Biochemical and Biophysical Research Communications **266,** 66–71 (1999)
Article ID bbrc.1999.1774, available online at http://www.idealibrary.com on IDEAL®

Compositional Correlation Studies among the Three Different Codon Positions in 12 Bacterial Genomes

S. Majumdar, S. K. Gupta, V. S. Sundararajan, and T. C. Ghosh[1]
Distributed Information Centre, Bose Institute, P 1/12, C.I.T. Scheme, VII M Calcutta 700 054, India

Received October 26, 1999

Compositional distributions in the three codon positions of the coding sequences of 12 fully sequenced prokaryotic genomes, which are publicly available, were investigated. A universal compositional correlation was observed in most of the genomes under investigation irrespective of their overall genomic GC contents. In all the genomes, the GC contents at the first codon positions are always greater than the overall GC contents of the genomes whereas the reverse is true in the case of second codon positions. GC contents at the third codon positions are higher than the overall genomic GC contents in high GC containing genomes, and the opposite situation was found in case of low GC genomes except for *Helicobacter pylori*. In high-GC rich genomes, the GC contents at the first + second codon positions are less than the GC contents at the third codon positions, and they are low in low-GC genomes except for *Helicobacter pylori*. The distributions of four bases at the three different positions were also investigated for all 12 organisms. It was observed that in high-GC genomes G is the most dominant base and in low-GC genomes A is the most dominant base in the first codon positions. But purine bases, i.e., (A + G), predominantly occur in the first codon position. In the second codon position, A is the most dominant base in most of the organisms and G is the least dominant base in all the organisms. There is no unique regular pattern of individual bases at the third codon positions; however, there are significant differences in the occurrences of (G + C) contents in the third codon positions among the different organisms. Calculations of dinucleotide frequencies in 12 different organisms indicate that in GC-rich genomes GG, GC, CC, and CG dinucleotides are the most dominant whereas the reverse is true in case of low-GC genomes. Biological implications of these results are discussed in this paper.

The number of available complete genomes is increasing at a fast pace and as a result it imposes a great challenge to the computational molecular biologist to reveal many fundamental questions hidden in the living system. The average guanine and cytocine content (G + C) of genomic DNA varies widely among the bacteria. It has also been suggested that the bacterial genomic (G + C) content is related to phylogeny (2, 3). GC levels of the first, second as well as the third codon positions is positively correlated with overall genomic GC levels (4). This type of general relationship is found not only inter-genomically but also intra-genomically and it exists both in highly compartmentalized genomes and weakly compartmentalized genomes (5–7). However, the order of correlation value differs among the three different codon positions. The magnitude of correlation values is highest in case of third codon positions and lowest in case of second codon position (8–10). It was also observed that the correlations between the GC levels of the third and first + second codon positions are linearly correlated in both the highly and weakly compartmentalized genomes (7). All these studies were carried out with a limited number of sequences. Availability of complete genomes made a tremendous opportunity to reinvestigate the compositional correlations among the three different codon positions. We have extracted the complete coding sequences from 12 bacterial genomes covering from low GC contents to high GC contents and made a thorough study on compositional correlations among the different codon positions with an intention that these results may shed light on the general common relationship as mentioned above.

MATERIALS AND METHODS

Complete genomes of *Mycobacterium tuberculosis, Treponema pallidum, Archaeoglobus fulgidus, Escherichia coli K-12, Aquifex aeolicus, Bacillus subtilis, Chlamydia trachomatis, Helicobacter pylori, Haemophilus influenza Rd, Mycoplasma genitalium, Methanococcus jannaschi,* and *Borrelia burgdorferi* of varying genomic GC content were downloaded from ncbi.nlm.nih.gov/genbank/genomes by anonymous ftp. For each individual bacteria we have extracted the coding

[1] To whom correspondence should be addressed. Fax: +91-334-3886. E-mail: tapash@boseinst.ernet.in.

