Available online at www.sciencedirect.com

SCIENCE DIRECT®

Biochemical and Biophysical Research Communications 330 (2005) 357–360

BBRC

www.elsevier.com/locate/ybbrc

Breakthroughs and Views

The correlation between genomic G + C and optimal growth temperature of prokaryotes is robust: A reply to Marashi and Ghalanbor

Hector Musto [a,b], Hugo Naya [a], Alejandro Zavala [a,b], Hector Romero [a,c], Fernando Alvarez-Valin [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 9 February 2005
Available online 11 March 2005

Abstract

We have recently shown that optimal growth temperature (T_{opt}) is one of the factors that influence genomic GC in prokaryotes. Our results have been disputed by Marashi and Ghalanbor, who claim that the correlations we show are not "robust" because the elimination of some points (arbitrarily chosen) leads, in some families, to variations in the correlation coefficients and/or significance of correlations. Here, we test whether the correlation between T_{opt} and genomic GC is robust by using two independent approaches: detection of possible outliers (using robust Mahalanobis distance) and usage of a non-parametric correlation coefficient that is not sensitive to the presence of outliers. The results presented here reinforce our previous proposal that T_{opt} is correlated with genomic GC in prokaryotes.

Keywords: Genome evolution; Isochores; DNA thermodynamic stability

In a recent paper [1], we have shown that when families of prokaryotes comprising at least 10 species are studied, positive correlations between optimal growth temperature (T_{opt}) and genomic GC are found for 15 out of 20 taxa analyzed. Furthermore, a comparative analysis by independent contrasts made within the families (in order to control for phylogenetic non-independence) showed qualitatively equivalent results. We concluded that T_{opt} is one of the factors that influence genomic GC in prokaryotes.

Our data (available at: http://oeg.fcien.edu.uy/Temperature/) were promptly re-analyzed by Marashi and Ghalanbor [2] in a paper entitled "Correlations between genomic GC levels and optimal growth temperatures are not robust" claiming that our "correlations can be considerably influenced by exclusion of very few (even as small as one) species from each dataset." Here we will rebut these criticisms.

In the first place, it is necessary to rely on a clear definition of robustness. According to [2] a correlation coefficient would be "robust," if after eliminating the point (or the few points) that contribute most to the correlation, the correlation still holds and remains statistically significant. The problem of robustness of statistical estimates has been largely studied (see [3] and references therein). Robust statistics methods are those essentially unaffected by the presence of extreme

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

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.02.133

values (stragglers and outliers). Since the problem of robustness has been studied in detail, there is little room for improvisation.

Unfortunately, Marashi and Ghalanbor's definition of robustness not only was improvised, but was also exceedingly severe and leading to wrong conclusions, in particular when the sample size is small. Note, for instance, that if the sample size is small (say 10 or less than 15), each point matters very much. It is even possible to make a correlation coefficient become non-significant just because of a decrease in the number of degrees of freedom, even if the r value remains relatively stable. Moreover, when only small datasets are available, these datasets very often show fortuitous grouping of points that give the appearance of outliers (e.g., see [4]).

Instead of using the whimsical criterion used by Marashi and Ghalanbor for eliminating points (namely a criterion that depends only on researcher's choice and is by no means an acceptable option), it is more logical to define, a priori, a sensible and objective rule, and then to apply the same criterion to all samples and see whether the r value remains stable. One objective way to determine whether the correlation coefficients are robust is to identify possible outliers by using a suitable test, and then to recalculate the correlation coefficient after eliminating those points that are outliers according to the test.

We have chosen the following two approaches for testing the robustness of our results: (a) conducting well-accepted tests for outlier detection (robust Mahalanobis distance [5]); (b) using a non-parametric correlation coefficient (Spearman's Rho), that is not sensitive to the presence of outliers besides not assuming any specific distribution.

Table 1 summarizes the results of these analyses. The first point to be remarked is that in 12 families (out of the 20 analyzed), there are points that could be considered as outliers. After eliminating these "outliers," the results remain qualitatively equivalent, namely most families exhibiting statistically significant correlation coefficients remain in the same situation. There are three exceptions: (a) Micrococacceae, where the correlation coefficients go from $r = +0.41$ ($p < 0.05$) to $r = +0.26$ (NS), after eliminating two putative outliers; (b) Microbacteriaceae, where the correlation coefficients go from $r = +0.37$ (NS) to $r = +0.58$ ($p < 0.05$), after eliminating one outlier; and (c) Neisseriaceae, where the correlation coefficient becomes statistically significant ($r = -0.38$, $p > 0.05$ to $r = -0.61$, $p < 0.01$), after eliminating two points.

As for the results obtained with the non-parametric correlation coefficient (last two columns of Table 1), the pattern again remains qualitatively equivalent, that is, most families that exhibited positive and significant Pearson's r values also exhibit positive and significant Spearman's Rho-values.

