Journal of Phylogenetics & Evolutionary Biology

Brocchieri, J Phylogen Evolution Biol 2014, 2:1
DOI: 10.4172/2329-9002.1000e108

Editorial Open Access

The GC Content of Bacterial Genomes

Luciano Brocchieri*

Department of Molecular Genetics & Microbiology and Genetics Institute, University of Florida, Gainesville FL, USA

Bacterial genomes exhibit a wide range of compositional diversity, most spectacularly represented by variation in genome GC content, which varies in different organisms from as low as 17% to as high as 75%. The nature of the biological processes underlying these differences has been long debated and two polarizing interpretations have been advanced, one proposing that GC content is driven by genome-specific mutational biases (the mutational hypothesis), and one that it reflects different selective processes in different organisms (the selectionist hypothesis). The hypothesis that differences in GC content are mostly driven by species-specific mutational biases [1] implies that smaller variation in GC content across genomes should be seen at positions that are most constrained by any form of purifying selection, and conversely that greatest variation should be observed in positions that are functionally neutral. Differences in GC content among prokaryotic genomes largely reflect on, or are driven by the GC content of protein coding sequences, which usually occupy the majority of the genome. When considering separately the GC content at the three codon positions of genes (GC_1, GC_2, GC_3), typical patterns are observed Figure 1-A. The GC content of all positions varies roughly linearly with the overall content of the genes, but variations in the first two codon positions, and especially in the second codon position, are much reduced compared to the variability observed in third codon positions, where the GC content spans across species almost all possible values from close to $GC_3 = 0.0$ to almost $GC_3 = 1.0$. These differences in variability are consistent with expected constraints imposed by the relation between codons and amino acids [2], with first and second codon positions mostly determining the amino acid type (and second codon position mostly determining the physico-chemical properties of the amino acid), and third codon position being mostly either synonymous, or encoding amino acids with similar properties. It is interesting to observe that the GC content of genomic intergenic regions closely correlates with the GC content of the coding sequences Figure 1 panel B, and it varies across genomes approximately to the same extent as it does in coding regions, and thus much less than in third codon positions.

A simple toy model relating mutational bias to codon compositional substitutability can be advocated to explain the overall contrasts and variability in GC content observed between codon positions of different genomes. In this model, coding regions are represented as sequences formed from a two-letter alphabet {S, W} in which bases are identified either as Strong (S = G or C) or as Weak (W = A or T). Each sequence position is assumed either to evolve freely by substitution between S and W states, or to be constrained by purifying selection either in state S or in state W. Each of the three codon-base-positions i (i = 1, 2, or 3) will be then characterized by a codon-position-specific fraction $f_s^{(i)}$ of sites constrained to be of type S, a fractions $f_w^{(i)}$ of sites constrained to be of base-type W, and a fraction $f_w^{(i)} = 1.0 - f_s^{(i)} - f_w^{(i)}$ of sites freely variable. The idea that a position is either variable or constrained independently of the state of the neighboring positions is an obvious simplification, but we assume that the approximation is sufficient to capture the compositional properties of codons we are interested in. Another assumption is that the fractions of constrained and variable sites do not depend on the genome, i.e., all genomes have identical frequencies of constrained and variable sites. This assumption may be violated more significantly, for example, in genomes that deviate more strongly from the linear relations of GC contents Figure 1-A, such as AT-rich small genomes of very reduced gene content. We finally assume that, within a genome, sequences evolve under the pressure of a homogeneous mutational process characteristic of each genome, defined by two substitution rates, one for substitutions $W \rightarrow S \quad AT \rightarrow GC$ and one for substitutions $S \rightarrow W \quad (q_{SW})$. The equilibrium frequencies corresponding to this substitution model are $\pi_s = q_{ws} / (q_{ws} + q_{sw})$ for nucleotide type S, and $\pi_w = q_{sw} / (q_{ws} + q_{sw})$ for nucleotide type W. Since we assume that any mutation occurring at a constrained site is removed by purifying selection, the mutational process results in substitutions only at variable positions, thus affecting only the fraction i of codon-base-positions i. At equilibrium, the GC content at codon-base-position i will be:

$$S_i = f_s^{(i)} + \pi_s f_v^{(i)}$$

and the total GC content of the coding sequence will be the average of the GC content at each codon-base-position:

$$S = \frac{1}{3}\sum_{i=1}^{3} s_i = \frac{1}{3}\sum_{i=1}^{3} f_s^{(i)} + \pi_s f_v^{(i)}$$