0006-291X/99 $30.00

TABLE 1

List of 12 Organisms, Their Identification Code, Number of Coding Genes, and GC% at the Different Codon Positions

Name of organism	Identification code	No. of coding sequences	GC% at (1st + 2nd + 3rd) codon position	GC% at 1st codon position	GC% at 2nd codon position	GC% at 3rd codon position	GC% at (1st + 2nd) codon position
Mycobacterium tuberculosis	MT	4000	65.79 (3.33)	68.04 (5.22)	50.10 (5.89)	79.21 (5.40)	59.07 (4.43)
Treponema pallidum	TP	1041	53.15 (4.36)	60.14 (5.61)	44.27 (5.96)	55.05 (6.65)	52.20 (4.63)
Escherichia coli K-12	EC	4288	51.28 (4.9)	58.39 (6.42)	40.67 (5.03)	54.78 (8.44)	49.53 (4.43)
Archaeoglobus fulgidus	AF	2436	48.99 (3.93)	52.76 (5.00)	36.23 (5.17)	57.98 (7.27)	44.49 (3.72)
Aquifex aeolicus	AA	1512	43.70 (3.90)	50.50 (5.22)	32.47 (4.94)	48.14 (6.35)	41.48 (4.06)
Bacillus subtilis	BS	4100	43.58 (4.61)	51.65 (6.0)	35.32 (5.49)	43.77 (7.79)	43.49 (4.52)
Chlamydia trachomatis	CT	894	41.72 (2.3)	51.58 (4.24)	38.77 (4.45)	34.82 (4.09)	45.17 (3.06)
Helicobacter pylori	HP	1590	39.34 (3.77)	44.29 (5.06)	32.12 (5.17)	41.61 (5.72)	38.20 (3.86)
Haemophilus influenzae Rd	HI	1743	38.62 (3.71)	50.61 (5.88)	35.93 (4.95)	29.32 (6.12)	43.27 (4.34)
Mycoplasma genitalium	MG	470	31.88 (3.59)	41.64 (4.94)	30.36 (5.06)	23.63 (6.27)	36.00 (3.88)
Methanococcus jannaschii	MJ	1738	31.91 (4.06)	41.28 (6.08)	29.48 (6.10)	24.96 (4.79)	35.38 (5.1)
Borrelia burgdorferi	BB	853	28.60 (3.93)	36.84 (6.77)	27.92 (5.66)	21.04 (3.27)	32.38 (5.41)

Note. Values in parentheses indicate standard deviation.

sequences by our own program developed in C++. We have not made any attempt to remove the ORFs of unknown functions. The base composition at the different codon positions was calculated by using GCUA (General Codon Usage Analysis) developed by James McInernay (1). All the figures were generated by using Microsoft Excel.

RESULTS AND DISCUSSION

Compositional Pattern (GC Levels) Analysis

The compositional patterns, i.e., GC distributions, at the three different codon positions as well as average GC levels, the average GC levels of first + second codon positions were calculated. Table 1 displays the name of the organisms, number of genes examined and GC distributions at the three different codon positions etc. From Table 1 it is obvious that the interspecies variation of GC levels (21.04–79.21) is highest in the third codon positions whereas it is lowest (27.92–50.10) in the second codon positions. The standard deviation of the GC levels of third codon positions is always greater than first + second codon positions and the coding sequences. This indicates that the variations of GC levels in the third codon positions is not as homogeneous as in the case of other codon positions even within the same species. In high-GC-rich genomes GC levels at the third codon positions is always higher than the second codon positions as well as the first + second codon positions whereas the opposite trend was observed in case of low-GC genomes. The transition between these two situations occurred at around 44% GC levels of the coding sequences. But an exception was observed in case of *Helicobacter pylori,* where it was found that GC levels at the third codon positions is higher than the second codon positions though the average GC level of the coding sequences is lower than 44%.

The abnormal compositional behavior of *Helicobacter pylori* is not surprising, since it was reported earlier that *Helicobacter pylori* genome is distantly related to the genomes of *Mycoplasma genitalium, Methanococcus jannaschii* (11).

