Available online at www.sciencedirect.com
SCIENCE DIRECT
Biochemical and Biophysical Research Communications 340 (2006) 90–94
BBRC
www.elsevier.com/locate/ybbrc

A test of Chargaff's second rule

David Mitchell [a,*], Robert Bridge [b,1]

[a] Vice Deanery of Genetics and Microbiology, Trinity College, Dublin, Ireland
[b] School of Chemistry, Trinity College, Dublin, Ireland

Received 21 November 2005
Available online 7 December 2005

Abstract

In 1968, Chargaff and his colleagues discovered a rule in *Bacillus subtilis*: in single stranded DNA, A = T and C = G. This rule has since been confirmed many times in other bacterial and eukaryotic genomes. To the best of our knowledge, this rule has not been tested before in either single stranded DNA or RNA genomes. Over 3400 genomic sequences were examined here and included for the first time both double and single stranded DNA and RNA genomes. We found that: (1) with the exception of the organellar DNA, this parity rule holds for all types of double stranded DNA genomes and (2) that this rule fails to hold for other types of genomes. The parity rule appears to be a selective force on genome evolution and codon use.

Keywords: DNA; Genome; Chargaff's rule

Chargaff and his colleagues [1,2] discovered that the base composition of single strands of DNA possessed similar relationships to those of double stranded DNA described earlier: to wit that A = T and G = C. The basis for the first rule was elucidated in the structure of DNA but that of the second remains elusive. Forsdyke and his colleagues have proposed a number of possible reasons for its existence [3]. Zhang and Zhang [4] have shown that this rule limits the GC content of the protein encoding sequences to lie between 20% and 80%.

Unlike the first parity rule, the second is not exact. Local deviations from parity were first described by Smithies et al. [5] and are normally small in magnitude [6]. These have been used to identify origins of replication in a number of organisms [7–10] and it has been proposed that these are due to both transcription and translational pressures [11,12].

In spite of the availability of many entire genomes, to the best of our knowledge, the presumed universal validity of this rule has never been examined before. When this rule has been tested, this to date has been only in double stranded DNA genomes. We have examined over three thousand chromosomal sequences from the viruses, archaea, bacteria, and eukaryotes, and have found that this parity rule is true for all double stranded DNA sequences examined with the exception of the organelle genomes. Instead, these genomes and the single stranded DNA viruses obey a different rule. The hepatoviruses obey the GC parity rule but not the AT parity rule. In the double stranded RNA sequences, C + T is approximately 50% but the parity rule is not obeyed. No identifiable parity rule applies to the remaining types of genomes.

Materials and methods

The genome sequences were downloaded from the NCBI server. The data set consisted of 2177 sequences from 1495 viral genomes, 835 organelle genomes, 231 bacterial, and 20 archaeal genomes. 164 sequences were examined from 15 eukaryotes: *Anopheles gambiense*, *Aradopsis thaliana*, *Candida albicans*, *Canis familiaris*, *Cryptococcus neoformans*, *Drosophilia*

* Corresponding author. Fax: +353 1 679 9294.
E-mail addresses: dmitchel@tcd.ie (D. Mitchell), bridgero@tcd.ie (R. Bridge).
[1] Fax: +353 1 671 2826.

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.11.160

melanogaster, *Encephalitozoon cuniculi*, *Eremothecium gossypii*, *Kluyveromyces lactis*, *Homo sapiens*, *Leishmania major*, *Plasmodium falciparum*, *Schizosaccharomyces pombe*, *Trypanosoma brucei*, and *Yarrowia lipolytica*. The GenBank Accession Nos. are given in the supplemental material.

The length and the A, C, G, and T counts were determined directly from the sequences. In addition to the NCBI number and the name of the sequence, it was also noted whether the sequences were circular or linear, single or double stranded, and if the genome was fragmented. For viruses the presence of an envelope, the number and type of hosts (plant, vertebrate, invertebrate, bacterial, and archaea) were also noted. The number of bases other than those present in the sequence was counted separately. Statistical analysis was done with Microsoft Excel 2003.

