Available online at www.sciencedirect.com
SCIENCE DIRECT
Biochemical and Biophysical Research Communications 342 (2006) 681–684
BBRC
www.elsevier.com/locate/ybbrc

Mini Review

On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: Data quality and confounding factors

Huai-Chun Wang [a,*], Edward Susko [a], Andrew J. Roger [b]

[a] *Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada B3H 3J5*
[b] *Canadian Institute for Advanced Research, Program in Evolutionary Biology, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, Canada B3H 4H7*

Received 1 February 2006
Available online 20 February 2006

Abstract

The correlation between genomic G+C content and optimal growth temperature in prokaryotes has gained renewed interest after Musto et al. [H. Musto, H. Naya, A. Zavala, H. Romero, F. Alvarex-Valin, G. Bernardi, Correlations between genomic GC levels and optimal growth temperatures in prokaryotes, FEBS Lett. 573 (2004) 73–77], reported that positive correlations exist in 15 families studied. We have reanalyzed their data and found that when genome size and data quality were adjusted for, there was no significant evidence of relationship between optimal temperature and GC content for two of the families that had previously shown strongly significant correlations. Using updated temperature optima for *Halobacteriaceae* species we found the correlation is insignificant in this family. For the family *Enterobacteriaceae* when genome size and optimal temperature are included in a multiple linear regression, only genome size is significant as a predictor of GC content. We showed that more profound statistical methods than simple two factor correlation analysis should be used for analyzing complex intrinsic and extrinsic factors that affect genomic GC content. We further found that a positive correlation between temperature and genomic GC is only evident in free-living species of low optimal growth temperatures.

Keywords: Prokaryotes; G+C content; DNA stability; Optimal growth temperature; Genome size; Life style; Environment

The relationship between genomic G+C content and optimal growth temperature in prokaryotes has been important in understanding whether nucleotide composition is under environmental temperature-based selection. Cross-species and phylogeny-based comparisons of genomic GC content and temperature from various sets of microbial species have established that there is no correlation between the two traits among the species [1–3]. However, a recent study by Musto et al. [4,5] shows that a positive correlation between growth temperature and GC content appears among species within 15 out of 20 families of prokaryotes. This novel finding was promptly challenged by several reports [6–9]. The central arguments are that the correlations observed within the families are generally not very strong, in some cases they are negative rather than positive, and are not robust because of small data sets that are subject to the influence of the outliers [6].

We have reanalyzed the data set of Musto et al. [4]. First, we examined the quality of the original data. The temperature data were largely (416 of 656 species) taken from Galtier and Lobry's collection [1], which was, in turn, collected from Bergey's Manual [10]. Although the latter is the commonly accepted resource for bacterial physiology data including optimal growth temperature and genomic GC content, the optimal temperatures are often given in a range, rather than unique values for different species. The optimal temperatures were calculated by averaging the lower bound and upper bound of the optimal temperature range and the average accuracy is within ±2.5 °C [1]. Although this approach

* Corresponding author. Fax: +902 494 5130.
E-mail address: hcwang@mathstat.dal.ca (H.-C. Wang).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.02.037

may be reasonable for a correlation analysis of temperature and GC content for a large data set [1–3], the data accuracy becomes accurate for a small data set with few data points. In Musto et al.'s data, several (archae)bacterial families contain only 11–15 data points [4].

For instance, an extremely halophilic archaeaon family *Halobacteriaceae* has 14 species in [4]. The Pearson correlation coefficient ($R = 0.67, p < 0.01$) for the optimal temperature and GC content of the family is amongst the highest in the 15 families with positive correlations, using both the cross-species comparison and the comparison based on phylogenetic contrasts [4]. However, the optimal temperature values listed in the original data set may not be up to date. A recent study [11] explicitly examined the temperature optimum in *Halobacteriaceae*; of which eight species are on Musto et al.'s species list. A comparison between optimal temperatures for the eight species from the two sources indicates five of which are very different (Table 1). Using the five updated temperature optima together with temperature data for the other nine species from the original data set, we re-calculated the Pearson coefficient for the correlation between temperature and GC content, resulting in an insignificant correlation ($n = 14$, $R = 0.33$, $p = 0.25$).

Our second concern is 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 [12] and extrinsic environmental factors. One of the intrinsic quantitative factors is the genome size. Large genomes tend to be GC-rich and small genomes tend to be AT-rich [13–15]. Therefore, the confounding factor of genome size should be taken into account when analyzing GC and temperature relationship. Based on Musto et al.'s data, we added genome size data from several sources, including NCBI Genome Projects database (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi), TIGR Genome Properties database [16] and a Prokaryote Genome Size Database provided by T. Ryan Gregory. Of the 20 bacterial families in Ref. [4], only *Bacillaceae* (containing 12 species) and *Enterobacteriaceae* (15 species) have information for all three metrics (temperature optimum, GC content, and genome size) for more than 10 species. For the two families, we made regression analyses of GC content with temperature and genome size, individually and combined (see Table 2).

