Available online at www.sciencedirect.com

SCIENCE DIRECT®

Biochemical and Biophysical Research Communications 347 (2006) 1–3

BBRC

www.elsevier.com/locate/ybbrc

Mini Review

Genomic GC level, optimal growth temperature, and genome size in prokaryotes

Héctor Musto [a,b], Hugo Naya [a], Alejandro Zavala [a,b], Héctor Romero [a,c], Fernando Alvarez-Valín [b,d], Giorgio Bernardi [b,*]

[a] *Laboratorio de Organización y Evolución del Genoma, Facultad de Ciencias, Montevideo, Uruguay*
[b] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Naples, Italy*
[c] *Escuela Universitaria de Tecnología Médica, Facultad de Medicina, Montevideo, Uruguay*
[d] *Sección Biomatemáticas, Facultad de Ciencias, Montevideo, Uruguay*

Received 30 May 2006
Available online 19 June 2006

Abstract

Two years ago, we showed that positive correlations between optimal growth temperature (T_{opt}) and genome GC are observed in 15 out of the 20 families of prokaryotes we analyzed, thus indicating that "T_{opt} is one of the factors that influence genomic GC in prokaryotes". Our results were disputed, but these criticisms were demonstrated to be mistaken and based on misconceptions. In a recent report, Wang et al. [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] criticize our results by stating that "all previous simple correlation analyses of GC versus temperature have ignored the fact that genomic GC content is influenced by multiple factors including both intrinsic mutational bias and extrinsic environmental factors". This statement, besides being erroneous, is surprising because it applies in fact not to ours but to the authors' article. Here, we rebut the points raised by Wang et al. and review some issues that have been a matter of debate, regarding the influence of environmental factors upon GC content in prokaryotes. Furthermore, we demonstrate that the relationship that exists between genome size and GC level is valid for aerobic, facultative, and microaerophilic species, but not for anaerobic prokaryotes.

Keywords: Prokaryotes; Genome evolution; Genome size; GC level; Optimal growth temperature

Fifty years ago, it was discovered that the genomes of prokaryotes cover a broad compositional range, GC levels being comprised between approximately 25% and 75% [1]. The first explanation for such a range was to attribute it to a mutational bias [2,3], a neutralist explanation. Then, thirty years ago, it was discovered that differences in GC levels also existed between the genomes of cold- and warm-blooded vertebrates [4]. The higher GC levels of the genomes from warm-blooded vertebrates were explained as being due to selection for higher thermodynamic stability required by DNA, RNA, and proteins at their body temperatures [5], a selectionist interpretation. As a consequence, it seemed reasonable to expand the latter interpretation to prokaryotes which display a variation in T_{opt} from about −15 to 120 °C.

Along this line, later investigations [6] failed to find a correlation between T_{opt} and GC (or GC_3 in protein-coding sequences of prokaryotes [7]), although positive correlations were observed for ribosomal and tRNA genes. These results were interpreted not only as strong evidence against the thermodynamic hypothesis for prokaryotes, but also as evidence against the corresponding hypothesis for vertebrates. This interpretation was criticized, however, because many factors have often contrasting inputs on genome composition of such a vast array of organisms as prokaryotes [8]. Indeed, it was pointed out [8] that "what Galtier and Lobry (1997) did was to use a sample of pro-

[*] Corresponding author. Fax: +39 081 2455807.
E-mail address: bernardi@szn.it (G. Bernardi).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.06.054

karyotes including *Archaea* and *Bacteria*, mesophiles, thermophiles and hyperthermophiles, aerobes and anaerobes (the latter two groups of prokaryotes being characterized by different GC levels [Naya et al., 2002]), without taking into consideration that many factors are collectively responsible for the base composition of a genome, and different strategies were developed by different prokaryotes to cope with high temperatures. This led to looking at very different and often contrasting inputs as far as genome composition is concerned. *Under these circumstances, plots of prokaryotic GC* (*or* GC_3) versus *optimal growth temperature are meaningless, because they mix different effects on genome composition*" (emphasis in the original text).

It is obvious that taking into account all the environmental factors that have a potential effect on genome composition is impossible, at least not for a sufficiently large number of prokaryotic species. A way to largely circumvent this problem is to restrict the analysis to groups of phylogenetically related prokaryotes and to assume that, apart from the parameter under study (in our case T_{opt}), all other variables are comprised within quite small limits, and hence variables that affect the GC level are likely to be more similar.

These considerations led us [9] to restrict our study of co-variation between T_{opt} and genomic GC to the family level, essentially because the phylogenetic relationships among prokaryotic families are still uncertain in several cases [10], whereas the phylogenetic relationships among species within each family are expected to be more accurate since the times of divergence are much smaller. In brief, we reported that when families comprising at least 10 species were studied, positive correlations were found in 15 out of 20 families. We therefore concluded that "*T_{opt} is one of the factors that influence genomic GC in prokaryotes*".