Results

A number of sequences had to be removed from the data set because they contained a significant proportion of unidentified bases: specifically the *H sapiens* chromosomes 9, 21, 22, and Y and *C. familiaris* chromosome 16 and 35 in which the proportion of unidentified bases exceeded 10%. The sequences were then divided into thirteen groups: bacteria, archaea, nuclear eukaryote, organellar (plastid and mitochondrial), double stranded DNA viruses, double stranded RNA viruses, partly double stranded DNA viruses, single stranded DNA viruses, satellite DNA viruses, satellite RNA viruses, negative strand RNA viruses, positive strand RNA viruses, and viroids. All these types of genomes are distinct and this division was chosen to identify the characteristics of each group.

Among the bacteria, the GC content varied from 22.5% in *Wigglesworthia* to 72.1% in *Streptomycetes* close to the limits estimated by Zhang and Zhang [4], and the parity rule was upheld throughout this range. Variation in the GC content among the archaeal genomes was more modest ranging from 36.0% in *Picrophilus* to 67.9% in *Halobacterium*. The GC range in the eukaryotic sample was larger than that of the archaea running from 18.5% in *P. falciparum* to 63.1% in *L. major*. Given the disparity in the size of samples, it is possible that the eukaryote and archaeal samples underestimate their real respective ranges. All three of these groups upheld the parity rule (Fig. 1).

The GC range in the organelles varied between 13.3% (*Aleurodicus dispersus*) and 52.2% (*Chirocentrus dorab*) but in contrast to other groups, the parity rule was not upheld. The AT plot shows a significant correlation ($p < 10^{-8}$) but the plot is markedly heteroskedastic and accordingly this result should be viewed with suspicion. The GC plot (Fig. 2) shows no significant correlation ($p > 0.1$). Plotting A against G (Fig. 3) and C against T gave a different picture of the relationships between the bases: to a first approximation the A + G and C + T content of the genomes were both equal to 50%.

These regressions (A–G and C–T) were also estimated for the bacterial, archaeal, and eukaryote genomes and not unexpectedly the correlations were found to be significant since approximate equality is a mathematical consequence of the parity rule. The difference between the regressions in the bacterial/archaeal groups and the eukaryotes is explainable in the proportion of bases that were not identified. The eukaryote sequences that had little ambiguity had an A–G coefficient of approximately −1 like bacterial and archaea, while those with a greater proportion lay further from this line (Fig. 3).

Four hundred and thirty-six double stranded DNA viral sequences were examined and the parity rule was found to hold in all but with a reduced R^2 compared with the other genomes. Residual variability was not significantly reduced by including dummy variables for the presence of an envelope, a segmented genome, a circular genome or for the type or number of hosts infected. The length of the genome was significantly negatively correlated in the single strand DNA viruses ($t = 4.78$, $p < 0.01$) but since the adjusted R^2 was increased only by ~1% after including this additional variable this raises questions about the biological significance of this particular correlation.

Three hundred and forty-one single stranded DNA virus genomes were examined and neither AT nor GC parity was found. In particular, the AT correlation failed to reach significance. In contrast, the A + G was approximately equal to C + T (Fig. 4).

Thirty-one hepatovirus genomes were also examined. These genomes are partly double stranded DNA and replicate via an RNA intermediate. [13] While the AT regression here failed to reach statistical significance, the GC content obeyed the parity rule. The AG correlation was sig-

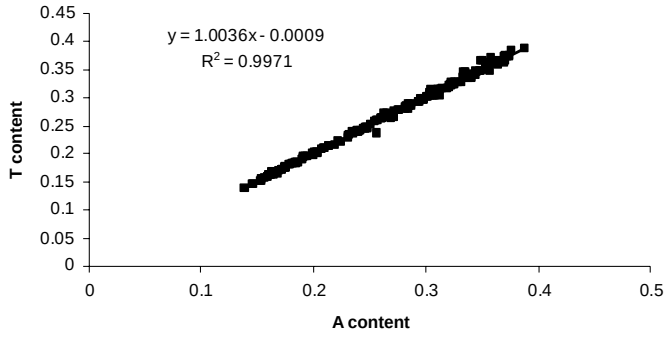

Fig. 1. Plot of thymidine (T) content against adenosine (A) in 231 bacterial genomes. $F = 85{,}447.40$, $p < 10^{-166}$.