Figure 1 shows the histograms of the compositional distributions of first, second and third codon positions of highest GC containing genome *Mycobacterium tuberculosis* and lowest GC containing genome *Borrelia burgdorferi* respectively. From the figure it is evident that the distributions of GC in the third codon positions lie in range of 40% (60%–100%) having peaks at around <80% and <90% in the highest GC rich genomes *Mycobacterium tuberculosis* whereas in the lowest GC containing genomes *Borrelia burgdorferi* the distributions of GC in the third codon positions lie in the range of 20% (20%–40%) having peaks at around <20% and <30%. The intermediate cases (data not shown) demonstrate that the distributions of GC levels at the third position of the codons vary directly with the genomic GC contents. GC levels at the first position of the codons are always higher than the second and third positions of the codons in all the genomes except *Mycobacterium tuberculosis* and *Archaeoglobus fulgidus.* It was also observed that the order of GC levels among the three codon positions is I > III > II for most of the GC-rich genomes and for all GC-poor genomes the order is I > II > III which is fully consistent with the earlier observations (10). The reason behind this is that in the low-GC genomes occurrence of amino acids

TABLE 2

Standard Deviation of the Frequencies of the Four Bases at Three Different Codon Positions

Codon position	G	A	T	C
1st	4.3	6.39	3.07	5.34
2nd	2.77	4.68	1.88	2.81
3rd	8.46	9.49	8.55	9.64

