Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 303 (2003) 65–68
BBRC
www.elsevier.com/locate/ybbrc

Reinvestigation on the causes of genomic GC variation between the orthologous genes of *Mycobacterium tuberculosis* and *Mycobacterium leprae*

S.K. Gupta and T.C. Ghosh*

Distributed Information Centre, Bose Institute, P 1/12, C.I.T. Scheme, VII M, Kolkata West Bengal 700 054, India

Received 16 January 2003

Abstract

Genomic GC (overall G + C content of the coding sequences) variations were reinvestigated between the orthologous genes of *Mycobacterium tuberculosis* and *Mycobacterium leprae* species. It was observed that overall genomic GC variation between the species mainly originates from the combined effects GC_1 and GC_2 variations. But codons having identical amino acids with different codons (IA) (between the orthologous codon pairs) are responsible for the genomic GC_3 variation between the organisms, whereas orthologous codons having different amino acids (DA) between the two organisms are responsible for the variation of GC_1 levels. Further analyses indicate that duets and quartets are going in the same direction with same magnitude in changing the GC_3 levels for IA category, whereas GC_1 levels of duets of DA category decreases significantly from the overall GC_1 levels but GC_1 levels of quartets increases significantly from the overall GC_1 levels. GC_3 levels of informational genes for the IA category decrease more rapidly than the other functional categories of genes. The biological implications of these results have been discussed in this paper.

The genomic (G + C) content among the prokaryotes varies substantially from 25% to 75% [1,2]. But the causes of inter-species variation are not clearly understood. It has been proposed that the increment of GC% could be advantageous for the organisms belonging to the thermophilic group or those that are exposed to UV radiation [3,4]. However, there is a lot of controversy in this line of argument [5]. Naya et al. [6] recently reported that aerobiosis increases genomic GC% in prokaryotes. Recently in-depth analyses performed by Bellgard and Gojobori [7] by taking orthologous codons between closely related species concluded that identical amino acids having different synonymous codons took the main role for the differentiation of the genomic GC variation among the bacteria and subsequently they argued that synonymous third positions of codons are hot spots for the overall GC variation among the bacteria. Further by analyzing the two strains of *Helicobacter pylori* Bellgard et al. [8] observed that codons belonging to either three, four, or six codon boxes are the main operational units in changing the GC_3 content in identical amino acids with different synonymous codons. But Bellgard et al. have performed their analyses only at the third positions of codons.

Here in this paper we have made detailed studies by taking orthologous sequences of *Mycobacterium tuberculosis* and *Mycobacterium leprae* genes, in three codon positions separately, in order to gain more insight into the genomic GC variation among the prokaryotic organisms.

Materials and methods

Eight hundred and fifty-eight orthologous gene pairs of *M. tuberculosis* and *M. leprae* were extracted from Cluster of Orthologus Genes (COG) database, available at NCBI (http://www.ncbi.nlm.nih.gov/COG). CLUSTALW [9] was used to align the protein sequences. Codon-based alignment program developed by us was used to align DNA sequences. Each pair of orthologous genes was divided into three categories according to the method of Bellgard and Gojobori [7]. IA represents identical amino acids having different codons for each

* Corresponding author. Fax: +91-2334-3886.
E-mail address: tapash@boseinst.ernet.in (T.C. Ghosh).

0006-291X/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0006-291X(03)00283-3

species; DA represent different amino acid having different codons for each species and IC represent identical amino acid having identical codons. Thus for a orthologous pair of gene there are five sets of sequences IC, IA1, DA1, IA2, and DA2, where 1 and 2 are two species of the orthologous pair.

GC plot, i.e., a plot of genomic GC (G + C content of the coding sequences) with each of GC_1 (G + C content at the first codon position), GC_2 (G + C content at the second codon position) and GC_3 (G + C content at the third codon position) was constructed using the bacterial and archeabacterial complete genomes currently available from NCBI (http://www.ncbi.nlm.nih.gov/genomes/data).

Translation, transcription, DNA replication, recombination, and repair genes were considered as information storage and processing genes as described in COG database [10].