The results demonstrate that, holding genome size fixed, there is still significant evidence that GC content tends to increase with increasing temperature in *Bacillaceae*, although the strength of the relationship becomes weaker. For *Enterobacteriaceae*, however, given genome size, temperature is no longer a significant predictor of GC content.

Furthermore, a recent metagenomics study has shown that GC content of complex microbial communities seems to be globally and actively influenced by the environment [17]. It is known that oxygen requirements [18,19], nitrogen utilization [20], and habitats [14,21] affect GC content. Salinity and alkalinity may also have effects. As these ecology data have been accumulating in the databases, e.g., [16,22], a generalized linear or nonlinear model involving several interacting factors (numerical or categorical) would be more appropriate than simple two factor correlation (e.g., GC versus temperature) in delineating the role of environmental factors (including temperature) and genomic G+C content. For instance, we generated a data set of 130 microbial species that contain information on genomic GC content, genome size (in mega base pairs), and oxygen requirements (aerobic, anaerobic, microaerophilic, and facultative). Using analysis of covariance (ANCOVA) model we found that the slopes of regression of GC content

Table 1
Optimal growth temperature (OPT) of eight *Halobacteriaceae* species from two resources [5,11]

Genus	Species	OPT (°C) in Ref. [4]	OPT (°C) in Ref. [11]
Natronomonas	*pharaonis*	45	43–45
Haloferax	*volcanii*	45	45
Halorubrum	*saccharovorum*	50	45
Natronococcus	*occultus*	37.5	45
Haloarcula	*vallismortis*	40	43–49
Natronobacterium	*gregoryi*	38.5	47
Halobacterium	*salinarum*	50	49–50
Halococcus	*morrhuae*	33.5	51

Table 2
Regression of genomic G+C on optimal growth temperature (°C) and genome size (mega base pairs)

Family	Factors	R^2[a]	Coeff.[b]	P-value[*]
Bacillaceae (12 species)	Temperature	0.77	0.39 ± 0.1	0.0035
	Genome size	0.70	-5.75 ± 1.86	0.0115
	Temperature + Genome size[c]	0.85	0.28 ± 0.1	0.02 (Temperature)
			-3.47 ± 1.66	0.07 (Genome size)
Enterobacteriaceae (15 species)	Temperature	0.67	1.22 ± 0.37	0.006
	Genome size	0.86	5.30 ± 0.88	<0.0001
	Temperature + Genome size[c]	0.89	0.49 ± 0.30	0.13 (Temperature)
			4.35 ± 1.01	<0.001 (Genome size)

[a] Adjusted R^2 for the regression.
[b] Regression coefficient ± standard errors.
[c] For multiple regression of G+C content on the combined factors (temperature and genome size), partial regression coefficient and their significance are evaluated separately for the two factors.
[*] P-value associated with the null hypothesis that regression coefficient = 0.

on temperature or on genome size are different for the different oxygen requirement groups, since the interaction terms of GC and oxygen or genome size and oxygen are both significant ($p < 0.0001$) in the ANCOVA. This indicates that individual regressions should be fit for the different oxygen requirement groups of microbes.

In a third analysis, we assembled a data set of genomic GC content and optimal temperature of 1065 species, 772 of which were from Ref. [1], and the rest 293 species were from Ref. [4], NCBI Genome Projects database, TIGR Genome Properties database, and German Collection of Microorganisms and Cell Cultures (http://www.dsmz.de). This data set and the other data we used in this study are available at http://www.mathstat.dal.ca/~hcwang/Research/Manuscript/genoGC. We separated the data into five temperature groups (less than 30 °C, 30–40 °C, 40–50 °C, 50–80 °C, and greater than 80 °C), corresponding to mesophiles of low temperature, mesophiles, moderate thermophiles, thermophiles, and hyperthermophiles. Surprisingly average genomic GC is highest in the temperature group of less than 30 °C (G30−) but lowest in the group of hyperthermophiles (greater than 80 °C, G80+). The three middle temperature groups have similar average GC (Fig. 1). Kruskal–Wallis rank sum tests show that the difference in average GC content between the G30 group and the other four higher temperature groups is significant ($p < 0.00001$) while there is no significant difference among the latter four groups ($p = 0.54$).

