J Mol Evol (2009) 69:203–206
DOI 10.1007/s00239-009-9230-9

LETTER TO THE EDITOR

Oxygen and Guanine–Cytosine Profiles in Marine Environments

Héctor Romero · Emiliano Pereira ·
Hugo Naya · Héctor Musto

Published online: 25 June 2009

Abstract One of the historic debates in molecular evolution concerns the strong variation in the genomic guanine–cytosine (GC) content of prokaryotes, which ranges from approximately 20–75%: Is this factor selectively neutral, or is it the result of natural selection? In a previous article published by our group, we showed that inside well-defined taxonomic groups of prokaryotes, strictly aerobic organisms tend to display higher genomic GC levels than strictly anaerobic species. In the present study, we examined the GC content of fragments of DNA obtained from microbial communities along a well-defined environmental gradient: a 4,000-m vertical profile in the North Pacific subtropical gyre. The patterns of GC distribution might be associated with oxygen concentrations in the seawater column. These results give further support to the link between a physiologic trait (aerobic respiration) and genomic GC content.

Keywords Aerobiosis · GC content · Metagenomics

Introduction

Genomic guanine–cytosine (GC) content in prokaryotes has been studied for >50 years. Since the pioneering works, which began in 1955, its variability has been an intriguing subject of study (Barbu et al. 1956; Belozersky and Spirin 1958; Sueoka 1962). The first considerations of GC content evolution speculated on the absence of a selective force behind this pattern of GC content variability (Sueoka 1962), a vision that was reinforced by the formulation of the neutral theory of molecular evolution (Kimura 1968; Singer and Ames 1970). Along other lines, selection was believed to play a role in the evolution of GC content, and some hypotheses were proposed within a Darwinian perspective (see later text). Since then, evolution of GC content has been framed within the neutralist-selectionist debate of molecular evolution.

Within the neutral framework, it has been argued that GC content variability should be a consequence of random mutation and fixation, in particular produced by the idiosyncratic biases of DNA replication-repair machinery, and therefore without selective value. In contrast, among the first ideas proposing a selectionist hypothesis, it was suggested that high GC content (low adenine–thymine) may constitute a selective advantage because it could decrease the formation of thymine dimmers, particularly in organisms exposed to ultraviolet radiation (Singer and Ames 1970). Later, also from the selectionist standpoint, it was proposed that increased GC content could be advantageous for thermophilic organisms (Argos et al. 1979). However, the selectionist hypotheses could not be unambiguously demonstrated in some cases, or solid counter-examples were found in other cases. In this context of lack of a clear physiologic or ecologic trait linked to GC content variability, this genomic trait was considered as a bastion of the neutral theory of molecular evolution.

With the advance of the "genomic era," plenty of high-quality data became available, and several works presented results compatible with selectionist hypotheses. For example, Rocha and Danchin (2002) proposed that intracellular

H. Romero (✉) · E. Pereira · H. Musto
Laboratorio de Organización y Evolución del Genoma, Secc.
Biomatemática Facultad de Ciencias, Universidad de la
República, Iguá 4225, Montevideo 11400, Uruguay
e-mail: eletor@fcien.edu.uy

H. Naya
Unidad de Bioinformática, Institut Pasteur de Montevideo,
Mataojo 2020, Montevideo 11400, Uruguay

parasitism may be linked to GC content. Our group found that strict aerobes were more GC-rich than strict anaerobes (Naya et al. 2002) and presented evidence that optimal growth temperature may be linked to higher GC values within prokaryotic families (Musto et al. 2004).

The above-mentioned analyses were confined to culturable prokaryotes. However, in recent years the field of environmental genomics has opened up new possibilities of studying unculturable organisms. Indeed, it is now possible to sequence tens of thousands of DNA fragments from preparations obtained directly from the environment (see Venter et al. 2004; Tyson et al. 2004; Tringe et al. 2005). Among the new possibilities fostered by this approach are the study of unculturable organisms; the search for new gene and protein functions; and, more interestingly, a new basis for proposing and testing particular ecologic and evolutionary hypotheses (see Kunin et al. 2008 and Tringe and Rubin 2005 for reviews).

The case of GC is no exception because data generated in environmental genomics studies constitute a unique opportunity to test the effect of specific environmental conditions on this particular molecular trait. Pioneering this approach, Foerstner et al. (2005) examined the GC content distribution of DNA fragments from different environments and concluded that the environment plays an important role in determining the GC content of corresponding microbial communities.

In the current work, we applied this approach to samples obtained during a genomic survey of microbial communities along a well-defined environmental gradient: a 4,000-m vertical profile in the North Pacific subtropical gyre obtained by DeLong et al. (2006). One particularity of this data set is the fact that high-quality measures of physical, chemical, and biologic variables are available together with the DNA sequences. In this context, clues to the evolutionary forces behind the evolution of GC content are presented that reinforce previous results.