These results were considered to be "not robust" by Marashi and Ghalanbor [11], a claim echoed by Basak et al. [12,13], because according to these authors the correlations relied on few points that the authors called outliers. This criticism was demonstrated to be mistaken [14] since after eliminating from the samples real outliers, identified using objective and well-established approaches, the results remained basically unchanged. Indeed, the correlations between T_{opt} and genomic GC remained positive and statistically significant for the majority of these families.

In a recent report, Wang et al. [15] state that "all previous simple correlation analyses of GC versus temperature have ignored the fact that genomic GC content is influenced by multiple factors including both intrinsic mutational bias and extrinsic environmental factors". This statement is surprising for at least two reasons. In the first place it is clear, even from the abstract of our paper that we did not ignore that temperature is *one* of many factors that affect GC level. Furthermore, the same consideration was clearly made before (see [8] and references therein). In the second place, this criticism applies not to ours, but to the work of others [6] and, ironically, also to the study of Wang et al. [15]. Indeed, these authors [15] analyzed two sets of prokaryotes in their Figs. 1 and 2 and Table 3. A set of 1065 species was analyzed in their GC and T_{opt} disregarding again the phylogenetic relationships, as had already been done by Galtier and Lobry [6]. Expectedly, no significant correlation was found in either study. Another set of 130 species that contain information on genomic GC, genome size, and oxygen requirement showed that the correlation between GC and temperature on genome size is different for the different oxygen requirement groups, since the interaction terms of GC and oxygen on genome size and oxygen are both significant.

The criticisms raised by Wang et al. [15] are in fact only two. The first one is that, taking into account very recent data [16], which comprises new estimates of T_{opt} from some species that belong to the *Halobacteriaceae* family used by us [1], the T_{opt}/GC-level correlation remains positive ($R = 0.33$), yet is no longer significant for this family. Such a result is not surprising if one considers the serious flaws made in the analysis. Apart from the fact that the sample size is very small ($N = 14$) and thus any manipulation of the data may produce drastic changes in the correlation values (due to the low number of degrees of freedom), the situation is even more delicate because the new estimates of T_{opt} are systematically higher than the old ones (for the same species), as stated in the paper where these estimations are presented [16]. As a consequence, the sample of *Halobacteriaceae* species used by Wang et al. [15] includes two subsets of data, one of which contains systematic biases in the estimation of T_{opt}, since either the estimates for the sub-group of species for which old T_{opt} are used are underestimations, or the values for the group with the new T_{opt} are overestimations. It should be evident that for purposes of comparison, mixing data measured using different criteria is not correct regardless of the "profundity" of the statistical method used. We doubt that any serious conclusion could be drawn from this kind of analysis.

The second criticism concerns the effect of genome size which putatively prevails upon the effect of temperature. Apart from the obvious prediction that if there is a strong correlation between genome size and GC, this may prevail over the correlation between temperature and GC, it is difficult to see why such prevalence in the correlation should rule out the functional significance of other genuine correlations such as that between T_{opt} and genomic GC. Besides it is not possible to establish a cause–effect relationship in the correlation of genome size versus GC.

In this regard, we should mention that the correlation between genome size and genomic GC holds only for aerobic, facultative, and microaerophilic prokaryotes, but not for anaerobic species, as is shown in Fig. 1. This observation gives further support to our previous finding, in that aerobiosis is among the most important factors shaping GC content in prokaryotes [17].

Concerning the genome size/GC level analyses conducted by Wang et al. [15], they restricted their study to only 2 of the 20 prokaryotic families used by us [9]: *Bacillaceae* and *Enterobacteriaceae*. Regrettably, only a fraction of

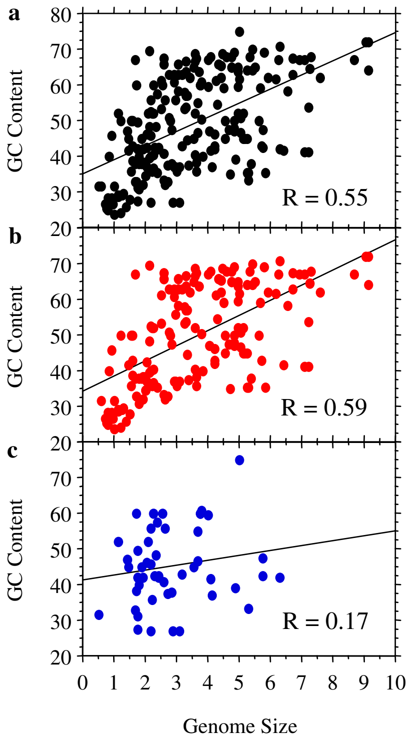

Fig. 1. The data for constructing this plot were taken from http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, which displays several physiological and ecological features for all completed prokaryotic genomes. When several species from the same genus were available, we included the organism with the highest genome size. (a) All species; (b) aerobic, facultative, and microaerophilic prokaryotes; (c) anaerobic prokaryotes.