The base compositions were calculated using codonW1.3 (J. Peden; http://molbiol.ox/ac.uk/Win95.codonW.zip). The *t* test was used to evaluate the significance of the pair-wise differences in nucleotide composition.

Results and discussion

GC plot with three different codon positions using available complete bacterial genomes

In Fig. 1 GC levels of first, second, and third codon positions were plotted against the overall (G + C) contents of the corresponding genome. From Fig. 1 it is evident that the GC contents of all the three codon positions have a strong positive linear correlation with the overall GC contents. The slopes for the three different codon positions are different and the magnitude of the slopes increases in the order of III > I > II. The most interesting point is that the pattern of relationships of GC contents of three different codon positions against the overall GC contents as observed by Muto and Osawa [2] with a limited number of sequences as well as by Bellgard and Gojobori [7] with 15 complete prokaryotic complete genomes is still valid when we did the analysis with 78 available complete genome sequences. The strong correlation between the GC contents at the third codon positions and the overall GC contents was thought to be driven by biased mutation pressure since most of the substitution at this position is synonymous and is under very weak purifying selection. Majority of substitution at the first two codon positions results in non-synonymous substitution and it was argued that first and second codon positions undergo strong purifying selection and between the two first codon positions second one takes the most dominant role in determining the three dimensional structure of proteins. From the above discussion it is clear that among the three codon positions, third and to some extent first codon positions are more prone to A.T/G.C substitution than the second codon position.

Overall compositional analyses

The average values of GC, GC_1, GC_2, and GC_3 for the *M. tuberculosis* and *M. leprae* and their corresponding values of IA, DA, and IC portions are shown in Table 1. The difference in overall GC levels between the two species is highly statistically significant ($p < 4.25 \times 10^{-250}$). From the table it is also evident that in *M. tuberculosis*, GC levels in all the three categories

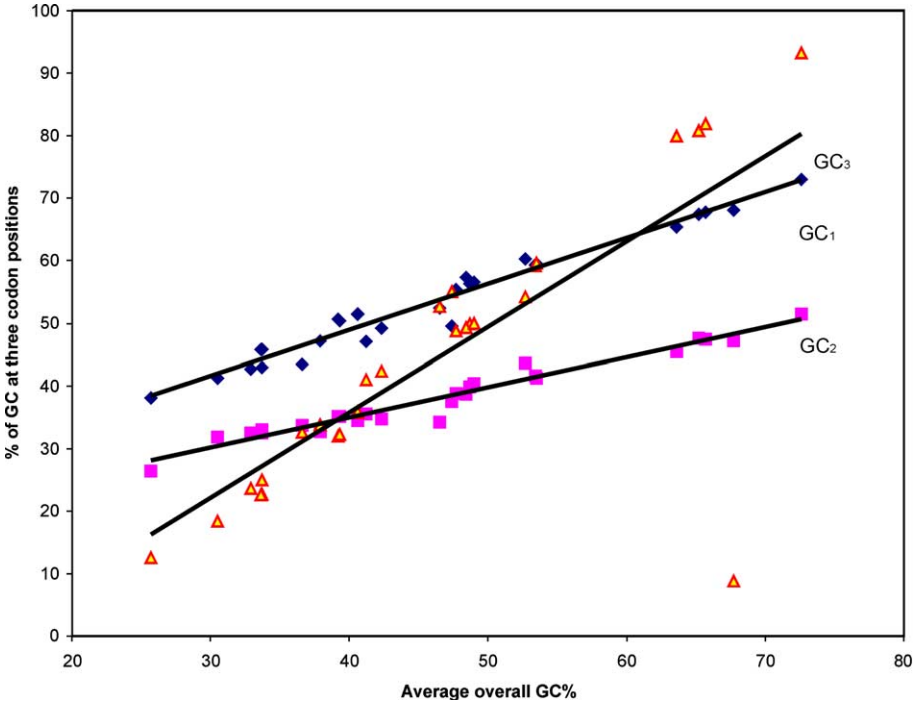

Fig. 1. Scatter plot of the GC levels between the overall genomic DNA and the first, second, and third codon positions.

