Molecular Phylogenetics and Evolution 61 (2011) 650–658

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

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

A close relationship between primary nucleotides sequence structure and the composition of functional genes in the genome of prokaryotes

Juan A.L. Garcia *, Antoni Fernández-Guerra, Emilio O. Casamayor

Department of Continental Ecology-Biogeodynamics & Biodiversity Group, Centre d'Estudis Avançats de Blanes, CEAB–CSIC, E-17300 Blanes, Girona, Spain

ARTICLE INFO

Article history:
Received 29 September 2010
Revised 31 May 2011
Accepted 5 August 2011
Available online 16 August 2011

Keywords:
DNA walk
Detrended fluctuation analysis (DFA)
Microbial ecology
Long-range correlation
Comparative genome analyses
Large-scale genome structure

ABSTRACT

Comparative genomics is an essential tool to unravel how genomes change over evolutionary time and to gain clues on the links between functional genomics and evolution. In prokaryotes, the large, good quality, genome sequences available in public databases and the recently developed large-scale computational methods, offer an unprecedent view on the ecology and evolution of microorganisms through comparative genomics. In this work, we examined the links among genome structure (i.e., the sequential distribution of nucleotides itself by detrended fluctuation analysis, DFA) and genomic diversity (i.e., gene functionality by Clusters of Orthologous Genes, COGs) in 828 full sequenced prokaryotic genomes from 548 different bacteria and archaea species. DFA scaling exponent α indicated persistent long-range correlations (fractality) in each genome analyzed. Higher resolution power was found when considering the sequential succession of purine (AG) vs. pyrimidine (CT) bases than either keto (GT) to amino (AC) forms or strongly (GC) vs. weakly (AT) bonded nucleotides. Interestingly, the phyla Aquificae, Fusobacteria, Dictyoglomi, Nitrospirae, and Thermotogae were closer to archaea than to their bacterial counterparts. A strong significant correlation was found between scaling exponent α and COGs distribution, and we consistently observed that the larger α the more heterogeneous was the gene distribution within each functional category, suggesting a close relationship between primary nucleotides sequence structure and functional genes composition.

1. Introduction

Comparative genetic analyses have become common practice to obtain clues on the links between functional genomics, evolution, and physiology following the accelerating discovery rate of new genomic information from biological species and environmental samples (Koonin and Wolf, 2008; Bentley and Parkhill, 2004). In a different approach, long-range correlations in DNA (i.e., an indicator of the self-similarity or fractal structure of the genome that may be a consequence of a large-scale genome structure) can also help to explore the links between genomic diversity and evolution (Voss, 1992, Bernaola-Galván et al., 2002, Garcia et al., 2008) and reduce the genome complexity to a few useful descriptors (Garcia et al., 2008). In the case of microorganisms, this is of special interest since prokaryotes encompass the major part of physiological and phylogenetic diversity, and the number of full sequenced available microbial genomes is by far larger than in other group of organisms (http://www.genomesonline.org). The causes behind the large-scale genome structure are however not fully understood, and it has been suggested that a wide range of evolutionary factors such as insertions, inversions, and horizontal gene transfers might be involved (Bernardi, 2004).

Genes contained in genomes provide essential information for understanding evolutionary relationships, as well as ecological and functional adaptations in microorganisms. Many studies have classified sets of orthologous sequences among different species, and therefore many databases of orthologous groups are available. NCBI Clusters of Orthologous Groups (COG) database for unicellular organisms (Tatusov et al., 2003) contains putative orthologous groups, mostly prokaryotes, and consistent information about orthology provides the basis for inferring phylogenetic relationships (Tatusov et al., 1997). Each COG consists of individual orthologous proteins and its delineation is achieved by both comparison of proteins encoded in different complete genomes from major phylogenetic lineages, and elucidation of consistent patterns of sequence similarities. In order to extract information from large sets of prokaryotic genomes, COGs can classify conserved genes according to their homologous relationships. Orthologs typically have the same function, allowing transfer of functional information from

* Corresponding author. Address: Centre d'Estudis Avançats de Blanes (CEAB–CSIC), Accés Cala St. Francesc, 14, E-17300 Blanes, Girona, Spain. Fax: +34 972337806.
E-mail address: jag@ceab.csic.es (J.A.L. Garcia).

1055-7903/$ - see front matter © 2011 Published by Elsevier Inc.
doi:10.1016/j.ympev.2011.08.011

one member to an entire COG. Each COG represents a functional pathway and changes in COGs content actually will determine changes in the ecological lifestyle and habitat (Sen et al., 2008).

In a recent study, we observed long-range correlation in microbial genomes using detrended fluctuation analysis (DFA, summarized by a scaling exponent significantly higher than random controls) and a close link with microbial lifestyle, mainly in the case of hyperthermophiles (Garcia et al., 2008). However, whether the genomic structures detected were more related to structural mechanisms than to the functional genes content could not be satisfactorily addressed. In fact, multiple origins of replication (Mrázek and Karlin, 1997), asymmetric DNA replication (Mackiewicz et al., 2002), and changes in GC content (Foerstner et al., 2005) may influence the genome structures, as well as the rates of spontaneous mutation (hydrolytic depurination or hydrolytic deamination) that are greatly accelerated at extremely high temperatures (Lindahl, 1993). In the present work, we deeply explored whether or not the emergence of genomic structures had a functional meaning using multivariate correlation analysis between scaling exponents and different structural, phylogenetic and ecophysiological factors covering genomes gene functionality (COGs), in 828 genomes from 548 microbial species. An outstanding correlation emerged between the structure of the prokaryotic genomes—represented by the scaling exponent—and functional genes composition.