In summary, there were eight families exhibiting positive and significant correlations coefficients using Pearson's r, out of which 7 (Acidaminococcaceae, Bacillaceae, Enterobacteriacea, Flexibacteriacea, Halobacteriaceae, Methanobacteriaceae, and Pseudomonadaceae) do not have the problem raised by Marashi and

Table 1
Correlations between Topt and genomic GC within 20 prokaryotic Families

Family	N_1	R_1	Significance	N_2	R_2	Significance	Rho	Significance
Acetobacteraceae	14	+0.34	NS	14	+0.34	NS	+0.59	—*
Acidaminococcaceae	**11**	**+0.77**	**	**10**	**+0.93**	****	**+0.52**	**NS**
Bacillaceae	**18**	**+0.80**	****	**17**	**+0.84**	****	**+0.80**	****
Chromatiaceae	**12**	**+0.21**	**NS**	**11**	**−0.23**	**NS**	**−0.10**	**NS**
Clostridiaceae	**59**	**+0.20**	**NS**	**55**	**+0.02**	**NS**	**+0.12**	**NS**
Comamonadaceae	**22**	**+0.02**	**NS**	**20**	**−0.11**	**NS**	**−0.04**	**NS**
Corynebacteriaceae	11	−0.67	*	11	−0.67	*	−0.32	NS
Enterobacteriaceae	**38**	**+0.54**	***	**34**	**+0.35**	*	**+0.48**	**
Eubacteriaceae	11	−0.21	NS	11	−0.21	NS	−0.36	NS
Flavobacteriaceae	15	−0.02	NS	15	−0.02	NS	−0.08	NS
Flexibacteriaceae	10	+0.75	*	10	+0.75	*	+0.75	*
Halobacteriaceae	14	+0.67	**	14	+0.67	**	+0.69	**
Methanobacteriaceae	**12**	**+0.57**	*	**11**	**+0.82**	**	**+0.56**	+
Microbacteriaceae	**15**	**+0.37**	**NS**	**14**	**+0.58**	*	**+0.36**	**NS**
Micrococcaceae	**25**	**+0.41**	*	**23**	**+0.26**	**NS**	**+0.25**	**NS**
Neisseriaceae	**23**	**−0.38**	**NS**	**21**	**−0.61**	**	**−0.60**	**
Pseudomonadaceae	13	+0.63	*	13	+0.63	*	+0.57	*
Rhodobacteraceae	15	+0.15	NS	15	+0.15	NS	−0.05	NS
Spirochaetaceae	**13**	**−0.49**	**NS**	**12**	**−0.28**	**NS**	**−0.15**	**NS**
Staphylococcaceae	**17**	**+0.46**	+	**16**	**+0.48**	+	**+0.55**	*

N_1, R_1 and N_2, R_2 are the numbers of species analyzed within each Family and the product–moment (Pearson) correlation coefficients, before and after eliminating possible outliers, respectively. The last two columns indicate the Spearman's Rho and its significance. Significances are as follows: NS, not significant; *, **, ***, and **** are significant at the 5%, 1%, 0.1%, and 0.01% levels, respectively. $^+$Indicates those coefficients that are at the limit of significance ($0:05 < p < 0:06$).

