Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 353 (2007) 207–210
BBRC
www.elsevier.com/locate/ybbrc

GC content and genome length in Chargaff compliant genomes

David Mitchell *

Vice Deanery of Genetics and Microbiology, Trinity College, Dublin, Ireland

Received 26 November 2006
Available online 11 December 2006

Abstract

Musto et al. [H. Musto, H. Naya, A. Zavala, H. Romero, F. Alvarez-Valin, G. Bernardi, Genomic GC level, optimal growth temperature, and genome size in prokaryotes, Biochem. Biophys. Res. Commun. 347 (2006) 1–3] recently reported a linear correlation between GC content and genome length. The regression model was heteroscedactic which suggested that the relationship might be more clearly defined. Alternative regression models ($R^2 > 0.95$) were fitted to a set of over 900 sequences compliant with Chargaff's second parity rule. The new models suggest that the relationship between GC content and genome length is more complex than was originally suggested. While similar models can be derived for non-Chargaff compliant genomes, their interpretation is likely to be more difficult.

Keywords: DNA; Genome; Chargaff's rule; GC content; Archaebacteria; Virus; Bacteria

In 1956 it was found that bacterial guanine and cytosine content varies between 25% and 75% [1]. Later it was suggested that this difference had arisen as a consequence of mutational bias [2,3]. In 1973 bovine DNA was found to form three bands on centrifugation that corresponded to guanine–cytosine content (GC content) of 39%, 48%, and 54% [4]. Subsequent investigation in 25 eukaryotic genomes found similar banding patterns in mammalian and avian genomes but not in the examined invertebrate or unicellar genomes: it was proposed that this intragenomic heterogeneity might be an adaptation to a higher body temperature [5]. More recently investigation of the relationship between temperature and GC content has been extended to prokaryotic genomes: as a surrogate for vertebrate body temperature the optimal growth temperature was used [6,7]. Currently opinion is divided on the existence of a relationship between these parameters [8,9].

While investigating this relationship Musto et al. [9] also reported a linear relationship between GC content and genome length within a subgroup of the genomes they examined: specifically in the aerobic, facultative, and microaerophilic bacteria. Two difficulties are apparent with this proposed relationship. The first is the presence of heteroscedacity in the model. One of the assumptions underlying the use of ordinary least squares regression is that the residuals are independent of the model's variables (homoscedacity) [10]. In cross-sectional studies, heteroscedacity is a common problem and may indicate that a significant variable is missing from the model, the relationships between the variables in the model have been misspecified or a combination of these. The presence of heteroscedacity may be formally tested with a number of tests including White's [11] but this may not be necessary. On inspection of the plots it is clear that the residuals near the centre are considerably greater than elsewhere. While this is alleviated somewhat when the model was applied to the subgroup, the heteroscedacity persists.

The model suggested that genome length is approximately proportional to the square root of the number of its guanine residues. Since bacterial obey Chargaff's second rule [12] to an excellent approximation

$$L = A + T + C + G = 2A + 2G$$

where L is the genome length and A, C, G, and T are the number of adenosine, cytosine, guanine, and thymidine

* Fax: +353 1 679 9294.
E-mail address: dmitchel@tcd.ie

0006-291X/$ - see front matter
doi:10.1016/j.bbrc.2006.12.008

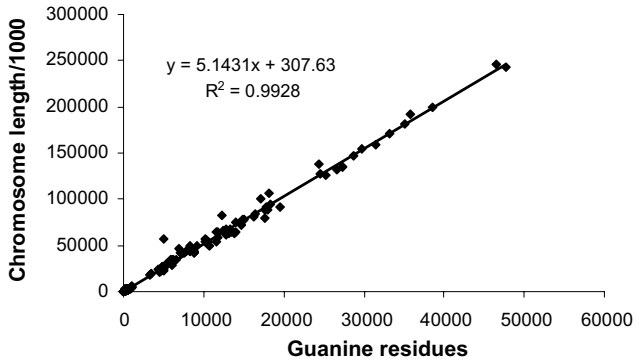

Fig. 1. Eukaryote chromosome length divided by 1000 plotted against genome guanine residue number. $N = 161$, $F = 22{,}204.6$, $p < 10^{-173}$.

