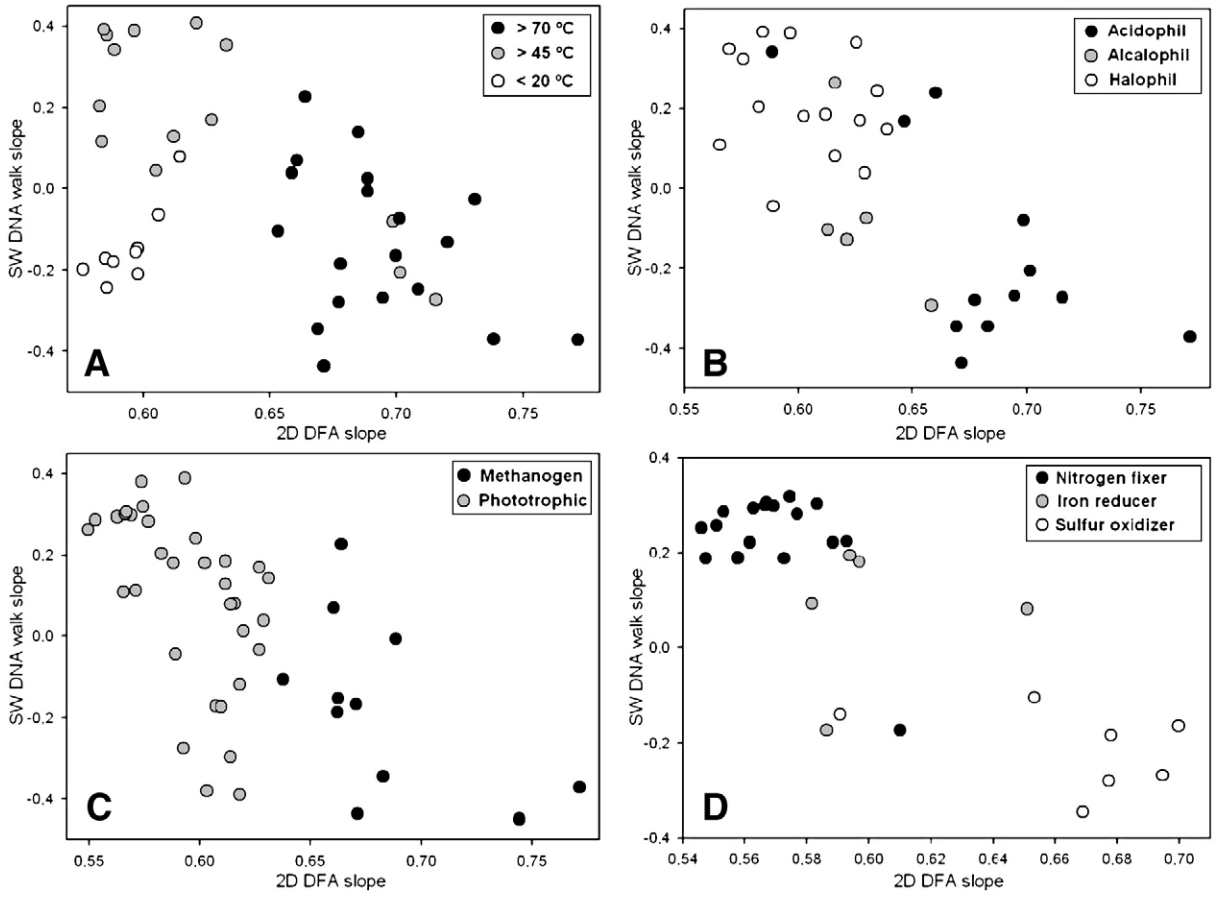

Fig. 3. Ecological and metabolic clusters detected after the SW DNA walk slope (a value proportional to the %GC content) and the DFA scaling exponents (a measure of the likelihood that one nucleotide will be followed by the same nucleotide) analyses on the entire genomes are plotted in combination. For all plots, the GC content increases on the *y* axis proportional to the SW DNA walk slope value. The scaling exponents (α value) represented on the *x* axis are a measure of persistent, long-range correlations in the DNA sequence for each genome. Note that the long-range correlations obtained from the available genomes (A) increased with temperature and (B) decreased with pH. Moreover, (C) methanogens (anaerobic Archaea) showed higher values than phototrophs (oxygenic bacteria), whereas (D) sulfur oxidizers presented on average higher long-range correlations than nitrogen fixers. (A) and (C) plots yielded the highest statistical confidence after discriminant analysis (see text for details).

correct prediction in 91% of methanogens and 100% of phototrophs. Photoautotrophy is an exclusive bacterial metabolism that implies complex enzymatic pathways and no representatives with similar photosystem equipment have been described within Archaea. On the other hand, methanogenesis is a feature present only in the archaeal world.

