ABSTRACT

After a gap of more than seven decades of discovery of DNA by Friedrich Miescher in 1869, Chargaff and co-workers found that the double stranded DNA contains as many purine (A, G) as pyrimidine bases (C, T). Further they also observed that the complementary nucleotides have the same abundance values. This vital discovery latter helped explaining the DNA double-helix model in which A pairs only with T and G pairs only with C. Moreover Chargaff and coworkers also observed that the relative frequency of the four DNA bases was different for different organisms. We can verify from the available genome sequences in the DDBJ data base that the GC content (% of G and C nucleotide in a DNA) ranges approximately from 16.6 to 74.9 in bacterial genomes. Burge, Campbell and Karlin observed that the relative frequencies of di and tri-nucleotides characterize a genome. Species differ with respect to this genomic signature. Certain oligonucleotide sequences are reported to be preferred/avoided in a genomic sequence. In this project, we did a thorough investigation of di-nucleotide frequencies in high expression (HE) genes and whole genome in 2^{nd} and 3^{rd} codon positions as well as in intergenic regions for seventeen bacterial genomes covering a wide range of genomic GC. We have ignored codon position 1 in this study because it is not so degenerate like other two codon positions. Di-nucleotides at codon position 2 are a part of a codon whereas di-nucleotide at codon position 3 considers nucleotides from two neighboring codons. Therefore we have considered both of these positions for a better understanding of di-nucleotide avoidance or preference in this study. We observed that most of the di-nucleotide frequencies are as expected in the intergenic regions which are similar to a simulated sequence, whereas certain di-nucleotide usages are more constrained in high expression genes than whole genome. Constrained di-nucleotide frequency in high expression genes can be attributed to high codon usage bias in these genes.