GC Content in the Water Column of a North Pacific Ocean

We examined 64,610 reads (length >500 nt) from the different depths sampled in DeLong's work (10, 70, 130, 200, 500, 770, and 4,000 m). GC content distribution of the fragments for each depth is displayed in Fig. 1a, and similarities and differences are clearly illustrated by the histograms. The general shape of the first three samples (10 m, 70 m, and 130 m) is rather similar, with a slight negative skew and a major peak at approximately 50% GC. The 200-m sample is also fairly symmetric, with a peak at approximately 50% GC. In contrast, 500-m and 4,000-m samples display bimodal distributions, with a major peak at approximately 55% GC and a minor peak at approximately 42% GC. Interestingly, the 770-m sample is also bimodal, with coincident modes (42% and 55%) but with a notable increase in the frequency at approximately 42%, which surpasses the peak at 55% GC.

Functional cluster analyses based on Clusters of Orthologous Groups (Tatusov et al. 2003) and KEGG (Kanehisa et al. 2008) classifications carried out by DeLong et al. (2006) grouped the 10-m, 70-m, and 130-m samples (corresponding to the photic zone) on one side and the 500-m, 770-m, and 4,000-m samples on the other side (the 200-m sample was left apart). As discussed previously, the first group also showed similar distributions in GC content. Therefore, these groups are not good candidates with which to study changes in GC values associated with environmental variation. In contrast, the distributions of the 500-m, 770-m, and 4,000-m samples showed noticeable differences. However, because they were clustered in one group, we decided to perform a thorough analysis of these three samples.

First, we tried to dissect the distributions. The rationale behind our procedure was that particular peaks could be the consequence of the overrepresentation of certain type of

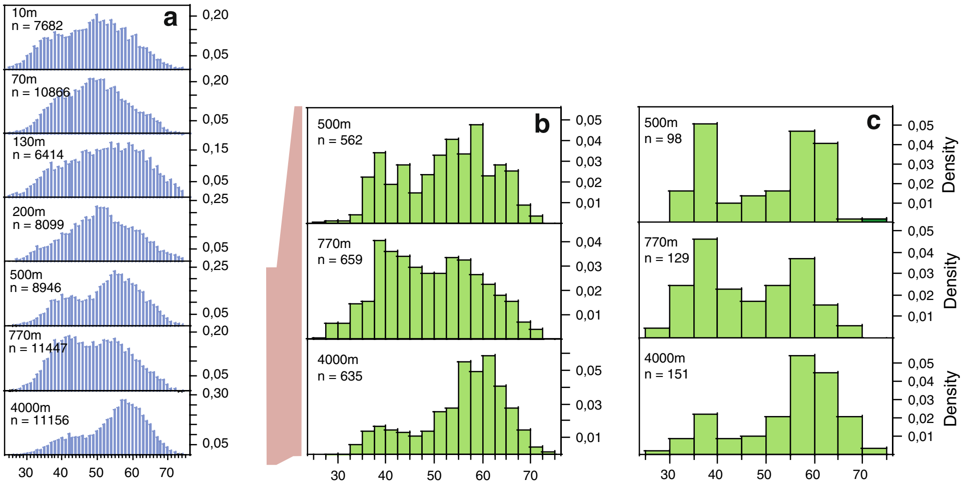

Fig. 1 **a** GC content distribution of each depth for all of the sequences. **b** GC content distribution of the three deepest sampling points constructed with the "informational core" of bacteria. **c** The same as **b** but for archaea. *n*—number of fragments analyzed

sequences. For example, the abundance of some genes or gene families (represented by similar sequences) may produce the observed peaks. In addition, the abundance of one, or a few, prokaryotic species dominating the community may also produce these peaks.

To investigate the first possibility, we conducted a BLASTX (Altschul et al. 1997) search (BLOSUM 45, e-value threshold 1e–15) against a "nonredundant" (approximately 80% identity) database of proteins from 625 completely sequenced genomes related to basic informational processes: translation (e.g., aminoacyl-tRNA synthetases, elongation factors, ribosomal proteins, etc.), transcription (subunits of the RNA-polymerase), and replication (proteins of the replication complex). In Fig 1b and c, the GC distributions for the 500-m, 770-m, and 4,000-m samples are shown for the sequences having a positive hit using this database, which classified each fragment into bacterial or archaeal origin if the first five BLAST hits included one of these domains. Remarkably, the distribution of this subset, which is approximately 10% of the total sequences (clearly not a random sample), is similar to the one obtained using the whole data set. This suggests that these distributions reflect the global GC content present in these environments. Strikingly, the profiles of archaeal and bacterial fragments were similar between one another for each depth as well as similar to the global data set. To our understanding, this is indicative of the robustness of these results and, more importantly, provides a hint that a process of molecular convergence might be taking place at this level. We are carrying out analyses on more samples to confirm and extend this latter hypothesis.