Similarly, nitrogen fixers and sulfur oxidizers showed opposite behavior in both DFA and SW DNA slopes (Fig. 3D), although both Bacteria and Archaea are able to carry out both processes. Discriminant analysis showed correct prediction of 94% for nitrogen fixers, 80% for iron reducers, and 86% for sulfur oxidizers. In fact, we detected two outliers from the general trend shown by both clusters, one from each: first, the cyanobacterium *Nostoc,* which located away from the remaining nitrogen fixers at the center bottom of the graph (higher DFA slope and lower GC content than the remaining bacteria, mostly from soils), and second, the mesophilic bacterium *Thiomicrospira crunogena,* which separated from the remaining sulfur oxidizers (all of them Archaea and thermophilic). Therefore, these conclusions could be biased for the limited number of nitrogen-fixer and sulfur-oxidizer genomes still available, but it seems that phylogeny and T_{opt} have stronger influence than these metabolic features in the genomic properties detected.

Overall, the combination of genometrics and physical statistic methods captured intrinsic ecological and phylogenetic patterns present in the likelihood that one nucleotide will be followed by the same nucleotide along the entire prokaryotic genome, offering clues to deciphering their biological significance. Although the application of fractal and time series analyses (e.g., self-similarity and fractional dimensionality) to genome data has been carried out for several years already, these techniques have not seen broad usage in genomics. The application of self-similarity parameters as a measure of persistent, long-range correlations in the DNA sequence relative to different ecophysiological lifestyles and other biological parameters (J.A.L. Garcia, A. Fernández-Guerra, and E.O. Casamayor, manuscript in preparation) would help to link physicists and statisticians' approaches with genomic microbiology aims. This work and other recent approaches (e.g., [8,44]) will provide microbial ecologists new tools for a better understanding of the naturally occurring genomic structure and variation and, together with detailed studies of the gene content, may help them to follow and understand the genetic adaptations to specific environments and the magnitude of the genetic reservoir present in the microbial world.

Methods

Four hundred fifty-six completely sequenced closed genomes from 309 different species of prokaryotes were downloaded from GenBank (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/Genbank) in May 2007. The prospected genomes belonged to three archaeal kingdoms (28 chromosomes) and 20 bacterial classes (for more details see Supplementary Tables S1 and S2). The run for DNA walks started at position 0 of the annotated sequence. For comparative purposes, we constructed several artificial genomes as controls by randomly mixing the order of bases from original genomes. For instance, for controls 1, 2, and 3 the following conditions were chosen. Control 1 had the 1,197,687 bases of *Anaplasma marginale* in random order and 50% GC content. Control 2 was the randomly ordered strain of *Mycoplasma mycoides* (1,211,703 bases in length and 24% GC content). Finally, control 3 was rich in GCs with the same length (1,849,735 bases) and GC percentage (70%) as the *Thermus thermophilus* chromosome.

DNA walks

We analyzed the sequential distribution of individual nucleotides along the genomes by the DNA walk method [18,33,48,49]. DNA walks are graphical representations of the fluctuations in nucleotide series and provide quantification on internal deviations of individual nucleotides along the genome. Every genome produces a specific DNA walk and there are several possibilities and rules for plotting genomic landscapes [33]. Here, we have used two types of representation. First, we translated the original nucleotide sequence onto a one-dimensional numerical series grouping the bases in pairs following the hydrogen bond energy rule (SW for strong–weak pair): n_i being the ith nucleotide of the genomic sequence and y_i the DNA walk value for the nucleotide n_i; if n_i is a strongly bonded pair (G or C), then $y_i=+1$, and if n_i is a weakly bonded pair (A or T), then $y_i=-1$. We mapped the resulting SW DNA walk series onto an orthogonal plane (see Supplementary Fig. S1). The SW walks were fitted by linear regression and the slopes were used as variables for subsequent analysis (SW DNA walk slope). For the second representation, we performed a 2D map in which each nucleotide defines one direction in a plane formed by two orthogonal axes (i.e., C versus G and T versus A). In this walk, the walker moves 1 unit onto the plane according to the four senses defined by the nucleotide read. This 2D DNA walk generates an irregular graph resembling a fractal landscape (see the example in Supplementary Fig. S2). The defining feature of the landscape is the statistical self-similarity of the plots obtained at various magnifications calculated with the DFA method.