From these relations, the GC content at each of the three codon-base-positions can be expressed as a linear function of the total GC content S:

$$S_i = \frac{f_v^{(i)}}{f_v} s + f_s^{(i)} - \frac{f_s}{f_v} f_v^{(i)}$$

where $f_s = \frac{1}{3}\Sigma_i f_s^{(i)}$ and $f_v = \frac{1}{3}\Sigma_i f_v^{(i)}$ are the fractions of S-constrained and variable sites in coding regions, respectively. From the observed relations between S_i and S, we can infer the fractions of variable and constrained sites in the three codon-base positions, the equilibrium frequencies at variable sites, and the rates of substitution in different genomes. From the distribution among genomes of GC content in third codon position, spanning almost all possible values

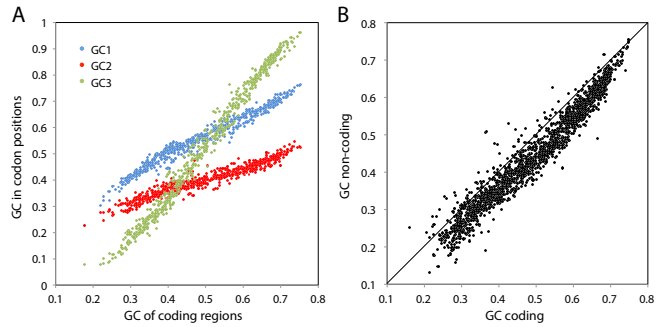

Figure 1: **A**) The GC content at the three codon positions in relation to GC content of coding regions within each genome.

B) For each genome the GC content of coding regions is compared to the GC content of intergenic regions.

***Corresponding author:** Luciano Brocchieri, Department of Molecular Genetics & Microbiology and Genetics Institute, University of Florida, Gainesville FL, USA, Tel: +1 352 273 8131; E-mail: lucianob@ufl.edu

Received March 30, 2014; **Accepted** March 31, 2014; **Published** April 10, 2013

Citation: Brocchieri L (2014) The GC Content of Bacterial Genomes. J Phylogen Evolution Biol 2: e108. doi:10.4172/2329-9002.1000e108

Copyright: © 2014 Brocchieri L. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

between 0.0 and 1.0, we deduce that the fraction of variable sites in third codon position is close to $S_i = a_i S + b_i$. The fractions of S, W, and V sites at all three codon positions can be estimated Table 1 from the equations above and from the coefficients of the linear regressions $S_i = a_i S + b_i$ obtained from the data (Figure 1-A). A similar model can also be applied to intergenic regions, suggesting that these regions harbor about 4-5% more W-constrained positions than coding regions, including, e.g., AT-rich promoters, and a fraction of variable sites similar to the overall fraction estimated for coding regions Table 1, thus, much less than in third codon positions. The model also predicts that the highest possible GC content of genomic coding regions is 75.7%, consistently with observations, and the lowest is 20.0%. The existence of genomes with coding regions of GC content lower than 20% can be explained assuming that these genomes have evolved different fractions of variable and constrained regions. This is not an unrealistic assumption, since genomes with lowest GC content are also very reduced in size and in number of genes [3].

The ratio R of mutational rates, q_{WS} / q_{SW}, in coding regions of different GC content, S, can be derived as:

$$R = \frac{S - f_s}{f_v - S + f_s}.$$

This relation between mutation-rate ratio and gene GC content (Figure 2) suggests that in genes of the lowest GC content the mutational rate towards AT is orders of magnitude higher than the rate towards GC. The very biased rate of mutation towards AT predicted for AT-rich coding regions is consistent with experimental analyses of mutational rates in repair-deficient constructs of *Salmonella typhimurium* [3] and with the deficiency of repair enzymes observed in AT-rich intracellular parasites and endosymbionts of reduced genome size. Conversely, the model predicts higher mutational rates towards GC bases in coding regions of the highest GC content (GC = 0.757), in which only $W \rightarrow S$ mutations are predicted to occur and $q_{SW} \approx 0.0$. However, evidence that this is not the case has been recently provided by the works of Hershberg and Petrov [4] and of Hildebrand and co-workers [5]. Hershberg and Petrov

	Codon position				Intergenic
	1	2	3	All	
f_S	0.339	0.262	0.000	0.200	0.180
f_W	0.247	0.484	0.000	0.244	0.287
f_V	0.414	0.254	1.000	0.556	0.533

Table 1: Codon position.

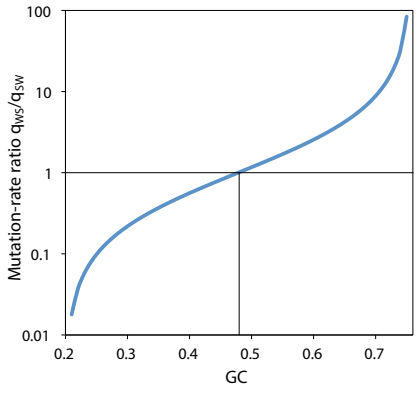

Figure 2: Estimated ratios of mutational rates between weak (W = AT) and strong (S = GC) bases as a function of GC content of the coding sequence, based on a toy model of mutational bias (see text).