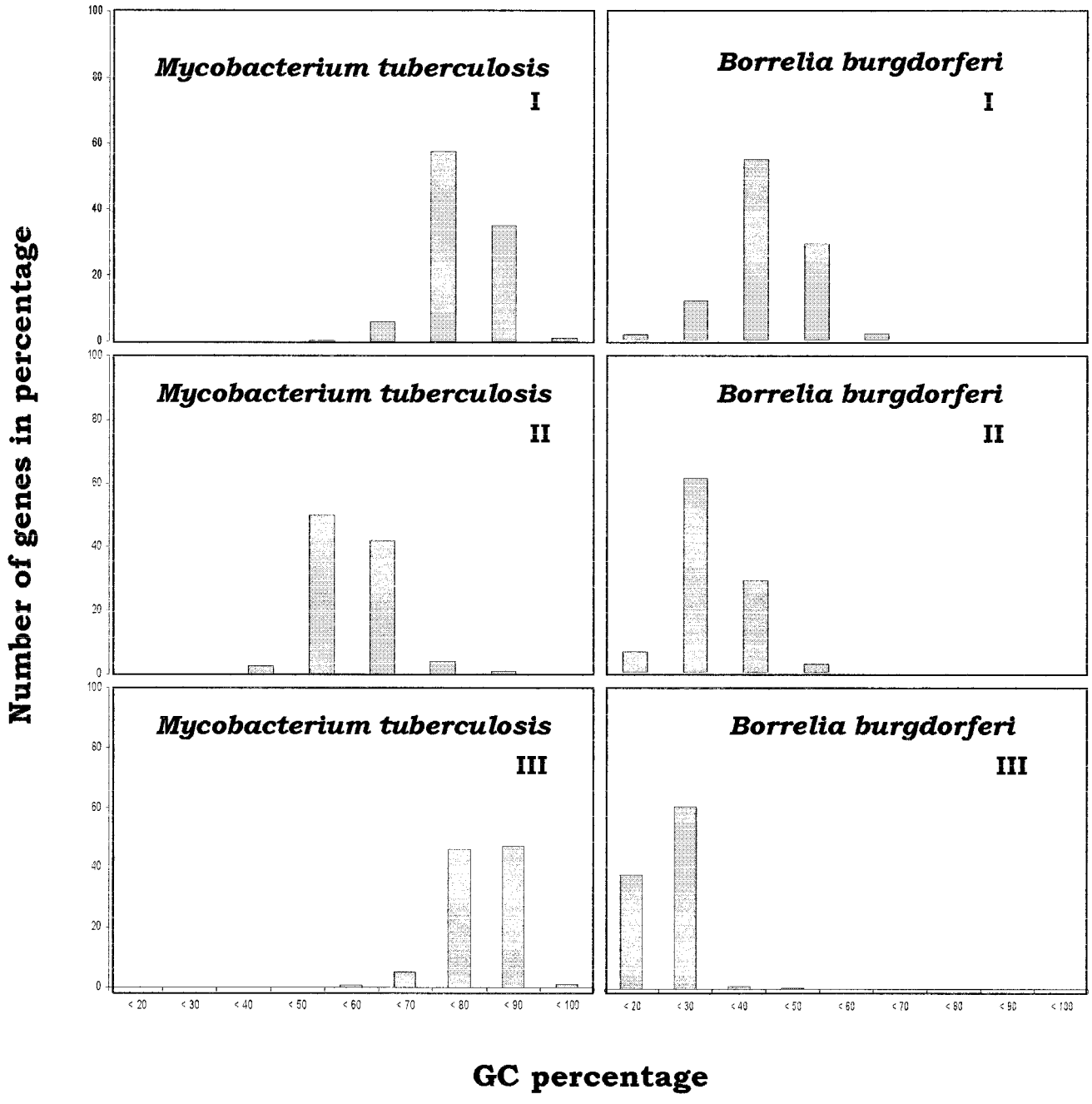

FIG. 1. Compositional distributions (GC levels) of three different codon positions (denoted as I, II, and III) of *Mycobacterium tuberculosis* and *Borrelia burgdorferi* genomes. GC levels and the number of sequences both are expressed as percentages and are drawn along the abscissae and ordinates, respectively.

having G or C at the first codon positions and A or T at the second codon position is more than the high-GC genomes as speculated by the earlier observers (10).

Mononucleotide Frequency Analysis

The frequency distributions of four base A, T, G and C at three different codon positions have also been calculated for 12 different bacteria. Figures 2–4 display the frequency plots of four bases at the three different sites of codons. From Fig. 2 it is evident that G is the most dominant base in all the high-GC containing genomes except in case of *Helicobacter pylori*, whereas in case of low-GC genomes A is the most frequent base in the first position of the codons. Purines are predominant over pyrimidine in the first po-