Table 1
Average overall GC levels at three different codon positions as well as of the corresponding IA, DA, and IC categories of codons

	Mycobacterium tuberculosis	*Mycobacterium leprae*
Overall		
GC	65.7	60.2
GC1	68.5	65.2
GC2	47.9	46.2
GC3	80.5	69.2
IA		
GC	66.2	59.1
GC1	70.5	67.7
GC2	49.7	48.4
GC3	78.5	61.1
DA		
GC	65.9	58.3
GC1	68.6	61.7
GC2	49.3	45.1
GC3	79.7	68.1
IC		
GC	65.2	65.2
GC1	68.2	68.2
GC2	42.1	42.1
GC3	85.2	85.2

(i.e., IA, DA, and IC) do not vary significantly from the overall GC levels whereas in case of *M. leprae* it was observed that there are significant differences of GC levels in all the three categories particularly, in IC category. In IA and DA categories GC levels decrease whereas in IC category the level of GC increases than the overall GC level. All the pair-wise comparisons were statistically significant and the lowest p value was that of GC levels of IA category vs. overall GC levels of *M. leprae* ($p < 1.12 \times 10^{-10}$). The average sequence lengths of IA, DA, and IC are, respectively, 49.52%, 33.12%, and 17.35%. Therefore, the observed GC variation between the two species is mainly due to the combined effects of IA and DA categories, though IC category influences in a minor way as the amounts of IC category are appreciably smaller than those of the other two categories.

Compositional analyses at the first codon position

Significant differences in GC levels at the first codon positions exist between the two species ($p < 2.18 \times 10^{-42}$). For IA category there is a significant increase in GC levels at the first ($p < 4.29 \times 10^{-12}$) codon position, whereas for DA and IC categories no significant differences were observed in *M. tuberculosis*. In case of *M. leprae* it was also observed that GC levels at the first codon position increase significantly for the IA and IC categories. But there is a significant decrease in GC content ($p < 1.94 \times 10^{-27}$) at the first codon position for the DA category. It is to be noted that GC_1 levels of IC remain the same as the overall GC_1 levels of *M. tuberculosis* whereas in case of *M. leprae* GC_1 levels of IC category increase significantly from the overall GC_1 levels. Since the average length of DA is twice that of IC it can be concluded that DA category of *M. leprae* genes is taking the major role for decreasing the overall GC levels of the genes.

Compositional analyses at the second codon position

From Table 1 it is evident that marginal difference in overall GC levels at the second codon position was observed between *M. tuberculosis* and *M. leprae* genes. This is fully consistent with the earlier observations that the second codon positions are well conserved throughout the genome evolution and they do not take part in the overall GC variation between the bacterial species [11,12].

Since there is no significant difference of overall GC_2 levels between the two species we have not made any attempt to understand the variation of IA, DA, and IC from the overall GC_2 levels of each species (Table 1).

Compositional analyses at the third codon position

Significant differences in GC_3 levels between the two species were observed. In DA category no significant differences were observed from the overall GC_3 levels of each species, and in IC category there is a significant increase in both species. But in IA category there is a significant decrease in GC_3 levels in *M. leprae* whereas there is no significant change in GC_3 levels in *M. tuberculosis* genes. Therefore, the observed differences in GC_3 levels between the two species mainly originate from the IA category of genes. Bellgard and Gojobori [7] observed exactly the same results with 146 homologous genes between *M. tuberculosis* and *M. leprae* and argued that synonymous positions are the main spots for the overall genomic GC variation among the organisms.