residues, respectively. With this approximation the Musto et al.'s [9] model is

$$L = a(G + C)/L + b + \varepsilon = 2aG/L + b + \varepsilon$$

where a and b are constants and ε is a residual term uncorrelated with the variables. Solving for L

$$L = [b + \varepsilon + \{(b + \varepsilon)^2 + 4aG\}^{0.5}]/2$$

Expanding to the second term and taking expectations

$$L = 0.5[b + \alpha + (b^2 + \sigma^2)/\alpha]$$

where $\alpha = (4aG)^{0.5}$, $E(\varepsilon^2) = \sigma^2$ a constant and $E()$ is the expectations operator. $E(2b\varepsilon) = 2bE(\varepsilon) = 0$ since b is constant. From the data in [9] Fig. 1b, $a = 0.5$ and $b = 0.6$ approximately.

Until a comprehensive theory of genomes is developed, it is difficult to know if this proposed relationship between genome length and GC content is correct. Intuitively one would anticipate that, at least to a first approximation, a linear relationship between GC content and genome length would instead exist (Fig. 1): this prompted a re-examination of this model. The results here suggest that revision of the original model may be indicated.

Materials and methods

Since archaebacteria, double stranded DNA (dsDNA) viruses and eukaryotes like bacteria obey Chargaff's second rule, these were also examined here. Twenty archaebacterial genomes, 436 dsDNA viral genomes, 164 eukaryotes chromosomes, and 231 bacterial genomes were examined. The full listing is given in Supplemental data. Lengths and guanine residue number in each sequence were determined. Since $G = C$ to an excellent approximation in these genomes, to the same approximation G/L is half the genome's GC content. The variables in the regression model were chosen as follows: G because it seems intuitive that genome length and guanine residue number would be related; $G^{0.5}$ because Musto et al.'s findings; and G/L to examine the relationship between genome length and GC content after the other variables were included.

For all four genome types, genome length was regressed against seven sets of parameters (a) G (b) $G^{0.5}$ (c) G/L (d) G and G/L (e) $G^{0.5}$ and G/L, (f) G and $G^{0.5}$, and (g) G, $G^{0.5}$ and G/L. To make the coefficients more readable, the L and G values were first divided by 1000. The G value used in the $G^{0.5}$ term was not divided before the square root was taken. Variables were eliminated from the regression model if their t value was <2.0. Competing regression models were compared with the F test [10]

$$F = [(\rho^2 - r^2)/p]/[(1 - \rho^2)/(n - m)]$$

where ρ^2 is the R^2 value of the new regression, r^2 is the R^2 value of the old regression, n is the number of data points, p is the number of additional variables introduced into the new model and m is the total number of parameters in the new model. F is distributed with p and $n - m$ degrees of freedom. Each group was examined separately. The regression analysis was carried out with Microsoft's Excel 2003.

Results

Results for all the models examined are given in Supplemental data. The best fitting model for the bacterial genomes included all three parameters (Table 1) and had an R^2 value of 0.990 ($p < 10^{-226}$). The next best model included only $G^{0.5}$ and G/L and had an R^2 value of 0.989. Comparing these models, F was 23.3 ($p < 10^{-3}$) and the t value for the G coefficient was 4.8 ($p < 10^{-3}$). While these were statistically significant, the small change in R^2 raises a question over the inclusion of this variable. This model was also the best fit to the dsDNA genome data ($R^2 = 0.953$). The next best model included G and G/L ($R^2 = 0.941$). The F value for the comparison was 111.3. The t value for the $G^{0.5}$ coefficient was 10.5 but including the $G^{0.5}$ term increased R^2 by only 0.011.

