Analysis of in-frame dinucleotides, encoded amino acids and synonymous codon choice in bacterial genomes reveals a common pattern

Amit Ranjan1,2, Akhauri Yash Sinha1,3, Badrinath Dubey1, Ankur Saikia1,4, Alak Kumar Buragohain1 and Suvendra Kumar Ray1,*

1Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur 784 028, India
2Present address: Advanced Centre for Treatment, Research and Education in Cancer, TMC, Kharghar, Navi Mumbai 410 210, India
3Present address: Department of Microbiology and Cell Biology, Indian Institute of Science, Bangalore 560 012, India
4Present address: Biology Group, Department of Drug Discovery, Jubilant Biosys Ltd, 450, 4th D Main, 12th Cross, Mahalakshmi Puram, Bangalore 560 086, India

A comparative analysis of in-frame dinucleotide abundance, amino acid abundance and codon usage pattern among bacterial genomes has been done in this study. Dinucleotide abundance studies revealed that the abundance value of a dinucleotide at the first position of codons is less variable than its abundance at the second or the third position in different bacterial genomes. This is correlated with the similarities observed in the relative abundance patterns of different amino acids in different bacterial proteomes. A similarity in preference for the occurrence of a keto nucleotide (G/T) at the third position for leucine family box codons and a pyrimidine nucleotide (C/T) at the third position for arginine family box codons was observed in different bacterial genomes. This indicates that apart from genome G + C content and tRNA abundance values, the first two nucleotides of a codon might influence the occurrence of a nucleotide at the third position of codons.

Keywords: Amino acid composition, bacterial genomes, codon usage pattern, dinucleotide composition.

Dinucleotide composition frequencies in a bacterial genome can be considered as genome signatures as the frequency values of a dinucleotide for different regions of a genome remain similar1–4. The reason for this similarity in genomes in terms of dinucleotide abundance frequency is not fully understood. Since bacterial genomes are mostly composed of coding sequences2, the dinucleotide signatures in the genome most probably have had their origins from the differential codon occurrence frequencies necessitated by the amino acid composition in the encoded proteomes. In-frame analysis of dinucleotide abundance in all open reading frames (ORFs) in genomes will enable determination of dinucleotide abundance at the first, second and third positions of codons. Dinucleotide abundance values at the first position will give a correlation among the abundance of different amino acids in a genome. In fact, the abundance of the dinucleotides GT, CC, AC, GC and GG at the first position has a direct correlation with the abundance of the amino acids valine, proline, threonine, alanine and glycine respectively, that represents corresponding amino acids (being encoded by synonymous codons with these dinucleotides at the first position) in a genome. Due to degeneracy among codons, study of dinucleotide abundance at the second and third positions (which involves the third nucleotides of codons) will provide insights into the low and high abundance value of a particular dinucleotide.

Synonymous codons in a family box of genetic code have the same nucleotides at the first and second positions. Since any of the four nucleotides can occur at the third nucleotide position in these codons, analysis of synonymous codon usage pattern in different family boxes of genetic code will reveal if there is any preference for the occurrence of a particular nucleotide in a family box6. The preferred nucleotide can be correlated with the genome G + C% and the tRNA occurrence, which are considered as the evolutionary forces acting on the codon usage pattern of an organism7–9. Moreover, comparative analysis of codon usage pattern across the genomes might reveal the presence of any other common factor playing an important role in determining codon usage in these organisms. In this study we have determined dinucleotide abundance frequency at the first, second and third positions in codons of all OREs in 14 different bacteria; amino acid abundance frequency in the whole proteomes of 13 different bacteria; synonymous codon usage patterns in genomes of 15 different bacteria and have done comprehensive analysis of these in-frame bacterial genomic and proteomic data to understand the evolutionary pattern in bacterial genomes.

*For correspondence. (e-mail: suven@tezu.ernet.in)
Table 1. Organisms analysed in this study