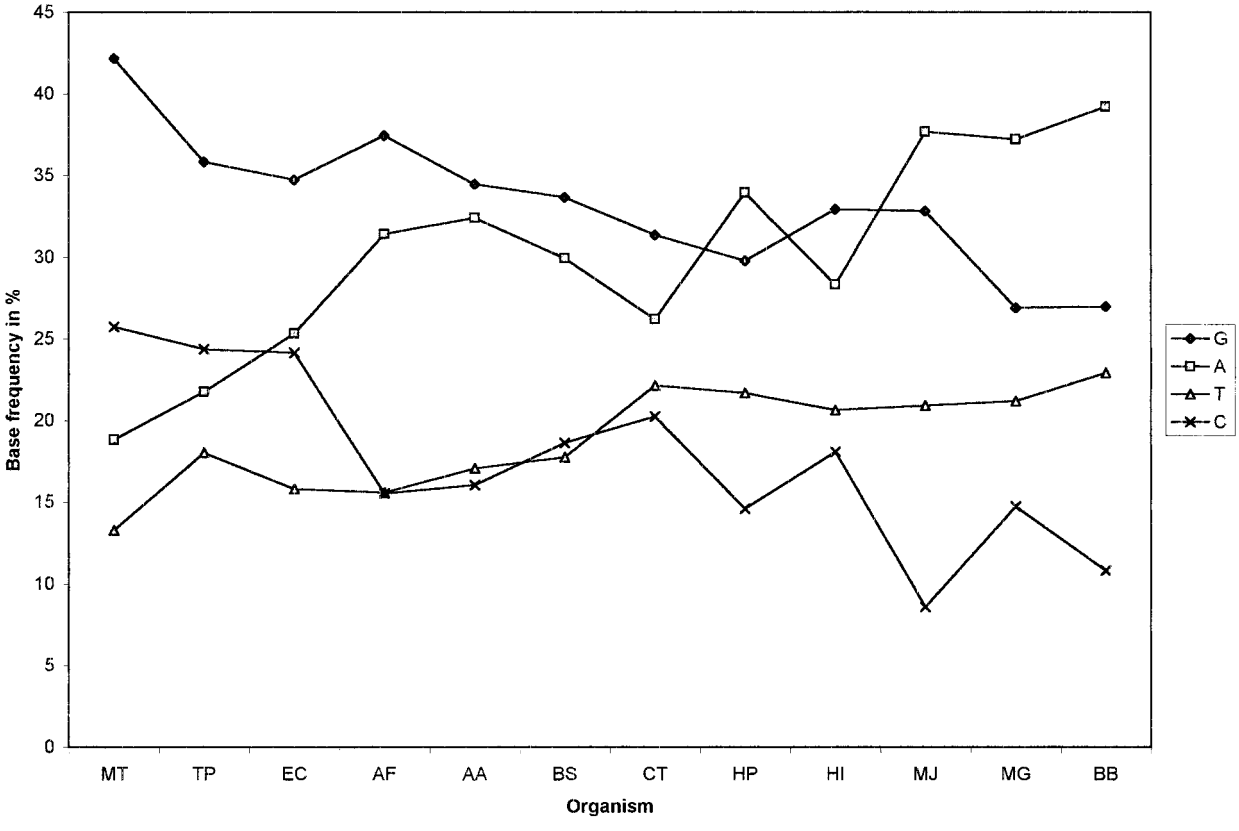

FIG. 2. Frequency distributions expressed in percentages of four individual bases (A,T,G,C) for each organism at the first codon position.

sition of the codon. From Fig. 3 where the frequencies of bases at the second positions of the codons are plotted it was observed that the frequencies of bases vary as A > T > C > G in most of the organisms. The frequency of G in the second codon position is always less than any other bases in all the organisms irrespective of their overall GC contents. The frequencies of bases at the third position shown in Fig. 4 demonstrate that there is no particular regular pattern of the occurrence of individual bases at this site. The more or less unique

TABLE 3

Dinucleotide Frequencies Expressed as Percentages in 12 Organisms

Dinucleotide	MT	TP	EC	AF	AA	BS	CT	HP	HI	MJ	MG	BB
AA	3.0699	5.7534	7.5326	8.2312	12.4699	10.8883	9.2915	13.404	12.3735	15.8346	15.792	16.6646
AT	3.6798	5.2687	6.4474	5.4972	5.2027	8.0343	7.6983	7.9505	9.0254	10.9346	8.9773	11.1078
AG	3.9871	5.9394	4.7282	8.8055	8.6187	6.2623	7.047	6.6452	5.1694	8.7461	6.2072	6.4068
AC	6.0695	5.271	5.4129	4.8586	6.2503	4.8152	4.4192	4.0884	4.7825	3.2097	5.3129	3.3407
GA	6.3225	6.293	6.3353	9.6716	9.0014	7.7654	7.1881	6.5086	5.8593	9.2509	5.8468	6.6486
GT	5.7054	6.7241	5.708	4.7299	4.6991	4.3914	4.7664	3.8175	5.3698	4.0417	5.0139	3.6569
GC	11.5774	8.5157	8.5998	6.1413	3.5646	6.3131	4.8875	6.1159	5.4466	2.7937	2.9952	3.0466
GG	10.0494	7.1409	6.7007	8.1129	7.1032	5.6348	4.8005	5.1892	4.408	4.974	2.9589	3.5402
TA	1.5991	4.1117	4.2499	3.8077	6.4164	4.8476	6.4775	6.6017	6.9004	9.3216	8.6562	9.5413
TT	3.0464	7.4314	6.4389	7.0535	7.8868	8.3288	10.2045	11.6339	10.9682	10.7587	12.8752	14.199
TG	6.7474	8.0756	8.1221	7.0238	4.5016	7.0518	6.3652	6.1317	7.3992	6.5904	6.7345	5.9537
TC	5.8994	5.5374	5.2306	5.3647	4.9357	5.4579	6.8551	3.9673	4.5226	2.8298	3.8646	3.8555
CA	5.8149	6.074	6.0032	5.6821	4.6539	6.4988	5.499	5.5741	6.2177	4.3179	5.9946	4.6655
CT	4.8605	5.7315	5.4472	5.9691	5.9519	4.9315	7.2329	4.9324	4.4271	3.7655	5.2639	4.5861
CG	12.8709	7.5185	7.7929	4.7135	4.1449	5.1558	3.4297	3.665	4.1074	0.7498	0.9141	0.991
CC	8.7003	4.6081	5.2502	4.3375	4.5988	3.6229	3.8376	3.773	3.0193	1.8811	2.5928	1.791