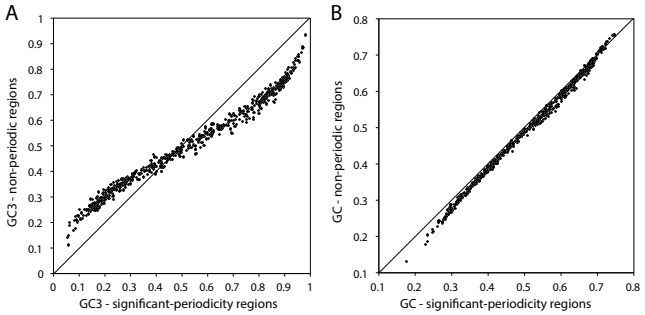

Figure 3: **A)** The GC content in third codon positions in gene regions with significant compositional contrasts among codon positions is compared to the same GC content in coding regions with less pronounced contrasts.

B) The overall GC content of coding-sequence regions with and without compositional contrasts.

[4] analyzed mutations in five clonal pathogens spanning a wide range of GC content and with no evidence of deficiencies in repair systems, and found that mutations were universally biased towards AT even in bacteria of high GC content, concluding that mutations are universally biased towards AT independently of GC content and that high level of GC content must be maintained by selection (or by selection-like processes). Similarly, Hildebrand and co-workers [5] examined mutations at 4-fold degenerate codon positions in a dataset of 149 phylogenetically diverse species, and also found a large excess of synonymous $AT \rightarrow GC$ mutations over $AT \rightarrow GC$ mutations in all but the most AT-rich bacteria. These data strongly suggest that variations of GC content across prokaryotic genomes are determined by selection or selection-like process, with weakest constraints against the prevailing mutational bias observed in parasitic bacteria evolving under relaxed-selection conditions [6,7]. Since compositional biases extend to intergenic regions, they seem not to be related to codon usage. Furthermore, Hildebrand and co-workers (2010) observe that "optimal" codons as used in genes that are highly expressed and hence supposedly under more intense selection, are generally more AT-rich than the average gene within the same genomes, and thus selection on codon usage cannot explain the bias in GC content observed of synonymous codon positions. Rocha and Feil [8] review several theories on environmental factors selecting for optimal genome-wide GC content in prokaryotes, frustrated by mediocre-at-best correlation of GC content with environmental variables [9-15]. Nevertheless, it is intriguing that GC content at third codon positions seems to vary balancing the constraints acting on the first two codon positions in such a way that the overall GC content of coding regions closely reflects the GC content in intergenic regions. To further investigate the possible balancing role of GC_3, we identified within individual genes sequence segments with significant compositional contrasts between codon positions, and compared GC content at these positions with the GC content of regions with non-significant contrasts. Not surprisingly, we found that within the same genome, non-contrasted regions have a reduced GC bias at third codon positions compared to contrasted regions Figure 3-A. However, we also found that the two regions maintained a very similar overall GC content Figure 3-B, suggesting that indeed GC_3 usage played a role in stabilizing the GC content of coding regions with variable constraints on GC usage at non-synonymous positions.

Acknowledgments

This work is supported by NIH Grant 5R01GM87485-2.

References

1. Sueoka N (1961) Correlation between Base Composition of Deoxyribonucleic Acid and Amino Acid Composition of Protein. Proc Natl Acad Sci U S A 47: 1141-1149.
2. Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A 84: 166-169.
3. Lind PA, Andersson DI (2008) Whole-genome mutational biases in bacteria. Proc Natl Acad Sci U S A 105: 17878-17883.
4. Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6: e1001115.
5. Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of selection upon genomic GC-content in bacteria. PLoS Genet 6: e1001107.
6. Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A 93: 2873-2878.
7. Andersson SG, Kurland CG (1998) Reductive evolution of resident genomes. Trends Microbiol 6: 263-268.
8. Rocha EP, Feil EJ (2010) Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? PLoS Genet 6: e1001104.
9. Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 44: 632-636.
10. McEwan CE, Gatherer D, McEwan NR (1998) Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. Hereditas 128: 173-178.
11. Hurst LD, Merchant AR (2001) High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc Biol Sci 268: 493-497.
12. Naya H, Romero H, Zavala A, Alvarez B, Musto H (2002) Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J Mol Evol 55: 260-264.
13. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, et al. (2004) Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett 573: 73-77.
14. Foerstner KU, von Mering C, Hooper SD, Bork P (2005) Environments shape the nucleotide composition of genomes. EMBO Rep 6: 1208-1213.
15. Wang HC, Susko E, Roger AJ (2006) On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors. Biochem Biophys Res Commun 342: 681-684.