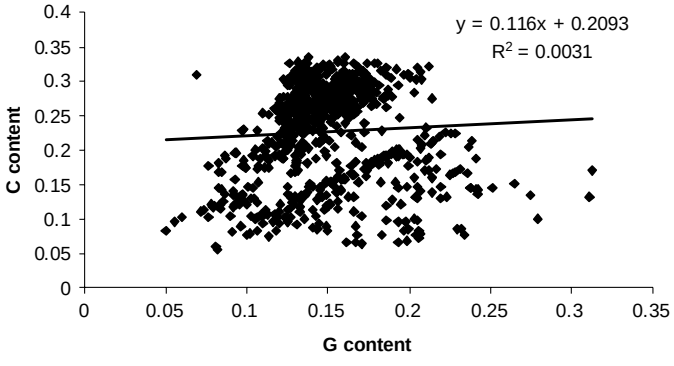

Fig. 2. Plot of guanine (G) content against cytosine (C) in 835 complete organellar genomes. $F = 2.6$, $p > 0.1$.

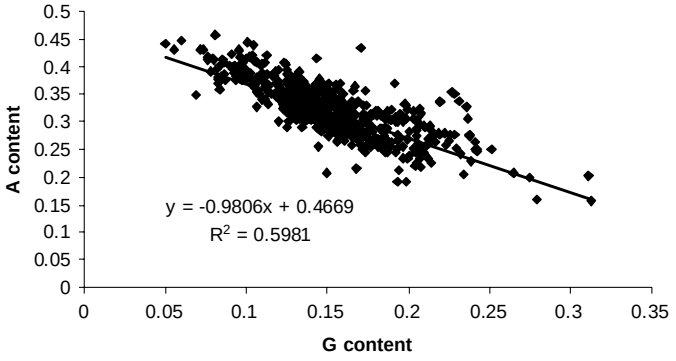

Fig. 3. Plot of the guanine (G) content against adenosine (A) in 835 complete organellar genomes. $F = 1{,}239.5$, $p < 10^{-130}$.

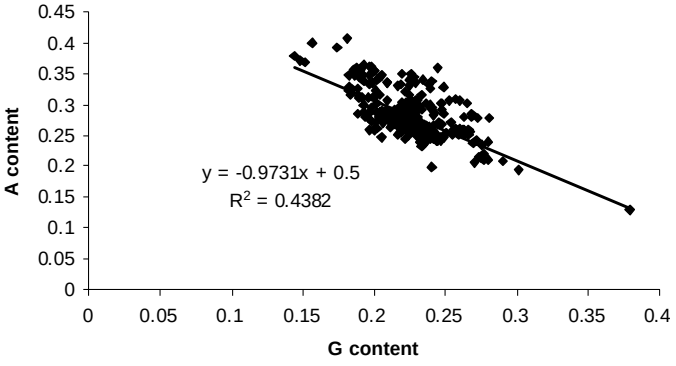

Fig. 4. Plot of the guanine (G) content against cytosine (C) in 340 single stranded DNA viral genomes. $F = 236.7$, $p > 10^{-43}$.

nificant, albeit with a regression coefficient greater than 1 but the TC correlation was not significant.

These results differed from those seen in the three hundred and five double stranded RNA sequences examined. Here, the AT correlation was significant but the plot is moderately heteroskedastic. The GC correlation failed to reach significance but both the AG and CT (Fig. 5) correlations were significant.

The positively and negatively single stranded viruses were considered separately in case strandedness should influence the results. Five hundred and seventeen positive strand and two hundred and forty negatively strand viral sequences were examined. In both sets of sequences, significant correlations were found in all four sets of regressions.

