Gene 536 (2014) 18–28

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Constraint on di-nucleotides by codon usage bias in bacterial genomes

Siddhartha Sankar Satapathy [a], Bhes Raj Powdel [c], Malay Dutta [a], Alak Kumar Buragohain [b,d], Suvendra Kumar Ray [b,*]

[a] Department of Computer Science and Engineering, Tezpur University, Tezpur, Assam 784 028, India
[b] Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur, Assam 784 028, India
[c] Department of Statistics, Darrang College, Tezpur, Assam 784001, India
[d] Dibrugarh University, Dibrugarh, Assam 786004, India

ARTICLE INFO

Article history:
Accepted 25 November 2013
Available online 11 December 2013

Keywords:
Relative di-nucleotide frequency
Codon usage bias
Gene expression
Inter-genic region
Codon context

ABSTRACT

It has been reported earlier that the relative di-nucleotide frequency (RDF) in different parts of a genome is similar while the frequency is variable among different genomes. So RDF is termed as genome signature in bacteria. It is not known if the constancy in RDF is governed by genome wide mutational bias or by selection. Here we did comparative analysis of RDF between the inter-genic and the coding sequences in seventeen bacterial genomes, whose gene expression data was available. The constraint on di-nucleotides was found to be higher in the coding sequences than that in the inter-genic regions and the constraint at the 2nd codon position was more than that in the 3rd position within a genome. Further analysis revealed that the constraint on di-nucleotides at the 2nd codon position is greater in the high expression genes (HEG) than that in the whole genomes as well as in the low expression genes (LEG). We analyzed RDF at the 2nd and the 3rd codon positions in simulated coding sequences that were computationally generated by keeping the codon usage bias (CUB) according to genome G+C composition and the sequence of amino acids unaltered. In the simulated coding sequences, the constraint observed was significantly low and no significant difference was observed between the HEG and the LEG in terms of di-nucleotide constraint. This indicated that the greater constraint on di-nucleotides in the HEG was due to the stronger selection on CUB in these genes in comparison to the LEG within a genome. Further, we did comparative analyses of the RDF in the HEG *rpoB* and *rpoC* of 199 bacteria, which revealed a common pattern of constraints on di-nucleotides at the 2nd codon position across these bacteria. To validate the role of CUB on di-nucleotide constraint, we analyzed RDF at the 2nd and the 3rd codon positions in simulated *rpoB*/*rpoC* sequences. The analysis revealed that selection on CUB is an important attribute for the constraint on di-nucleotides at these positions in bacterial genomes. We believe that this study has come with major findings of the role of CUB on di-nucleotide constraint in bacterial genomes.

1. Introduction

The four different nucleotides composing the DNA sequence result in sixteen different di-nucleotides. Unlike the mono-nucleotides, the di-nucleotides in a DNA double helix are constrained by the physio-chemical properties such as, helical orientation, base stacking (Yakovchuk et al., 2006), propeller twist, roll, and slide (Calladine et al., 2004). The differences in the properties like stacking in di-nucleotides with identical mono-nucleotide composition such as in TA with pyrimidine–purine sequence order in DNA and AT with purine–pyrimidine sequence order in DNA are attributed to these constraints (Calladine et al., 2004). Variability in the di-nucleotide frequencies in a genome, not proportional to the composition of mononucleotides is therefore expected.

Karlin and his group studied constraint on the di-nucleotides by finding out relative di-nucleotide frequency (RDF) (Karlin, 2001; Karlin et al., 1998). RDF is calculated by dividing the actual di-nucleotide frequencies with the expected di-nucleotide frequencies. The RDF value ≤ 0.78 indicates that the di-nucleotide is avoided while a value ≥ 1.23 indicates the preference for the di-nucleotide (Karlin et al., 1998). Study in bacterial genomes revealed that RDF value is similar in different parts of a chromosome while the value is variable among different bacteria. Therefore, this constancy of the RDFs within a genome was termed as "genome signature" (Burge et al., 1992; Karlin et al., 1998). Genome signatures may be indicative of the presence of anomalous gene clusters and pathogenicity islands in diverse bacterial genomes (Karlin, 2001).