The eukaryotic data were best fitted by the G and G/L model ($R^2 = 0.994$). When the full model was applied to the eukaryotes the $G^{0.5}$ coefficient had a t value of 0.93 ($p > 0.35$). In all cases the coefficient of the G/L coefficient was significant and less than zero. The archaebacterial data were best fitted by the model with the $G^{0.5}$ and G/L parameters ($R^2 = 0.986$). The next best model included the G and G/L parameters ($R^2 = 0.985$). Comparing these models, $F = 1.5$ $p > 0.10$. When all three parameters were included the regression became inconsistent with $t < 2$ for both G

Table 1
Coefficients and constants for the best regression models

	Constant	$G^{0.5}$	G	G/L	R^2	F
Bact	428.95 (6.0)	210.12 (36.9)	0.45 (4.8)	−13660.3 (−47.6)	0.990	7526.1
Vir	63.96 (16.0)	2.56 (14.1)	0.53 (10.6)	−411.2 (−29.1)	0.953	2969.0
Arch	2315.1 (14.1)	4.63 (34.6)	*	−10538.7 (−15.1)	0.985	625.4
Euk	7104.35 (4.8)	*	5.12 (156.9)	−32871.3 (−4.8)	0.994	12602.1

The value of G and L were first divided by 1000 before the coefficients were calculated. The t values for the coefficients are given in the parentheses. Adjusted R^2 and F values for each regression are given. $p < 10^{-15}$ for all regressions. Asterisk (*) indicates that the parameter was omitted from the model. *Abbreviations:* bact, bacterial genomes; vir, dsDNA viral genomes; arch, archaebacterial genomes; euk, eukaryotic chromosomes.

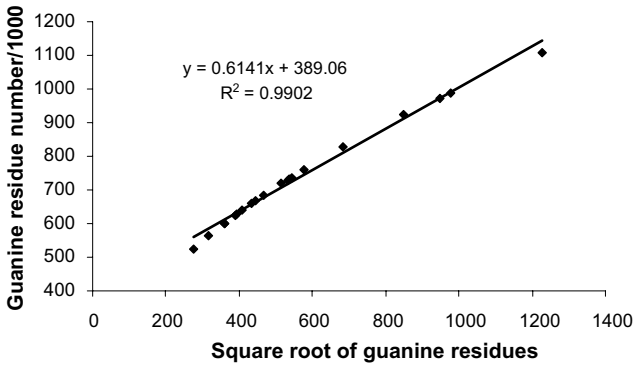

Fig. 2. Square root of the number of guanine residues and the number of guanine residues in the archaebacterial genomes.

and $G^{0.5}$. Inconsistency in the regression when all three parameters were included suggested that an unrecognised collinearity between the variables might exist. On plotting the G and $G^{0.5}$ values (Fig. 2) that these two were linearly dependent. The best fitting model with only one of these parameters was then chosen.

Discussion

The main findings here are: (1) complex relationships exists between the length of a Chargaff complaint genome and its composition, (2) the relationship differs between bacteria and double stranded DNA viruses on one hand and eukaryotes on the other, and (3) the GC content coefficient in all these models is less than zero. The best fitting model for the archaebacteria is distinct from both those for bacteria and eukaryotes. While the negative coefficient of the GC content here contrasts with the findings of Musto et al. [9], this sign change is a consequence of the heteroscedacity in the earlier model. The relatively poor fit for the dsDNA viruses here may reflect the fact that Chargaff's rule is less exact in these genomes than in the other groups [12].

Despite the models' fit here caveats exist. Since genomics is in its infancy little theory exists to guide the creation of exploratory models. Correlation does not imply causation so it remains a possibility that these purely empirical models may not reflect the real relationship between the variables but rather that of an unidentified confounding variable. Secondly although the data set here includes almost all available genomes, it is debatable if this is a statistically proper sample of all existing genomes. As more data become available it may be worth revisiting these models.

Assuming that the analysis presented here is in fact correct, there are several points that seem worth noting. Although the models here were based solely on Chargaff complaint genomes, similar models can be developed for the other genomes. One difficulty with these is the multitude of independent variables $-C$, G, $C+G$, $C^{0.5}$ etc—that can be included in the model. While choosing between competing models on statistical grounds is possible, in absence of theory it is difficult to know which—if any—of these models would reflect the underlying relationship. Chargaff's second rule both simplifies the analysis and imposes restrictions on the models making it intuitively more probable that the models described here are in fact correct.

Since guanine residues are part of the genome, a positive correlation between their number and genome length must exist. Since all sequenced prokaryotic genomes are of similar length but vary considerably in GC content, in these genomes any relationship between the variables must be non linear. Although variable choice here was guided by Musto et al.'s findings, the model's fit was unexpected (Fig. 3).