Detrended fluctuation analysis

DFA is a scaling analysis method providing a simple integrative parameter—the scaling exponent α—to represent the correlation properties of numerical series. The scaling exponent is also called the self-similarity parameter. An object is self-similar if its subsets can be rescaled to resemble statistically the original object itself. A numerical sequence is considered stationary if the mean, standard deviation and correlation functions are invariant under space translation [13,14,24]. Sequences that do not fit these conditions are nonstationary. DFA allows detection of long-range correlations embedded in seemingly nonstationary series, and it avoids the spurious detection of apparent long-range correlations that are an artifact of nonstationarity [50]. The scaling exponent quantifies the amount and range of the correlations. In a given sequence, a change in the scaling exponent indicates changes in the correlations through different scales.

Scaling exponents were calculated from the 2D DNA walks, using Euclidean distances from the origin of the graph to every x–y point representing a step of the walk, as follows. First, the entire sequence of length N, understood as the Euclidean distances for each step of the walk, was divided into boxes of equal length, n, each containing l steps of the walk. We defined the "local trend" in each box by fitting a least-squares linear model (proportional to the compositional bias in the box) to the data. Second, we defined the "detrended walk" as the difference between the original walk $y(n)$ and the local trend. Next, we calculated both the variance on the detrended walk for each box and the average of these variances over all the boxes of size l, denoted $F(n)$. Such computation was repeated over all time scales (box sizes) to provide a relationship between $F(n)$ and the box size n. Typically, $F(n)$ increases with box size n. A linear relationship on a log–log graph indicates the presence of scaling (long-range correlations). Obtaining linear log–log plots of the integrated and detrended time series versus "box size" ($F(n)$ vs n) can help to establish the appropriateness of the DFA method to all nonstationary data encountered. Under these conditions, fluctuations can be characterized by the scaling exponent (α), i.e., the slope of the line relating log $F(n)$ to log n. The minimum box size (n_{min}) does not depend on N. On the contrary, the maximum box size (n_{max}) scales as $n_{max}=N/10$ [50].

For an ideal sequence of infinite length, $\alpha=0.5$ indicates the absence of long-range correlation (random walk), where the value of one nucleotide is completely uncorrelated with any previous values, whereas α different from 0.5 indicates long-range correlation (see [51] for more details on the method). For a sample of finite length, statistical fluctuations due to finite size should be taken into account. Therefore, we considered a DNA sequence to exhibit long-range correlation only if a value was significantly different from the α value of the random finite control sequences. The α values in the range $0.5<\alpha<1$ indicate persistent long-range power–law correlations suggesting the existence of repetitive patterns in the sequence and that finding a particular nucleotide on a sequential position depends on the previous nucleotides.

Finally, discriminant analysis [52] was used to construct the Fisher discriminant function (a linear combination of the variables whose coefficients make maximum the distance between the populations) for species classification into one of two or more groups on the basis of the 2D DFA slope and SW DNA walk slope variables. Computations were carried out with SAS/STAT release 9.1 statistical package (SAS Institute, Inc., Cary, NC, USA).

Acknowledgments

This work was financed by Projects VIARC REN2003-08333 and CRENYC CGL2006-12058 to E.O.C. and Grant SAF2004-06134 to J.G. from the Spanish Ministerio de Educación y Ciencia, Marine Genomics Europe Network of Excellence Grant GOCE-CT-2003-505403 from the EU-FP6 (to E.O.C.), and Thematic Network on Environmental Microbial Genomics Grant 2004-XT-00012 from the Autonomous Government of Catalonia (to E.O.C.). We are very grateful to the staff at the Department of Physics, Boston University, for their help in running calculations and to the Centre de Supercomputació de Catalunya (www.cesca.es) for supercomputing facilities. Constructive comments from the reviewers are also acknowledged. J.A.G. is supported by an FI predoctoral scholarship and a travel grant to Boston from the Catalonian Departament d'Universitats, Recerca i Societat de la Informació, and by Marine Genomics Europe training courses on

bioinformatics. E.O.C. is a Fellow of the Program Ramon y Cajal from the Spanish Ministerio de Educación y Ciencia and FEDER. All the data and graphs generated are freely available by request at our Web site: http://nodens.ceab.csic.es/ecogenomics.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.03.001.