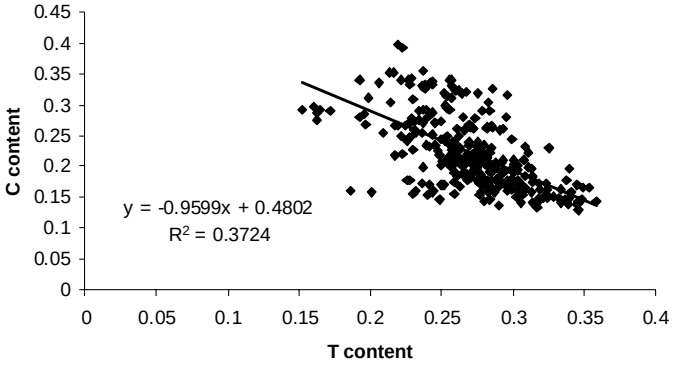

Fig. 5. Plot of thymidine (T) content and cytosine (C) in 305 double stranded RNA sequences. $F = 179.8$, $p < 10^{-31}$.

The viroids and satellite viruses are small single strand DNA and RNA virus like agents that do not encode proteins.[14] The viroids known to date infect only plants and unlike satellite viruses, the infectious material is naked RNA. Seventy-eight satellite virus (43 DNA and 35 RNA) and thirty six viroid genomes were examined. A summary of these correlation coefficients is presented in Table 1.

Discussion

The parity rule was found to hold for four of the five types of double stranded DNA genomes examined here: bacterial, archaeal, nuclear eukaryote, and viral, and it fails to hold for organellar DNA. In these latter genomes, another rule holds: to a first approximation $A + G = C + T = 0.5$. This second rule holds automatically in other genomes as a mathematical consequence of the parity rule but also holds in the single stranded DNA viruses where the parity rule fails. Only the GC parity rule holds for the partially double stranded viruses. Why this approximate equality of purines and pyrimidines in these DNA genomes should hold is not clear.

The organelles are thought to have been derived from free-living bacterial ancestors. [15] Given the universality of the parity rule in extant bacteria, it is likely that the ancestors of the organelles also obeyed Chargaff's rule and this rule no longer applies in the organelles. This rule imposes some form of evolutionary restraint on the double stranded genomes that has been relaxed in the organelles and instead replaced with a less stringent condition: that the $A + G$ and $C + T$ content of the genomes are approximately equal—a rule that is also obeyed by the single stranded DNA viruses.

The poor correlation in the eukaryote plots was unexpected and it would appear likely that the as yet unidentified bases in these genomes are likely to be largely G and T.

The remaining genomes are more difficult to understand. The good correlation between the G and C content in the hepatoviruses contrasts strongly with the lack of a similar correlation between A and T. At the moment no explanation is forthcoming.

The lack of such correlations in the double stranded virial genomes suggests that if an evolutionary constraint imposed by the parity rule does exist that it does not apply to RNA genomes. Presently an explanation for this quite marked difference is not clear. The satellite viruses and viroids do not as a rule encode proteins and so are likely to be under different constraints to the other viral genomes. If the proximate cause of this parity rule is connected with the protein encoding genes in some as yet unknown fashion then the lack of correlations between the base composition found here may be more understandable. Alternatively there may be another explanation that awaits discovery.

Table 1
Summary of the correlations. Correlation coefficients of the relationships between the bases in the several genome types