Moreover, we used Lowess, polynomial and cubic spline smoothing methods to plot the nonlinear relationship between genomic GC and temperature (Fig. 2). Although there are fewer organisms in the G30− group and a large amount of variation, all three methods suggest a positive correlation. For the G30_40 group the methods indicate a negative correlation. There is essentially no correlation in the moderately thermophilic, thermophilic, and hyperthermophilic groups. This is also shown in the correlation analyses for the five temperature groups (Table 3). The

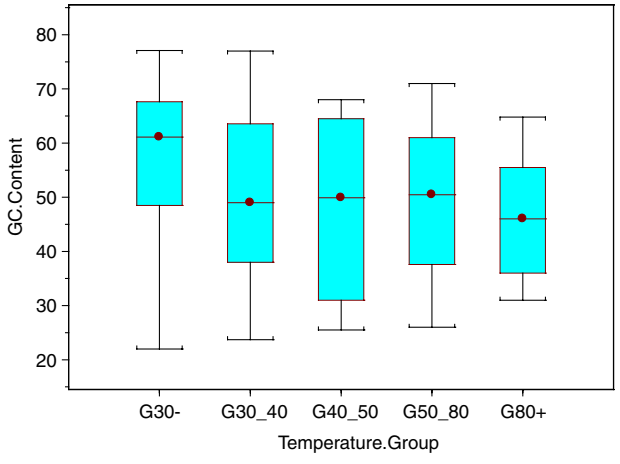

Fig. 1. The box-and-whisker plot of genomic G+C content in five temperature groups. The box represents the first quartile, the median, and the last quartile; the whiskers extend to the most extreme data points.

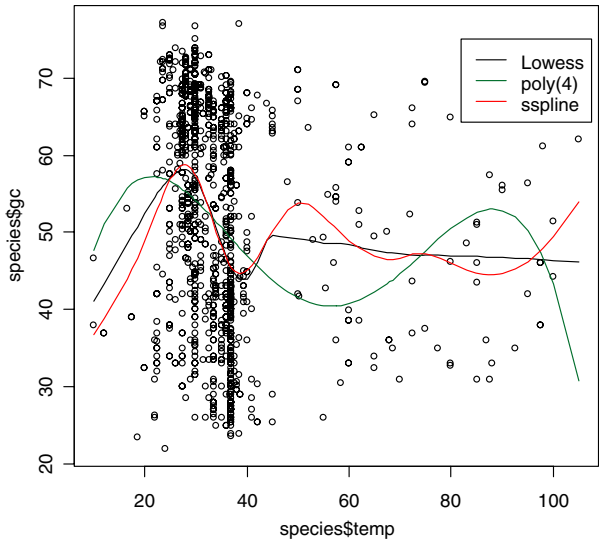

Fig. 2. The regression of genomic G+C on optimal growth temperature using three smoothing methods: Lowess, polynomial to the power of four, and cubic spline.

Table 3
Correlation of genomic GC and optimal growth temperature in five temperature groups

Temperature group (°)	Species number	*R*	Significance
<30	296	0.29	$<10^{-6}$
30–40	653	−0.38	$<10^{-6}$
40–50	35	0.14	0.41
50–80	54	−0.21	0.12
⩾80	27	0.23	0.25

R, Pearson's correlation coefficient.

G30_40 group contains many human pathogenic species and endosymbiotic bacteria which are known to be GC-poor [14,21]. Thermophilic species also tend to be GC-poor because of selection to avoid spontaneous cytosine deamination at high temperatures [23]. These results suggest GC content is strongly affected by the life style of the organisms.

The above analyses have shown that a positive correlation between growth temperature and genomic GC content appears in certain phylogenetic groups (e.g., *Bacillaceae*) and at low temperature range (i.e., temperature less than 30 °C). This correlation is not present in the *Enterobacteriaceae* and *Halobacteriaceae*, contrary to the claim by Musto et al. [4,5]. However, the accuracy of the current and previous analyses was compromised by the quality of the available temperature data, as the real temperature optimum is usually hard to determine and therefore can be different among various sources. The fact that a bacterial species can sustain a wide range of temperatures suggests that environmental temperature is not a primary factor affecting genomic DNA stability. Indeed, for most bacteria, the range from minimum to maximum growth temperature is 30 °C [24]. On the other hand, for the same temperature optimum (e.g., 37 °C), genomic GC content

can be from 23.7% in *Mycoplasma bovoculi* to 69.5% in *Pseudomonas pseudomallei*. In light of this, it would likely be better to use maximum growth temperature instead of optimal temperature in the correlation analyses, as the maximum temperature of a species has a narrower range than the optimal range [11,24] and the DNA thermal stability, if any, may be best tested by the maximum heat that microbes can sustain. Overall, it is clear 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 [25], including for instance, elevated G+C in structural RNAs [1–3], increased frequency of purines and polypurine tracts in message RNAs [26,27], codon usage and amino acid usage biases [8,28,29], and the presence of unique proteins [30,31].