Note. The identification code for each organism is described in Table 1.

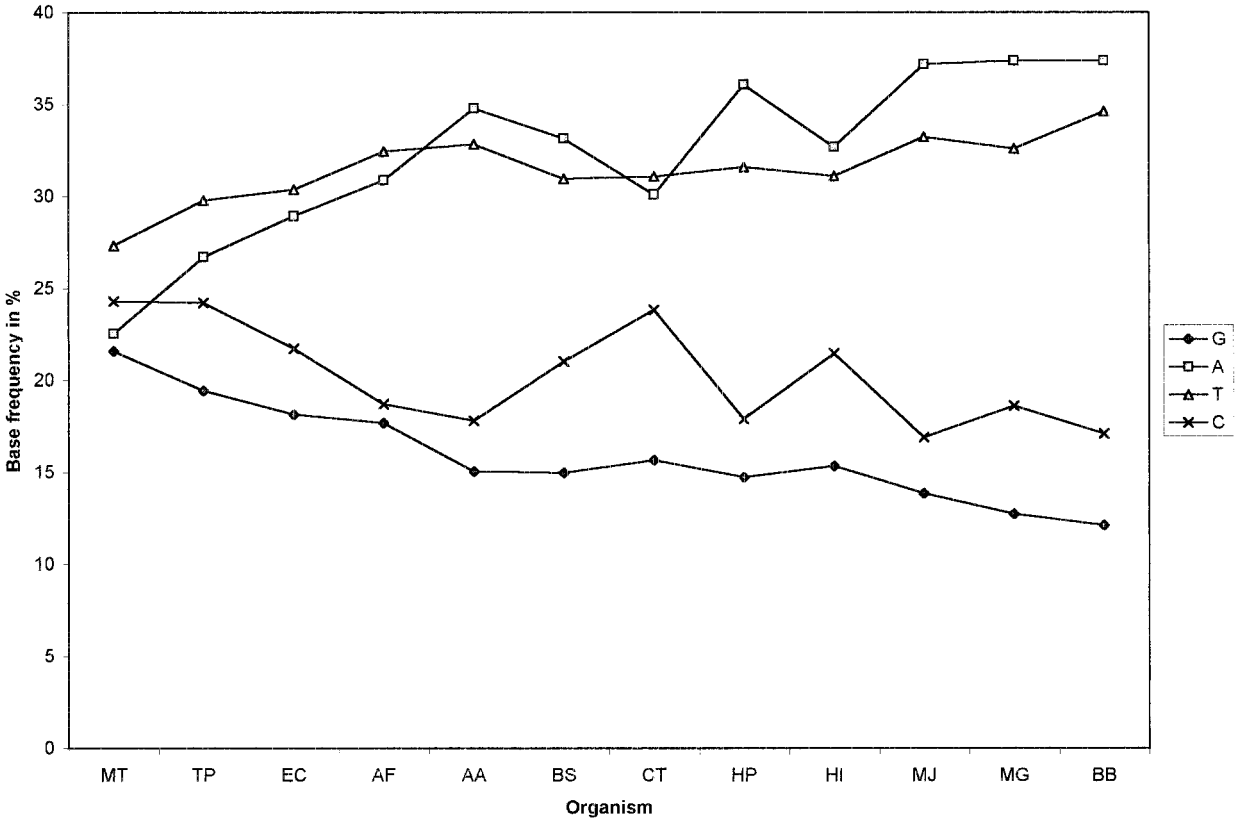

FIG. 3. Frequency distributions expressed as percentages of four individual bases (A,T,G,C) for each organism at the second codon position.