The functional significance of any di-nucleotide per se is not known in bacteria. The di-nucleotide CG is known to regulate gene expression in vertebrate genomes. It is pertinent to point out that di-nucleotides are

Abbreviations: RDF(s), relative di-nucleotide frequency (ies); CUB, codon usage bias; HEG, high expression genes; LEG, low expression genes.

* Corresponding author. Tel.: +91 3712 275406; fax: +91 3712 267005x006.
E-mail address: suven@tezu.ernet.in (S.K. Ray).

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.11.098

embedded in the functionally well-defined tri-nucleotide order (codon) in DNA sequences. However, the significance of di-nucleotides in genomes can be appreciated in the observation of TA avoidance in bacterial genomes (Karlin et al., 1998). In addition, the RDF is an important feature in genomes.

Di-nucleotides in genome can be a target of context dependent mutation. The best example of this is witnessed in the formation of pyrimidine dimer in the presence of ultraviolet light. However, pyrimidine dimer is not avoided in bacteria growing in light (Lobry, 1995) suggesting that the various DNA repair systems are sufficient to take care of this lesion. An independent comparative study by Ochman (2003) revealed that cytosine deamination is of greater occurrence than pyrimidine dimer in chromosomes. The avoidance of CG in mammalian genomes has been thought to be a result of deamination of methylated-C. There are different studies suggesting the context dependent mutation in *Escherichia coli* and mitochondria (Bulmer, 1990; Jia and Higgs, 2008). In a study involving distantly related organisms such as *E. coli*, *Saccharomyces*, and *Drosophila*, Antezana and Kreitman (1999) put forward the hypothesis that the tri and di-nucleotide motif preference in coding regions can explain the structuring of codon preferences. But in none of the above studies comparison has been made between coding and non-coding regions for any conclusive inference.

Genome is made up of both coding as well as non-coding sequences: in prokaryotes the proportion of coding sequence is larger than the non-coding sequences. It is known that coding sequences are relatively under greater selection pressure than the non-coding sequences (Bulmer, 1991; Hershberg and Petrov, 2009). To have a better insight towards the functional significance of di-nucleotides, we decided to study RDF in coding and non-coding regions in bacterial genomes. Analysis of the RDF in non-coding as well as in coding regions might help in understanding the contribution of di-nucleotide constraints towards codon usage bias (CUB) in a genome (Satapathy et al., 2012). It would also be helpful to understand the contribution of translational selection (Gouy and Gautier, 1982; Sharp and Li, 1986a, 1986b), codon context (Fedorov et al., 2002), and CUB (Ermolaeva, 2001) towards di-nucleotide constraints in bacterial genomes.

In this study we observed that the inter-genic sequences and the coding sequences within a genome are remarkably different with respect to di-nucleotide constraints within a genome. Further analysis of di-nucleotides at the 2nd and 3rd codon positions in coding sequences as well as in simulated coding sequences suggested that the CUB is a major attribute towards di-nucleotide constraints in genomes.

2. Materials and methods

2.1. Nucleotide sequences of the inter-genic and the coding sequences of bacterial genomes

Gene sequences of different bacterial genomes were downloaded from DDBJ site (www.gib.genes.nig.ac.jp). Intergenic regions were downloaded from Comprehensive Microbial Resource (CMR) website (http://cmr.jcvi.org/cgi-bin/CMR/CmrHomePage.cgi). Intergenic regions were analyzed in two different ways. In one, the entire intergenic regions were analyzed. In the other, only intergenic regions more than or equal to 200 nucleotides sizes were analyzed ignoring 50 nucleotides each from the 5′ end and the 3′ ends so as to focus only on the sequences that are without any possible role in gene expression regulation.

To analyze di-nucleotide abundance in the high expression genes (HEG) and in all the genes in the genome, we extracted information of gene expression data. Transcriptomic data were downloaded from NCBI GEO website (ftp://ftp.ncbi.nih.gov/pub/geo/). For *E. coli*, we used proteome data produced by Ishihama et al. (2008). In total seventeen bacterial genomes belonging to seven bacterial groups covering a genomic G+C content from 32.84% to 72.00% and genome size from 1.83 mbp to 9.11 mbp (Table 1) were analyzed.