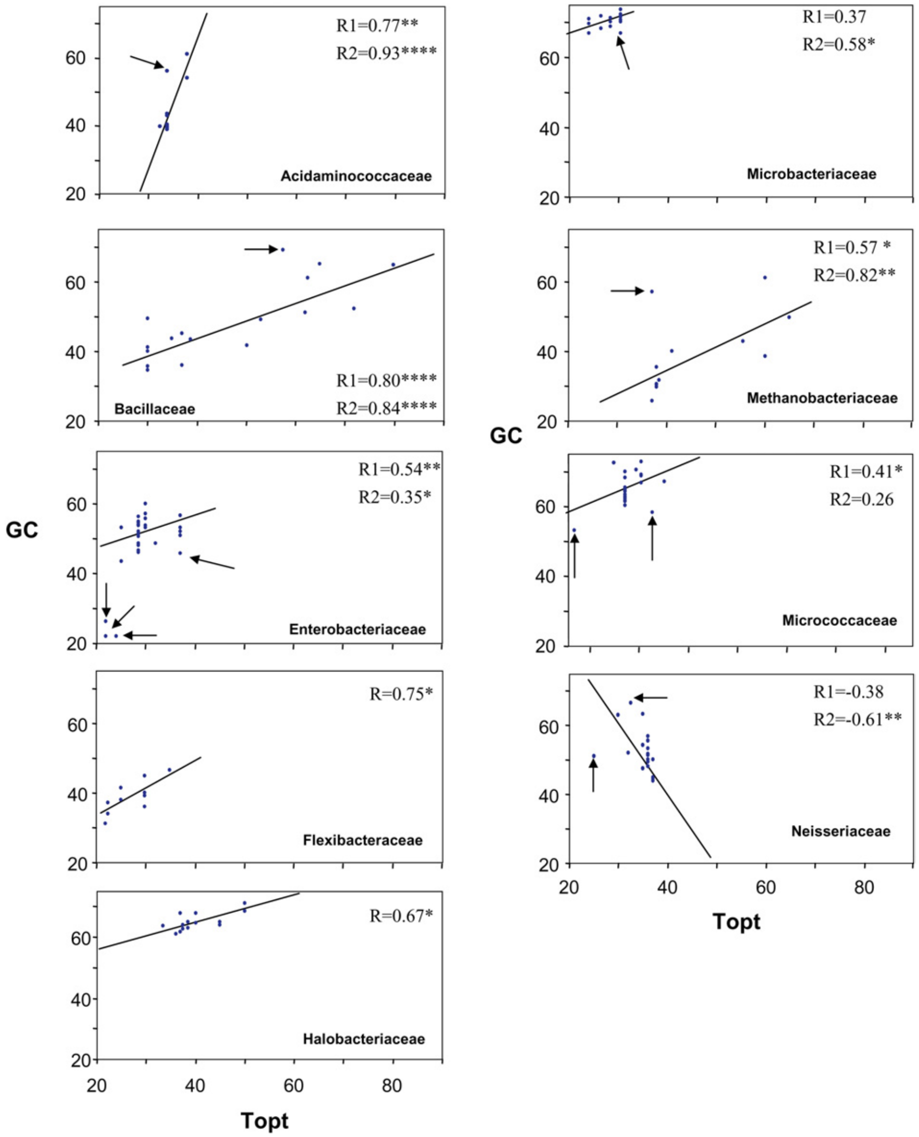

Fig. 1. Scatterplots between genomic G + C content and optimal growth temperature in nine prokaryotic families that exhibit significant correlation coefficients, as well as in those where the elimination of possible outliers led to a change in the level of significance. Arrows indicate those points that according to the Mahalanobis distance are outliers. R_1 and R_2 correspond to the Pearson's r values before and after eliminating possible outliers. The regression line was estimated after the elimination of outliers. Please note that the graph of *Micrococacceae* starts at 15° T_{opt}.

Ghalanbor. To make our point even clearer, Fig. 1 shows the scatterplots between T_{opt} and GC for those families that displayed statistically significant Pearson's r values, as well as for those where the robust Mahalanobis distance found "outliers" and recalculation or Pearson's r led to a change in the significance of the r value (for obvious reasons we do not include in this figure those families that did not exhibit significant r values, and remain with non-significant r values after eliminating possible outliers).

The criticisms raised by Marashi and Ghalanbor do not only refer to the reliability of correlations within families, but also to the soundness of the analysis of independent contrasts. Surprisingly enough, these authors draw their conclusions without conducting any analysis. Indeed, these authors state "Here, we have shown that in the first study, Pearson's correlation coefficients depend on the absence or presence of certain points. Hence, one can conclude that the results of the second study are sensitive to omission of such points, as well."

Two clarifications are in order. First, this statement has no basis because it does not rely on any analysis, but on the assumption that if some points "misbehave" in the first analysis they also will "misbehave" in the second one. Second, and more important, in this second study we showed that the results remain essentially the same (Table 1) after applying an objective criterion for eliminating possible outliers or using a statistics unaffected by outliers. It is noteworthy that, if one correlates the results obtained using independent contrasts with those obtained after eliminating the "outliers" detected by the robust Mahalanobis distance, the *r* value is nearly the same to that reported in [1].

Finally, we would like to mention that in our opinion outlier tests should be used with care and, of course, identified data points should only be removed if a technical reason can be found for their aberrant behavior. A caution is in order concerning the use of discretionary criteria for eliminating points, since by doing so it is possible to demonstrate or refute almost anything. Here, we have recalculated the correlation coefficients after eliminating putative outliers simply to show that the *r* values we presented in [1] are reliable since the positive results do not depend on the putative misbehavior of a few points in each family.

References

[1] 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.

[2] 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.

[3] F.A. Alqallaf, K.P. Konis, R.D. Martin, R.H. Zamar, Scalable robust covariance and correlation estimates for data mining, in: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. Edmonton, Alberta, Canada, 2002, pp. 14–23.

[4] S. Burke, Missing values, outliers, robust statistics and non-parametric methods. LC*GC Europe Online Supplement (1999) 19–24.

[5] K.I. Penny, A comparison of multivariate outlier detection methods for clinical laboratory safety data, Statistician 50 (2001) 295–308.