Acknowledgments

HCW was supported by postdoctoral funding from a Genome Atlantic/Genome Canada Large-scale Project "a comparative understanding of prokaryotic evolution and diversity: from genomics to metagenomics". This research was supported by Discovery grants awarded to E.S. and A.J.R. by the Natural Sciences and Engineering Research Council of Canada. A.J.R. and E.S. are fellows of the Canadian Institute for Advanced Research Program in Evolutionary Biology. A.J.R. is supported by a fellowship from the Alfred P. Sloan foundation and the Peter Lougheed New Investigator Award from the Canadian Institutes of Health Research and the Peter Lougheed Medical Research Foundation.

References

[1] N. Galtier, J.R. Lobry, Relationships between genomic GC content, RNA secondary structures and optimal growth temperature in prokaryotes, J. Mol. Evol. 44 (1997) 632–636.
[2] 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.
[3] H.-C. Wang, D.A. Hickey, Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes, Nucleic Acids Res. 30 (2002) 2501–2507.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] H.-C. Wang, The effects of nucleotide bias on genome evolution. PhD thesis, University of Ottawa, Ottawa, 2005.
[10] J.G. Holt, N.R. Krieg, P.H.A. Sneath, J.T. Staley, S.T. Williams, Bergey's Manual of Determinative Bacteriology, William and Wilkins, Baltimore, 1994.
[11] 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.
[12] N. Sueoka, On the genetic basis of variation and heterogeneity of DNA base composition, Proc. Natl. Acad. Sci. USA 48 (1962) 582–592.
[13] N.A. Moran, Microbial minimalism: genome reduction in bacterial pathogens, Cell 108 (1962) 583–586.
[14] E.P. Rocha, A. Danchin, Base composition bias might result from competition for metabolic resources, Trends Genet. 18 (2002) 291–294.
[15] U. Bastolla, A. Moya, E. Viguera, R.C. van Ham, Genomic determinants of protein folding thermodynamics in prokaryotic organisms, J. Mol. Biol. 343 (2004) 1451–1466.
[16] D.H. Haft, J.D. Selengut, L.M. Brinkac, N. Zafar, O. White, Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics, Bioinformatics 21 (2005) 293–306.
[17] K.U. Foerstner, C. von Mering, S.D. Hooper, P. Bork, Environments shape the nucleotide composition of genomes, EMBO Rep. 6 (2005) 1208–1213.
[18] 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.
[19] J.R. Lobry, Life history traits and genomic structure: aerobiosis and G+C content in bacteria, Lect. Notes Comput. Sci. 3039 (2004) 679–686.
[20] C.E. McEwan, D. Gatherer, N.R. McEwan, Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus, Hereditas 128 (1998) 73–78.
[21] M. Woolfit, L. Bromham, Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes, Mol. Biol. Evol. 20 (2003) 1545–1555.
[22] J. Hughes, D. Field, Ecological perspectives on our complete genome collection, Ecol. Lett. 8 (2005) 1334–1345.
[23] M. Ehrlich, K.F. Norris, R.Y. Wang, K.C. Kuo, C.W. Gehrke, DNA cytosine methylation and heat-induced deamination, Biosci. Rep. 6 (1986) 387–393.
[24] L. Rosso, J.R. Lobry, J.P. Flandrois, An unexpected correlation between cardinal temperatures of microbial growth highlighted by a new model, J. Theor. Biol. 162 (1993) 447–463.
[25] D.A. Hickey, G.A.C. Singer, Genomic and proteomic adaptations to growth at high temperature, Genome Biol. 5 (2004) 117.
[26] R.J. Lambros, J.R. Mortimer, D.R. Forsdyke, Optimum growth temperature and the base composition of open reading frames in prokaryotes, Extremophiles 7 (2003) 443–450.
[27] A. Paz, D. Mester, I. Baca, E. Nevo, A. Korol, Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes, Proc. Natl. Acad. Sci. USA 101 (2004) 2951–2956.
[28] G.A.C. Singer, D.A. Hickey, Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content, Gene 317 (2003) 39–47.
[29] P.J. Haney, J.H. Badger, G.L. Buldak, C.I. Reich, C.R. Woese, G.J. Olsen, Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species, Proc. Natl. Acad. Sci. USA 96 (1999) 3578–3583.
[30] P. Forterre, A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein, Trends Genet. 18 (2002) 236–237.
[31] K.S. Makarova, Y.I. Wolf, E.V. Koonin, Potential genomic determinants of hyperthermophily, Trends Genet. 19 (2003) 172–176.