As a last control, to rule out that this particular behavior is produced by a single, or a few, highly abundant species with a given GC content, we estimated species richness and diversity present at each depth. We aligned the sequences of the small subunit RNA (16S-RNA) genes against a profile obtained from the SILVA database (Pruesse et al. 2007) using the ARB package (Ludwig et al. 2004). Then we calculated distances with ARB using the Olsen correction. After this, the sequences were clustered using the DOTUR furthest-neighbor algorithm with different cut-off values of pairwise sequence identity (99%, 97%, and 95%). Finally, we estimated species richness using the Chao1 and Ace indexes and diversity using Shannon's and Simpson's diversity indices as implemented in the DOTUR package (Schloss and Handelsman, 2005). We also built rarefaction curves using DOTUR. Together these results indicate that these habitats are diverse enough to rule out the possibility of the dominance of a single or few species (data not shown). These results are rather robust because they are consistent among the total samples (the archaea and bacteria subsets) and do not indicate the relative abundance of certain gene families or a single, or a few, dominant species.

Standing on the robustness of the previous results, and given the availability of several environmental variables measured at each depth, we compared the first three samples with the three deeper samples. One of the most striking differences was the variability of oxygen concentration, with a low figure at 770 m (the oxygen minimum zone is around this depth at these coordinates [DeLong et al. 2006]). In this context, it is tempting to link these results to our previous work (Naya et al. 2002), which connects anaerobiosis with low GC and aerobiosis with high GC. Particularly, the skewed distribution toward high GC values, as well as the peak at approximately 58% of oxic environments in 500-m and 4,000-m samples, correspond with the mode of strictly aerobic prokaryotes. In contrast, the presence of the peak at 45% GC of the anoxic 770-m sample is consistent with the mode of strictly anaerobic prokaryotes. This coincidence of the two modes is rather striking and gives further support to our previous findings.

In summary, we have shown that the distributions of GC content of a given depth are consistent within the different subsets. In addition, our results might be connected to an environmental variable, i.e., oxygen concentration, which is concordant with our previous observations showing that strictly aerobic prokaryotes tend to be more GC-rich than strictly anaerobic ones. From a more general perspective, the importance of the environmental-genomics approach will become evident when well-studied and comparable habitats can be connected to evolutionary or ecologic hypotheses.

Acknowledgments This work was partially supported with Grant No. FCE200/05 and a fellowship from Programa de Desarrollo de las Ciencias Básicas (H.R.).

References

Altschul SF, Madden TL, Schffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Argos P, Rossman MG, Grau UM et al (1979) Thermal stability and protein structure. Biochemistry 18:5698–5703

Barbu E, Lee KY, Wahl R (1956) Content of purine and pyrimidine base in desoxyribonucleic acid of bacteria. Ann Inst Pasteur 91:212–224

Belozersky AN, Spirin AS (1958) A correlation between the compositions of deoxyribonucleic and ribonucleic acids. Nature 182:111–112

DeLong EF, Preston CM, Mincer T et al (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311:496–503

Foerstner KU, von Mering C, Hooper SD et al (2005) Environments shape the nucleotide composition of genomes. EMBO Rep 6:1208–1213

Kanehisa M, Araki M, Goto S et al (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36:D480–D484

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kunin V, Copeland A, Lapidus A et al (2008) A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev 72:557–578

Ludwig W, Strunk O, Westram R et al (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32: 1363–1371

Musto H, Naya H, Zavala A et al (2004) Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett 573:73–77

Naya H, Romero H, Zavala A et al (2002) Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J Mol Evol 55:260–264

Pruesse E, Quast C, Knittel K et al (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188–7196

Rocha EPC, Danchin A (2002) Base composition bias might result from competition for metabolic resources. Trends Genet 18: 291–294

Schloss PD, Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol 71:1501–1506

Singer CE, Ames BN (1970) Sunlight ultraviolet and bacterial DNA base ratios. Science 170:822–825

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582–592

Tatusov RL, Fedorova ND, Jackson JD et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41

Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet 6:805–814

Tringe SG, von Mering C, Kobayashi A et al (2005) Comparative metagenomics of microbial communities. Science 308:554–557

Tyson GW, Chapman J, Hugenholtz P et al (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43

Venter JC, Remington K, Heidelberg JF et al (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74