Sequence type	A–T		G–C		A–G		C–T	
	Coeff	R^2	Coeff	R^2	Coeff	R^2	Coeff	R^2
Bacteria	1.003	0.997	1.000	0.997	−0.996	0.999	−0.996	0.977
Archaea	0.989	0.997	0.992	0.997	−0.998	0.999	−0.998	0.999
Eukaryote	0.997	0.990	0.993	0.988	−0.899	0.787	−1.018	0.874
Organellar	0.145	0.040	0.116*	0.003	−0.981	0.598	−1.100	0.857
ds DNA	0.992	0.752	1.020	0.855	−1.047	0.892	−1.011	0.372
ss DNA	−0.065*	0.004	0.570	0.192	−0.973	0.438	−1.016	0.262
Partial ds DNA	−0.81*	0.114	0.998	0.435	−1.567	0.414	0.237*	0.025
Sat DNA	−0.658	0.244	−0.119*	0.020	−0.864	0.412	−0.454	0.149
ds RNA	0.461	0.128	0.077*	0.002	−0.554	0.178	−0.960	0.372
ss +ve RNA	0.087	0.010	−0.197	0.010	−0.189	0.062	−0.875	0.455
ss −ve RNA	−0.191	0.044	0.177	0.036	−0.124	0.019	−0.144	0.037
Sat RNA	0.163*	0.026	0.432*	0.010	−0.721	0.598	−0.563	0.505
Viroids	0.300	0.154	0.558	0.465	−0.660	0.390	−0.683	0.595

Abbreviations: Coeff, correlation coefficient of the second base against the first (Bacterial A–T: A = 1.003 T + constant); R^2, adjusted R^2 value for the correlation; bacterial, bacterial genomes; archaea, archaeal genomes; eukaryote, eukaryote genomes; organellar, plastid and mitochondrial genomes; ds DNA, double stranded viral DNA genomes; ss DNA, single stranded viral genomes; partial ds DNA, partly double stranded DNA genomes; Sat DNA, satellite DNA virus genomes; ds RNA, double stranded viral RNA genomes; ss +veRNA, single positive strand viral sequences; ss −ve RNA, single negative strand viral sequences; Sat RNA, satellite RNA virus genomes; viroids, viroid genomes.

* Correlation with $p > 0.05$.

Szybalski et al. [16] described a transcriptional rule that has been confirmed many times since: that in coding sequences purines (A + G) exceed pyrimidines (C + T). This rule in combination with the second parity rule has acted to shape genomes. Consider a hypothetical gene-rich genome (90% coding). As purines are in excess on the coding strand, their complementary pyrimidines will be in excess in the noncoding strand. To obey the second parity rule, the genes must either be distributed approximately equally between the strands or have a differential use of codons on the strands. To see this more clearly, consider a genome where all the coding sequences lie on one strand. On the coding strand as a consequence of the transcriptional rule A + G content would exceed the C + T and violate the parity rule. An approximately symmetrical distribution of coding sequences allows for greater flexibility in codon use and this is the pattern found in many bacteria.

In the bacterium *Treponema pallium*, there is an unequal distribution of the coding sequences with 64% of the coding sequences lying on the leading strand [17]. Consistent with the above hypothesis codon use is very biased in this genome [9]. In the related spirochete *Borrelia burgdorferi* the coding sequences are approximately equally distributed on the two strands [19]. Codon use in this bacterium is also biased suggesting that the parity rule is but one influence on codon use, albeit one that has not been considered before [18].

A second consequence of these rules suggests a possible reason for the evolution of introns. Consider an ancestral organism whose genome is growing in complexity as it acquires new genes. As the number of genes increases so does the risk of a violation of the parity rule. To prevent violation of this rule, either genes must be inserted to maintain approximately equal coding lengths in both strands, codon bias must evolve rapidly, the intergenic spaces must be comparatively large enough to act as a "buffer" to reduce the impact of parity violations or the genes themselves must carry their own pyrimidine "buffers" in the form of introns. These mechanisms are not mutually exclusive and any combination of them may have evolved and should not be understood to mean that introns evolved 'early' or 'late' in evolution but rather as a reason for their original inclusion into genes.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2005.11.160.