the species was analyzed: only 15 (out of 38) *Enterobacteriaceae* and only 12 (out of 18) *Bacillaceae*. They found that in spite of the remarkably smaller sample sizes, the correlation between T_{opt} and GC content still holds and remains statistically significant for these two subsets of species analyzed, but while in *Enterobacteriaceae* the correlation between genome size and GC prevails over the correlation between T_{opt} and GC, in *Bacillaceae* the prevailing correlation is that between T_{opt} and GC. Therefore, such a criticism cannot be taken as well founded, to say the least. Disregarding the problem of the serious reduction in sample size, accepting this criticism would be equivalent to denying the importance of the correlation between genome size and GC content in *Bacillaceae* simply because the predominant correlation with GC is that displayed by T_{opt}.

Finally, the statement that "genomic adaptation to elevated environmental temperature in prokaryotes is not generally achieved by increased overall genomic G+C, but involves many molecular processes at the transcriptome and proteome levels" [15] is a trivial consideration if one considers the well-established compositional correlations between the genome, the transcriptome, and the proteome [5]. The real question is which one is the "*primum movens*" among those factors, the relative weight among them and their interplay [8].

In conclusion, we do not see how the analyses done by Wang et al. [15], given their lack of consistency, weaknesses, and poor data management, can disprove our conclusion that "*T_{opt} is one of the factors that influence genomic GC in prokaryotes*".

References

[1] K.Y. Lee, R. Wahl, E. Barbu, Contenu en bases puriques et pyrimidiques des acides désoxyribonucléiques des bactéries, Ann. Inst. Pasteur. 91 (1956) 212–224.

[2] E. Freese, On the evolution of base composition of DNA, J. Theor. Biol. 3 (1962) 82–101.

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

[4] J.P. Thiery, G. Macaya, G. Bernardi, An analysis of eukaryotic genomes by density gradient centrifugation, J. Mol. Biol. 108 (1976) 219–235.

[5] G. Bernardi, G. Bernardi, Compositional constraints and genome evolution, J. Mol. Evol. 24 (1986) 1–11.

[6] 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.

[7] 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 (2001) 493–497.

[8] G. Bernardi, Structural and Evolutionary Genomics. Natural Selection in Genome Evolution, Elsevier, Amsterdam, 2004.

[9] H. Musto, H. Naya, A. Zavala, H. Romero, F. Alvarez-Valin, G. Bernardi, Correlations between genomic GC levels and optimal growth temperatures in prokaryotes, FEBS Lett. 573 (2004) 73–77.

[10] J.G. Holt, N.R. Krieg, P.H.A. Sneath, J.T. Staley, S.T. Williams, Bergey's Manual of Determinative Bacteriology, ninth ed., William & Wilkins, Baltimore, 1994.

[11] 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.

[12] S. Basak, S. Mandal, T.C. Ghosh, Correlations between genomic GC levels and optimal growth temperatures: some comments, Biochem. Biophys. Res. Commun. 327 (2005) 969–970.

[13] S. Basak, T.C. Ghosh, On the origin of genomic adaptation at high temperature for prokaryotic organisms, Biochem. Biophys. Res. Commun. 330 (2005) 629–632.

[14] H. Musto, H. Naya, A. Zavala, H. Romero, F. Alvarez-Valin, G. Bernardi, The correlation between genomic G+C and optimal growth temperature of prokaryotes is robust: a reply to Marashi and Ghalanbor, Biochem. Biophys. Res. Commun. 330 (2005) 357–360.

[15] 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.

[16] J.L. Robinson, B. Pyzyna, R.G. Atrasz, C.A. Henderson, K.L. Morrill, A.M. Burd, E. Desoucy, R.E. Fogleman III, J.B. Naylor, S.M. Steele, D.R. Elliott, K.J. Leyva, R.F. Shand, Growth kinetics of extremely halophilic Archaea (family *Halobacteriaceae*) as revealed by Arrhenius plots, J. Bacteriol. 187 (2005) 923–929.

[17] H. Naya, H. Romero, A. Zavala, B. Alvarez, H. Musto, Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes, J. Mol. Evol. 55 (2002) 260–264.