While the results here seem to suggest that an 'optimal' eukaryotic GC content of ~40% may exist, this conclusion may be premature. The *Plasmodium falciparum* and *P. vivax* genomes have GC contents of ~20% and ~60%, respectively [13]. Despite apparently similar life cycles and having arisen from a common ancestor ~150 million years ago [14], within this relatively short period their genomes have undergone significant changes in GC content, suggesting that this parameter may be more flexible—at least in eukaryotes—than is generally realised. Alternatively, since *Plasmodium* is a parasitic genus, it may instead be that this 40% GC rule applies only to non parasitic genera. Additional data will be needed to resolve this issue.

The possibility of two models for archaebacterial genomes was unexpected. The collinearity between the variables was unanticipated. This is curious and in contrast to the prokaryotic and eukaryotic models. Resolving this problem is difficult as both models are good fits to the data. These organisms are known to have both prokaryotic and eukaryotic features and it may be this previously unrecognised difference is simply another feature differentiating them from both groups. When additional genomes from this group become available it may be possible to determine which—if either—is the correct model.

While it is somewhat surprising that in this large and disparate group of organisms these simple equations

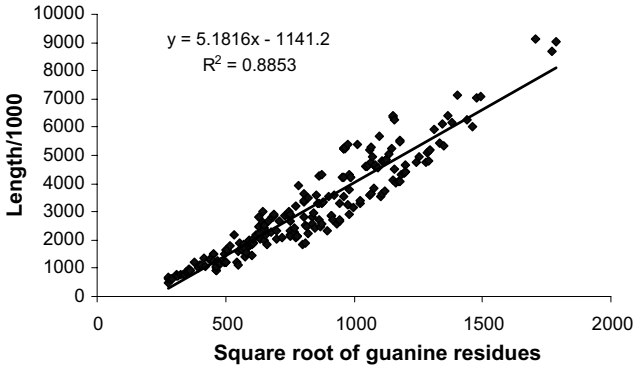

Fig. 3. Bacterial genome length divided by 1000 plotted against the square root of the guanine residue number. $N = 231$, $F = 1767.2$, $p < 10^{-108}$.

appear to govern the relationships between these variables, this suggests that there are common underlying factors—presumably of a thermodynamic nature—that await discovery. The $G^{0.5}$ term's presence in prokaryotic and dsDNA virus equations and its absence in the eukaryotic equation hint at fundamental differences in genome organisation between these groups. While this is obvious on biological grounds, precisely how this is translated into structural principles is not yet known.

In sum it seems that the relationship between genome GC content and length is not simple as had been proposed earlier and that this relationship differs significantly between eukaryotes and bacteria. As Musto et al. note [9] many influences act on genomic GC content of which this is but one. Uncovering the basis for this relationship will require additional theoretical and experimental work.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2006.12.008.

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. Filipski, J.P. Thiery, G. Bernardi, An analysis of the bovine genome by Cs_2SO_4–Ag density gradient centrifugation, J. Mol. Biol. 80 (1973) 177–197.
[5] J.P. Thiery, G. Macaya, G. Bernardi, An analysis of eukaryotic genomes by density gradient centrifugation, J. Mol. Biol. 108 (1976) 219–235.
[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. Biol. Sci. 268 (2001) 493–497.
[8] 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.
[9] H. Musto, H. Naya, A. Zavala, H. Romero, F. Alvarez-Valin, G. Bernardi, Genomic GC level, optimal growth temperature, and genome size in prokaryotes, Biochem. Biophys. Res. Commun. 347 (2006) 1–3.
[10] D.N. Gujarati, Basic Econometrics, third ed., McGrawHill, Singapore, 1995.
[11] H. White, A heteroscedacity consistent covariance matrix estimator and a direct test of heteroscedacity, Econometrica 48 (1980) 817–818.
[12] D. Mitchell, R. Bridge, A test of Chargaff's second rule, Biochem. Biophys. Res. Commun. 340 (2006) 90–94.
[13] M.T. McIntosh, R. Srivastava, A.B. Vaidya, Divergent evolutionary constraints on mitochondrial and nuclear genomes of malaria parasites, Mol. Biochem. Parasitol. 95 (1998) 69–80.
[14] A.A. Escalante, F.J. Ayala, Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences, Proc. Natl. Acad. Sci. USA 91 (1994) 11373–11377.