2. Material and methods

In this study 828 complete genomes from 548 prokaryotic species obtained from GenBank (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/Genbank, download April 2010) were used. Overall, we analyzed 54 genomes from five main archaea phyla, and 774 genomes from 20 bacteria phyla. These genomes are representative of most of the bacteria and archaea that have cultured representatives. On the whole, we covered a wide range of phylogenies and ecophysiological lifestyles related with different optimal growth temperatures, pH ranges, and respiration and metabolism features. Ancillary data were obtained from the taxonomy database at NCBI (www.ncbi.nlm.nih.gov) and the Bergey's Manual of Systematic Bacteriology (Garrity et al., 2001). For more details see Supplementary material.

2.1. Clusters of orthologous genes (COGs) database

For all genomes, we obtained the distribution of encoded proteins within different functional categories following the COG database (http://www.ncbi.nlm.nih.gov/COG/). COG, is a systematic grouping of gene families where each COG contains individual orthologous proteins or orthologous sets of paralogs from at least three different lineages and therefore matching an ancient conserved domain. Thus, COGs are grouped according to cell functionality characters. The main functional groups were divided in four categories: (1) information storage and processing genes, (2) cellular processes coding genes, (3) metabolic genes, and (4) genes from poorly characterized proteins. Each category contained different subgroups such as genes associated to translation, transcription, and DNA replication for category 1, or genes related with cell division, post-translational modification, cell envelope biogenesis, cell motility, ion transport, and signal transduction mechanisms for category 2. Category 3 included genes associated with energy production, carbohydrate, amino acid, nucleotide, coenzyme and lipid metabolism, respectively. Finally, category 4 allocated genes with unknown functions. These groups were formulated after comparing protein sequences of known sources with those proteins properly annotated from genomes.

2.2. DNA walks

For each genome we plotted the sequential distribution of individual nucleotides following the DNA walk genometric method (Lobry, 1996; Cebrat and Dudek, 1998; Grigoriev, 1998; Lobry, 1999). DNA walks represent the fluctuations in nucleotides series and provide quantification on internal deviations of individual nucleotides along the genome. There are several rules to plot genomic landscapes (Buldyrev et al., 1995). In a previous work, we run a two-dimensional DNA walk (i.e., each nucleotide fits one of the four cardinal directions) to optimize and test the DFA method (Garcia et al., 2008). In the present work, we enlarged the approach looking for higher resolution power and we used three types of one-dimensional plots translating the original nucleotide sequence onto a numerical series as follows. First, we grouped the bases in pairs following the hybrid rule (KM): being n_i the i nucleotide of the genomic sequence and y_i the DNA walk value for the nucleotide n_i, if n_i is a keto forms (G or T) then $y_i = +1$ and if n_i is a amino forms (A or C) then $y_i = -1$. Next, we grouped the bases following the purine–pyrimidine (RY) rule. If n_i is a pyrimidine (C or T) then $y_i = +1$ and if n_i is a purine (A or G) then $y_i = -1$. Finally, the bases were grouped depending on the hydrogen bond energy rule (SW). 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 DNA walk series onto an orthogonal plane. The DNA walk generated an irregular graph resembling a fractal landscape. The defining feature for the landscape is the statistical self-similarity of the plots obtained at various magnifications calculated with the Detrented Fluctuation Analysis (DFA) method. The scaling exponent-α-calculated by DFA represents the correlation properties of the numerical series. The α value within the range $0.5 < \alpha < 1$, indicated persistent long-range power-law correlations, i.e., a particular nucleotide is more likely to be followed by the same nucleotide than by any of the others. $\alpha = 0.5$ indicated absence of long-range correlation where the value of one nucleotide is completely uncorrelated with any previous values. For more details on the application of this method to prokaryotic genomes see Garcia et al. (2008). Fig. 1 represents an example of the plots generated by the three types of DNA walks (KM, RY and SW) for the bacteria *Oceanobacillus iheyensis*, and the scaling exponents associated to each walk. Overall, we calculated three scaling exponents for each genome, following the three types of DNA walk. In addition, we assigned to each genome 3 control variables with random values comprised between the lowest and the highest scaling exponent calculated by DFA using KM, RY and SW rules, respectively. Controls were used to emphasize the influence of factors as phylogeny, ecology or functional genomics in the genome structure.

2.3. Multivariate analysis

For each genome, we defined up to 56 features that were grouped in six different categories: taxonomy, genometry, genome, ecology, COG and DFA (Table 1). Among these features, we included taxonomy level, genome length, percentage of each single nucleotide, number of genomes per species, proteins and RNAs, coding and non-coding length mean, percentage of genes in 5′–>3′, habitat, oxygen requirement, optimal growth temperature (OGT), pH, salinity, COGs, and the three types of DFAs. Using the whole data set we created a multivariate dataset (see Supplementary material). The categorical factors describing the ecology of each microorganism were divided into each of those branches indicated in Table 2.

Two canonical analyses, Canonical Correlation Analysis (CCA) and Redundancy Analysis (RDA), were carried out to explore the relationships between the scaling exponents and the remaining