pattern of bases at the first two codon positions is expected, since majority of substitution at the first two codon positions results in non-synonymous substitution, resulting in the change of protein structure and consequently the folding pattern of proteins. It was also urgued by Zang and Chou (12) that for stable and native folding structure of a protein the occurrence of frequency of four bases at the first two codon positions is important. Calculations of standard deviation (s.d.) of the frequencies of the four bases at the three different positions of the codons shown in Table 2 show that the values of the standard deviation at the third positions of the codon is much higher than the first and second positions of the codon, and between the two first codon positions the second one has the lowest standard deviation, indicating that the frequencies of the four bases among the different organisms are not as homogeneous as the first two positions of the codons. From these results it can be said that the first and second positions of the codons 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 protein, whereas most nucleotide substitutions at the third positions are synonymous and is under weak purifying selection or in other words it can be said that first and second positions of the codons are the structure determining positions and the third position of the codon is the species determining position.

Dinucleotide Frequency Analysis

Dinucleotide frequencies are strongly related to diverse phenomena. It has been shown that the choice of low usage codons is strongly influenced by the dinucleotide biases in the various organisms (13). It was also observed that the level of gene expression, CpG methylation and conformational structure of DNA is dependent on dinucleotide frequencies of DNA (13–16). It was also reported that codon choices in eukaryotes are governed to a large extent by the dinucleotide frequencies (17). In order to determine whether there exist any unique relationship of the dinucleotide frequencies with the compositional constraints of the genome or not, we have calculated the dinucleotide frequencies of the 12 organisms. Table 3 shows the dinucleotide frequencies expressed as percentages for the 12 organisms. It is evident from Table 3 that GC-rich genomes are richer in GG, GC, CC and CG dinucleotides whereas in GC-poor genomes AA, AT, TA and TT dinucleotides are the most predominant. This is a clear