References

[1] E.O. Casamayor, C. Pedrós-Alió, G. Muyzer, R. Amann, Microheterogeneity in 16S rDNA-defined bacterial populations from a stratified planktonic environment is related to temporal succession and to ecological adaptations, Appl. Environ. Microbiol. 68 (2002) 1706–1714.
[2] E.F. De Long, Microbial population genomics and ecology: the road ahead, Environ. Microbiol. 6 (2004) 875–878.
[3] K.E. Nelson, The future of microbial genomics, Environ. Microbiol. 5 (2003) 1223–1225.
[4] W.R. Streit, R.A. Schmitz, Metagenomics—the key to the uncultured microbes, Curr. Opin. Microbiol. 7 (2004) 492–498.
[5] A. Muto, S. Osawa, The guanine and cytosine content of genomic DNA and bacterial evolution, Proc. Natl. Acad. Sci. 84 (1987) 166–169.
[6] L.D. Hurst, A.R. Merchant, High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes, Proc. R. Soc. Lond., B Biol. Sci. 268 (2000) 493–497.
[7] S.A. Marashi, Z. Ghalanbor, Correlations between genomic GC levels and optimal growth temperatures are not 'robust', Biochem. Biophys. Res. Commun. 325 (2004) 381–383.
[8] K.U. Foerstner, C.V. Mering, S.D. Hooper, P. Bork, Environments shape the nucleotide composition of genomes, EMBO Rep. 6 (2005) 1208–1213.
[9] H. Musto, H. Naya, A. Zavala, H. Romero, F. Alvarez-Valín, G. Bernardi, Genomic GC level, optimal growth temperature, and genome size in prokaryotes, Biochem. Biophys. Res. Commun. 347 (2006) 1–3.
[10] A.G. Pedersen, L.J. Jensen, S.B. Hans-Henrik Staerfeldt, D.W. Ussery, A DNA structural atlas for Escherichia coli, J. Mol. Biol. 299 (2000) 907–930.
[11] C.A.H. Roten, P. Gamba, J.L. Barblan, D. Karamata, Comparative Genometrics (CG): a database dedicated to biometric comparisons of whole genomes, Nucleic Acids Res. 30 (2002) 142–144.
[12] P. Bernaola-Galvan, P. Carpena, R. Román-Roldán, J.L. Oliver, Study of statistical correlations in DNA sequences, Gene 300 (2002) 105–115.
[13] C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Long-range correlations in nucleotide sequences, Nature 356 (1992) 168–170.
[14] C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M. Simons, H.E. Stanley, Statistical properties of DNA sequences, Physica, A 221 (1995) 180–192.
[15] C.A. Chatzidimitriou-Dreismann, D. Larhammar, Long-range correlations in DNA, Nature 361 (1993) 212–213.
[16] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, C.K. Peng, M. Simons, Scale invariant features of coding and noncoding DNA sequences, in: P. Iannaconne, M.K. Khokha (Eds.), Fractal Geometry in Biological Systems: an Analytical Approach, CRC Press, Boca Raton, FL, 1996, pp. 15–30.
[17] R.C. Elston, A.F. Wilson, Genetic linkage and complex disease: a comment, Genet. Epidemiol. 7 (1990) 17–19.
[18] J.R. Lobry, Genomic landscapes, Microbiol. Today 26 (1999) 164–165.
[19] W. Li, K. Kaneko, Long-range correlation and partial $1/f^{\alpha}$ spectrum in a non-coding DNA sequence, Europhys. Lett. 17 (1992) 655–660.
[20] C.L. Berthelsen, J.A. Glazier, M.H. Skolnick, Global fractal dimension of human DNA sequences treated as pseudorandom walks, Phys. Rev., A 45 (1992) 8902–8913.
[21] M.S. Vieira, Statistics of DNA sequences: a low-frequency analysis, Phys. Rev., E 60 (1999) 5932–5937.
[22] V.V. Anh, K.S. Lau, Z.G. Yu, Recognition of an organism from fragments of its complete genome, Phys. Rev., E 66 (2002) 031910.
[23] Z.G. Yu, V. Anh, K.S. Lau, K.H. Chu, The genomic tree of living organisms based on a fractal model, Phys. Lett., A 317 (2003) 293–302.
[24] C.K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Mosaic organization of DNA nucleotides, Phys. Rev., E 49 (1994) 1685–1689.
[25] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.K. Peng, M. Simons, H.E. Stanley, Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis, Phys. Rev., E 51 (1995) 5084–5091.
[26] R.F. Voss, Evolution of long-range fractal correlations and $1/f^{\alpha}$ noise in DNA base sequences, Phys. Rev. Lett. 68 (1992) 3805–3808.