References

[1] R. Rudner, J.D. Karkas, E. Chargaff, Separation of *B. subtilis* DNA into complementary strands, III, Proc. Natl. Acad. Sci. USA 60 (1968) 921–922.
[2] E. Chargaff, Structure and function of nucleic acids as cell constituents, Fed. Proc. 10 (1951) 654–659.
[3] J.D. Watson, F.H.C. Crick, A structure for deoxyribose nucleic acid, Nature 171 (1953) 737–738.
[4] D.R. Forsdyke, J.R. Mortimer, Chargaff's legacy, Gene 261 (2000) 127–137.
[5] C.-T. Zhang, R. Zhang, A nucleotide composition constraint of genome sequences, Comp. Biol. Chem. 28 (2004) 149–153.
[6] O. Smithies, W.R. Engels, J.R. Devereux, J.L. Slightom, S. Shen, Base substitutions, length differences and DNA strand asymmetries in the human G gamma and A gamma fetal globin gene region, Cell 26 (1981) 345–353.
[7] D.M. Prescott, S.J. Dizick, A unique pattern of intrastrand anomalies in base composition of the DNA in hypotrichs, Nucleic Acids Res. 28 (2000) 4679–4688.
[8] J.R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria, Mol. Biol. Evol. 13 (1996) 660–665.

[9] B. Lafay, A.T. Lloyd, M.J. McLean, K.M. Devine, P.M. Sharp, K.H. Wolfe, Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases, Nucleic Acids Res. 27 (1999) 1642–1649.
[10] A. Grigoriev, Strand-specific compositional asymmetries in double-stranded DNA viruses, Virus Res. 60 (1999) 1–19.
[11] M.J. McLean, K.H. Wolfe, K.M. Devine, Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes, J. Mol. Evol. 47 (1998) 691–696.
[12] M.P. Francino, H. Ochman, Strand asymmetries in DNA evolution, Trends Genet. 3 (1997) 240–245.
[13] P.H. Hayashi, J.B. Zeldis, Molecular biology of viral hepatitis and hepatocellular carcinoma, Compr. Ther. 19 (1993) 188–196.
[14] B. Ding, A. Itaya, X. Zhong, Viroid trafficking: a small RNA makes a big move, Curr. Opin. Plant Biol. 8 (2005) 606–612.
[15] J. Filee, P. Forterre, Viral proteins functioning in organelles: a cryptic origin? Trends Microbiol. 13 (2005) 510–513.
[16] W. Szybalski, H. Kubinski, P. Sheldrick, Pyrimidine clusters on the transcribing strands of DNA and their possible role in the initiation of RNA synthesis, Cold Spring Harb. NY Symp. Quant. Biol. 31 (1966) 123–127.
[17] C.M. Fraser, S.J. Norris, G.M. Weinstock, O. White, G.G. Sutton, R. Dodson, M. Gwinn, E.K. Hickey, R. Clayton, K.A. Ketchum, E. Sodergren, J.M. Hardham, M.P. McLeod, S. Salzberg, J. Peterson, H. Khalak, D. Richardson, J.K. Howell, M. Chidambaram, T. Utterback, L. McDonald, P. Artiach, C. Bowman, M.D. Cotton, C. Fujii, S. Garland, B. Hatch, K. Horst, K. Roberts, M. Sandusky, J. Weidman, H.O. Smith, J.C. Venter, Complete genome sequence of *Treponema pallidum*, the syphilis spirochete, Science 281 (1998) 375–388.
[18] J.O. McInerney, Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*, Proc. Natl. Acad. Sci. USA 95 (1998) 10698–10703.
[19] C.M. Fraser, S. Casjens, W.M. Huang, G.G. Sutton, R. Clayton, R. Lathigra, O. White, K.A. Ketchum, R. Dodson, E.K. Hickey, M. Gwinn, B. Dougherty, J.F. Tomb, R.D. Fleischmann, D. Richardson, J. Peterson, A.R. Kerlavage, J. Quackenbush, S. Salzberg, M. Hanson, R. van Vugt, N. Palmer, M.D. Adams, J. Gocayne, J. Weidman, T. Utterback, L. Watthey, L. McDonald, P. Artiach, C. Bowman, S. Garland, C. Fuji, M.D. Cotton, K. Horst, K. Roberts, B. Hatch, H.O. Smith, J.C. Venter, Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*, Nature 390 (1997) 580–586.