Compositional effects in duets and quartets

In order to determine how the different codon families namely duets and quartets are affecting the GC_1 variation in DA category and GC_3 variation in IA category for *M. leprae* genes we have separated IA and DA of *M. leprae* genes into two groups. For simplicity we have included all the three duets of sextets (serine, arginine, and leucine) in duet family and all the three quartets of sextets and isoleucine codon family are in quartet family. It was observed that there is a large decrease in GC_1 levels of DA category for the duets (average value of GC_1 for duet is 50.01) and appreciable increases in GC_1 levels for the quartets (average levels of GC1 is 71.84). But GC_3 levels of IA category do not

change significantly between duets and quartets (average value of GC_3 for duets 59.10 and that for quartets is 61.23). Therefore, GC_3 levels of duets and quartets cannot be distinguished from the overall IA of *M. leprae* genes, which led us to the proposal that the genomic GC evolution at the third codon positions has no effect on the duet and quartet codon families. In other words we can say that genomic GC_3 variation took place in the same direction throughout the gene homogeneously for the IA category.

Compositional analyses in different functional groups of genes

In order to find how compositional variations in three different codon positions are affected by different functional categories of genes we have calculated compositional parameters for the IA, IC, and DA categories of informational as well as other functional categories of genes separately (data not shown) for both the species. Interestingly we observed that for IA category of *M. leprae* genes, GC_3 values for informational genes as well as other functional genes are significantly lower than the overall GC_3 values and between the two functional categories, GC_3 values of informational genes have much lower values than the other genes. This indicates that all the genes in *M. leprae* have not changed to the same extent and synonymous positions of informational genes are more prone to A.T/G.C substitution than the other genes. On the other hand for DA category GC_1 levels for the two functional groups of genes have significant difference from the overall GC_1 levels in *M. leprae* genes and GC_1 levels between the two functional groups do not change significantly from each other. This clearly demonstrates that different functional groups of genes have no preference on the GC_1 variation for DA category.

In conclusion it can be said that observed overall GC variation between the two species mainly originates from the GC levels at the first as well as at the third codon positions. GC levels at the second codon position between the two species do not vary significantly. Identical amino acids with different codons (IA) are responsible for the genomic GC_3 variation between the organisms, whereas different amino acids (DA) between the two organisms are responsible for the variation of GC_1 levels. GC_3 levels of duets and quartets are more or less same with the overall GC_3 levels of IA category, whereas GC_1 levels of duets of DA category decrease significantly from the overall GC_1 levels but GC_1 levels of quartets increase significantly from the overall GC_1 levels. GC_3 levels of informational genes for the IA category decrease more rapidly than the other functional categories of genes. These results suggest that strong selective pressure is operational in the evolutionary changes of the genomic GC contents.

Acknowledgment

Authors are thankful to the Department of Biotechnology, Government of India, for the financial help.

References

[1] N. Sueoka, Proc. Natl. Acad. Sci. USA 48 (1962) 582–592.
[2] A. Muto, S. Osawa, Proc. Natl. Acad. Sci. USA 84 (1987) 166–169.
[3] P. Argos, M.G. Rossman, U.M. Grau, H. Zuber, G. Frank, J.D. Tratschin, Biochemistry 18 (1979) 5698–6703.
[4] C.E. Singer, B.N. Ames, Science 170 (1970) 822–825.
[5] N. Galtier, J.R. Lobry, J. Mol. Evol. 44 (1997) 632–636.
[6] H. Naya, H. Romero, A. Zavala, B. Alvarez, H. Musto, J. Mol. Evol. 55 (2002) 260–264.
[7] M.I. Bellgard, T. Gojoboru, Gene 238 (1999) 33–37.
[8] M. Bellgard, D. Schibeci, E. Trifonov, T. Gojobori, J. Mol. Evol. 53 (2001) 465–468.
[9] D. Higgins, J. Thompson, T. Gibson, J.D. Thompson, D.G. Higgins, T.J. Gibson, Nucleic Acids Res. 2 (1994) 4673–4680.
[10] R.L. Tatusov, D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, E.V. Koonin, Nucleic Acids Res. 29 (2001) 22–28.
[11] S. Majumdar, S.K. Gupta, V.S. Sundrarajan, T.C. Ghosh, Biochem. Biophys. Res. Commun. 266 (1999) 66–71.
[12] S.K. Gupta, S. Majumdar, T.K. Bhattacharya, T.C. Ghosh, Biochem. Biophys. Res. Commun. 269 (2000) 692–696.