[27] A. Grosberg, Y. Rabin, S. Havlin, A. Neer, Crumpled gobule model of the three-dimensional structure of DNA, Europhys. Lett. 23 (1993) 373–378.
[28] N. Nagai, K. Kuwata, T. Hayashi, H. Kuwata, S. Era, Evolution of the periodicity and the self-similarity in DNA sequence: a Fourier transform analysis, Jpn. J. Physiol. 51 (2001) 159–168.
[29] G.M. Garrity, D.R. Boone, R.W. Castenholz, Bergey's Manual of Systematic Bacteriology, 2nd ed. Springer, Berlin, 2001.
[30] A. Heddi, H. Charles, C. Khatchadourian, G. Bonnot, P. Nardon, Molecular characterization of the principal symbiotic bacteria of the weevil Sitophilus oryzae: a peculiar G+C content of an endocytobiotic DNA, J. Mol. Evol. 47 (1998) 52–61.
[31] N.A. Moran, Microbial minimalism: genome reduction in bacterial pathogens, Cell 108 (2002) 583–586.
[32] E.P. Rocha, A. Danchin, Base composition bias might result from competition for metabolic resources, Trends Genet. 18 (2002) 291–294.
[33] A. Grigoriev, Analyzing genomes with cumulative skew diagrams, Nucleic Acids Res. 26 (1998) 2286–2290.
[34] E.P.C. Rocha, Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol. 10 (2000) 393–395.
[35] P. Worning, L.J. Jensen, P.F. Hallin, H.H. Stærfeldt, D.W. Ussery, Origin of replication in circular prokaryotic chromosomes, Environ. Microbiol. 8 (2006) 353–361.
[36] W. Li, T. Marr, K. Kaneko, Understanding long-range correlations in DNA-sequences, Physica, D 75 (1994) 392–416.
[37] P. Mackiewicz, M. Kowalczuk, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, A. Laszkiewicz, M. Dudek, S. Cebrat, Replication associated mutational pressure generating long-range correlation in DNA, Physica, A 314 (2002) 646–654.
[38] J. Mrázek, S. Karlin, Strand compositional asymmetry in bacterial and large viral genomes, Proc. Natl. Acad. Sci. 95 (1997) 3720–3725.
[39] L.M. Kelman, Z. Kelman, Multiple origins of replication in archaea, Trends Microbiol. 12 (2004) 399–401.
[40] T. Lindahl, Instability and decay of the primary structure of DNA, Nature 362 (1993) 709–715.
[41] K.L. Jacobs, D.W. Grogan, Rates of spontaneous mutation in an archaeon from geothermal environments, J. Bacteriol. 197 (1997) 3298–3303.
[42] D.T. Pride, R.J. Meinersmann, T.M. Wassenaar, M.J. Blaser, Evolutionary implications of microbial genome tetranucleotide frequency biases, Genome Res. 13 (2003) 145–158.
[43] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, F.O. Glockner, Application of tetranucleotide frequencies for the assignment of genomic fragments, Environ. Microbiol. 6 (2004) 938–947.
[44] A.C. McHardy, H.G. Martin, A. Tsirigos, P. Hugenholtz, I. Rigoutsos, Accurate phylogenetic classification of variable-length DNA fragments, Nat. Methods 4 (2007) 63–72.
[45] H. Musto, H. Naya, A. Zavala, H. Romero, F. Alvarez-Valín, G. Bernardi, Correlations between genomic GC levels and optimal growth temperatures in prokaryotes, FEBS Lett. 573 (2004) 73–77.
[46] N. Galtier, J.R. Lobry, Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes, J. Mol. Evol. 44 (1997) 632–636.
[47] H.C. Wang, E. Susko, A.J. Roger, On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors, Biochem. Biophys. Res. Commun. 342 (2006) 681–684.
[48] J.R. Lobry, A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria, Biochimie 78 (1996) 323–326.
[49] C. Cebrat, M.R. Dudek, The effect of DNA phase structure on DNA walks, Eur. Phys. J., B Cond. Matter Phys. 3 (1998) 271–276.
[50] K. Hu, P.C. Ivanov, Z. Chen, P. Carpena, H.E. Stanley, Effect of trends on detrended fluctuation analysis, Phys. Rev., E 64 (2001) 011114.
[51] B.B. Mandelbrot, J.W. Van Ness, Fractional Brownian motions, fractional noises and applications, SIAM Rev. 10 (1968) 422–437.
[52] A.A. Afifi, V.A. Clark, S. May, Discriminant analysis, Computer-Aided Multivariate Analysis, Chapman & Hall/CRC, New York, 2004, pp. 249–279.