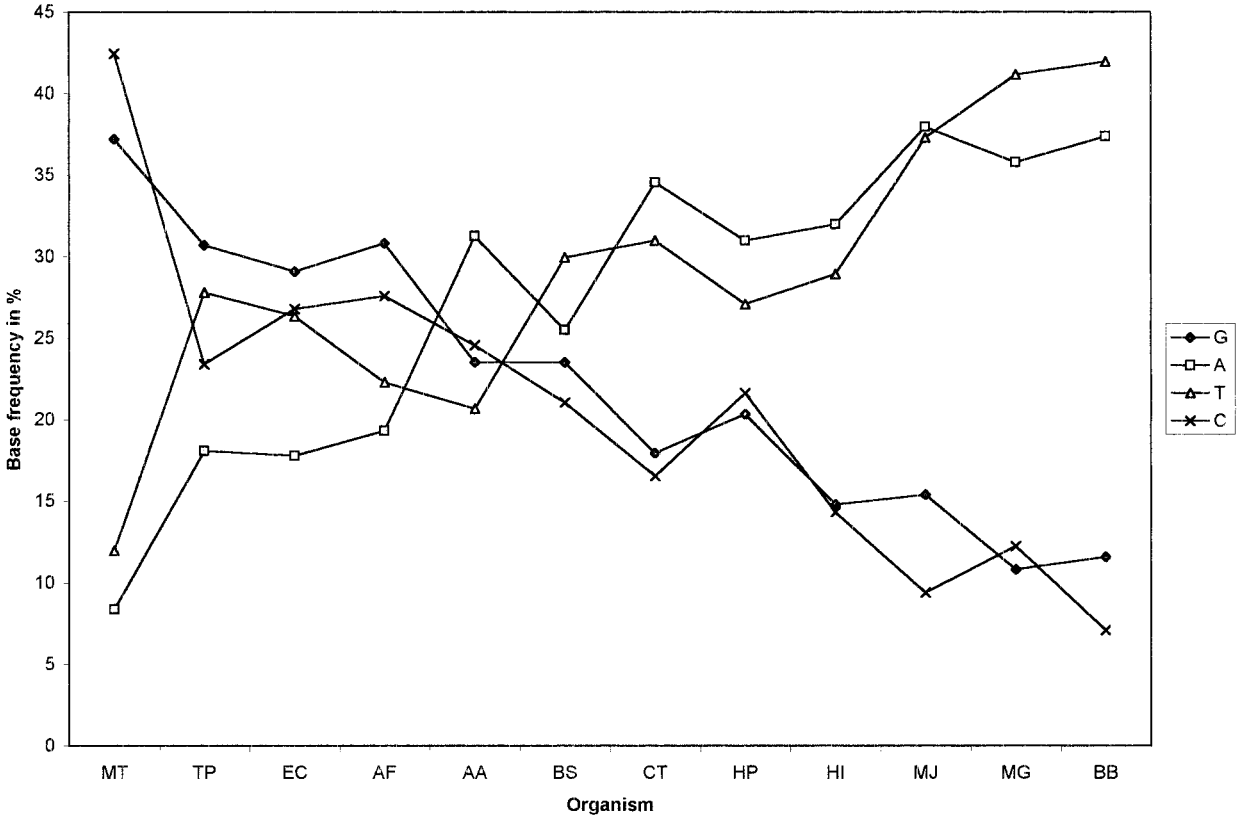

FIG. 4. Frequency distributions expressed as percentages of four individual bases (A,T,G,C) for each organism at the third codon position.

indication that the dinucleotide variability is also evident in compositionally compartmentalized genomes.

ACKNOWLEDGMENT

The authors are thankful to the Department of Biotechnology, Government of India, for funding the BTIS programme at Bose Institute, Calcutta.

REFERENCES

1. McInerney, J. O. (1998) *Bioinformatics* **14,** 372–373.
2. Lee, K. Y., Wahi, R., and Barbu, E. (1995) *Ann. Inst. Pasteur (Paris)* **91,** 212–214.
3. Sueoka, N. (1961) *J. Mol. Biol.* **3,** 31–40.
4. Bernardi, G., and Bernardi, G. (1985) *J. Mol. Evol.* **22,** 363–365.
5. Aissani, B., D Onofrio, G., Mouchiroud, D., Gardiner, K., Gautier, C., and Bernardi, G. (1991) *J. Mol. Evol.* **32,** 493–503.
6. D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C., and Bernardi, G. (1991) *J. Mol. Evol.* **32,** 504–510.
7. Bernardi, G., and Bernardi, G. (1991) *J. Mol. Evol.* **33,** 57–67.
8. Bernardi, G., and Bernardi, G. (1986) *J. Mol. Evol.* **24,** 1–11.
9. Muto, A. and Osawa, S. (1987) *J. Mol. Evol.* **84,** 166–169.
10. D'Onofrio, G., and Bernardi, G. (1992) *Gene* **110,** 81–88.
11. Tomb, J-F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKennery, K., Fitzegerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weidman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Borodovsky, M., Karp, P. D., Smith, H. O., Fraser, C. M., and Venter, J. C. (1997) *Science* **388,** 539–547.
12. Zhang, C. T., and Chou, K. C. (1994) *J. Mol. Evol.* **238,** 1–8.
13. Nussinov, R. (1984) *Nucleic Acids Res.* **12,** 1749–1763.
14. Hanai, R., and Wada, A. (1990) *J. Mol. Evol.* **30,** 109–115.
15. Bird, A. P. (1980) *Nucleic Acids Res.* **8,** 1499–1504.
16. Hunter, C. (1993) *J. Mol. Biol.* **230,** 1025–1054.
17. Nussinov, R. (1981) *J. Mol. Biol.* **149,** 125–131.