Table 1
List of the seventeen bacteria analyzed in this study.

Sl. No	Bacteria	Group	Genome size (bp)	G+C%
1	*Bradyrhizobium japonicum*	α proteobacteria	9,105,828	64.06
2	*Bifidobacterium longum*	Actinobacteria	2,260,266	60.13
3	*Bacillus subtilis*	Firmicutes	4,214,630	43.52
4	*Desulfovibrio vulgaris Hildenborough*	δ proteobacteria	3,773,159	63.28
5	*Escherichia coli*	γ proteobacteria	4,639,675	50.00
6	*Haemophilus influenzae*	γ proteobacteria	1,830,069	38.15
7	*Listeria monocytogenes*	Firmicutes	2,944,528	37.98
8	*Lactobacillus plantarum*	Firmicutes	3,348,625	44.42
9	*Nitrosomonas europaea*	β proteobacteria	2,812,094	50.72
10	*Pseudomonas aeruginosa*	γ proteobacteria	6,264,404	66.56
11	*Pseudomonas syringae*	γ proteobacteria	6,538,260	58.34
12	*Rhodopseudomonas palustris*	α proteobacteria	5,467,640	65.03
13	*Rhodobacter sphaeroides*	α proteobacteria	4,603,060	68.79
14	*Staphylococcus aureus*	Firmicutes	2,903,636	32.84
15	*Streptomyces coelicolor*	Actinobacteria	9,054,847	72.00
16	*Streptococcus mutans*	Firmicutes	2,030,921	36.83
17	*Thermus thermophilus*	Deinococcus-Thermus	2,116,056	69.50

Hypothetical genomic sequences were generated using Hidden Markov Model (HMM) based software GenRGenS (Ponty et al., 2006).

2.2. Calculation of RDFs at the 2nd as well as at the 3rd codon positions

Di-nucleotide frequencies were calculated as described in Karlin et al. (1998). Di-nucleotide frequencies at the 1st, 2nd and the 3rd positions of codons in the seventeen bacterial genomes were calculated. The relative frequencies of the di-nucleotides were calculated as follows: f_X, f_Y and f_Z denote frequencies of the nucleotides X, Y and Z at the 1st, 2nd and the 3rd codon positions respectively, where X, Y and Z may be any of the four nucleotides A, C, G and T. The actual di-nucleotide frequencies embedded within codons can be denoted as f_{XY}, f_{YZ}, f_{ZX}. In f_{ZX}, Z is the nucleotide at the 3rd position of a codon and X is the nucleotide at the 1st position of the next codon towards the 5′ end. We will have sixteen such di-nucleotides at each position of codons studied. The di-nucleotides at 1st position of codons were not analyzed here because this position is less degenerate and the di-nucleotide frequencies at this position will be a mere reflection of amino acid composition. So considering the nucleotide frequencies f_X, f_Y and f_Z, the expected di-nucleotide frequencies at the 2nd and the 3rd positions of the codons are as follows:

At the 2nd codon position: $Ef_{YZ} = f_Y.f_Z$

At the 3rd codon position: $Ef_{ZX} = f_Z.f_X$.

The RDFs at the 2nd and the 3rd codon can be calculated by dividing the observed di-nucleotide frequency by the expected di-nucleotide frequencies as follows.

At the 2nd codon position: $Rf_{YZ} = f_{YZ}/Ef_{YZ}$

At the 3rd codon position: $Rf_{ZX} = f_{ZX}/Ef_{ZX}$.

Using a computer program written in C language, we have calculated these RDFs in three sets of genes: (i). the top 100 HEG, (ii) the bottom 100 LEG and (iii) all the genes in the genome.

2.3. Calculation of RDFs in intergenic region

Di-nucleotide frequencies in inter-genic regions were calculated in the same way as described in the genic regions but codon specificity was not used. With the help of a computer program written in C language, we found out the frequencies of different nucleotides in all the intergenic regions and then used these values to calculate the expected di-nucleotide frequencies. Further with the help of the computer