<table>
<thead>
<tr>
<th>Serial no.</th>
<th>Organism (strain)</th>
<th>Genome G + C%</th>
<th>Genome/ chromosome size (nt)</th>
<th>Proteome size (aa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Agrobacterium tumefaciens C58 (U. Washington)</td>
<td>59</td>
<td>2841490</td>
<td>829342 (87.54)*</td>
</tr>
<tr>
<td>2</td>
<td>Bacillus subtilis 168</td>
<td>43</td>
<td>4214630</td>
<td>NA</td>
</tr>
<tr>
<td>3</td>
<td>Clostridium acetobutylicum ATCC824</td>
<td>30</td>
<td>3940880</td>
<td>1123604 (85.53)</td>
</tr>
<tr>
<td>4</td>
<td>Escherichia coli K12 MG1655</td>
<td>50</td>
<td>4639675</td>
<td>1355794 (87.66)</td>
</tr>
<tr>
<td>5</td>
<td>Erwinia caratovora subsp. atroseptica SCRI1043</td>
<td>50</td>
<td>5064019</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>Helicobacter pylori 26695</td>
<td>38</td>
<td>1667867</td>
<td>498269 (89.62)</td>
</tr>
<tr>
<td>7</td>
<td>Haemophilus influenzae Rd KW20</td>
<td>38</td>
<td>1830138</td>
<td>52827 (85.70)</td>
</tr>
<tr>
<td>8</td>
<td>Lactobacillus acidophilus NCFM</td>
<td>30</td>
<td>1995364</td>
<td>572815 (86.19)</td>
</tr>
<tr>
<td>9</td>
<td>Listeria monocytogenes 4b F2365</td>
<td>38</td>
<td>2905390</td>
<td>NA</td>
</tr>
<tr>
<td>10</td>
<td>Mycobacterium leprae TN</td>
<td>57</td>
<td>3268203</td>
<td>587773 (49.46)</td>
</tr>
<tr>
<td>11</td>
<td>Neisseria meningitidis Z2491</td>
<td>51</td>
<td>2184406</td>
<td>587023 (80.62)</td>
</tr>
<tr>
<td>12</td>
<td>Nitrosomonas europaea ATCC 19718</td>
<td>50</td>
<td>2812094</td>
<td>799993 (85.34)</td>
</tr>
<tr>
<td>13</td>
<td>Pseudomonas syringae pv. phaseolicola</td>
<td>58</td>
<td>5928787</td>
<td>1811390 (91.65)</td>
</tr>
<tr>
<td>14</td>
<td>Pseudomonas aeruginosa PA01 1448A</td>
<td>66</td>
<td>6264404</td>
<td>1859565 (89.05)</td>
</tr>
<tr>
<td>15</td>
<td>Ralstonia solanacearum GM1000</td>
<td>67.9</td>
<td>3716413</td>
<td>1096514 (88.02)</td>
</tr>
<tr>
<td>16</td>
<td>Xanthomonas campestris pv. campestris ATCC 33913</td>
<td>65</td>
<td>5076188</td>
<td>1436776 (84.91)</td>
</tr>
<tr>
<td>17</td>
<td>Xylella fastidiosa 9a5c</td>
<td>52.61</td>
<td>2679006</td>
<td>742365 (82.88)</td>
</tr>
<tr>
<td>18</td>
<td>Yersinia pestis KIM</td>
<td>47</td>
<td>4600755</td>
<td>NA</td>
</tr>
</tbody>
</table>

*Numbers in parentheses indicate percentage of chromosome/genome encoding the proteome. Organisms having genome G + C% above 55 and below 45 are considered as higher genome G + C content (shown in bold) and lower genome G + C content (shown in italics) respectively, in this study.

NA, Not analysed; Serial nos 7, 14, 15 and 17 have not been analysed for dinucleotide abundance; Serial nos 2, 5, 6, 9, and 18 have not been analysed for amino acid abundance, and serial nos 7, 14 and 17 have not been analysed for codon usage pattern.

Materials and methods

Only the potential protein coding regions of genome sequences [DNA sequences of CDS, excluding the introns (FASTA format)] and proteome sequences [amino acid sequences of CDS, (FASTA format)] of all the bacteria analysed in this study were downloaded from the DNA Data Bank of Japan (DDBJ), Genome Information Broker site (http://gib.genomes.nig.ac.jp). Gene names written before the potential ORF were replaced with ‘XXX’ and similarly, names of poly-peptides in a proteome were replaced with ‘X’. Dinucleotide analysis of all the ORFs of an organism was performed using the ‘Dinssearch’ program[10]. This program finds out the abundance values of all the 16 dinucleotides at the first, second and third positions of codons of a given DNA sequence. Frequency of dinucleotides was calculated by dividing the abundance value of a dinucleotide at the first nucleotide position by the total abundance of dinucleotides at the first position of codons. Similar method was used for dinucleotides at the second and third positions of codons. Amino acid abundance values in a proteome were calculated using the computer program ‘Seqsearch’[11]. This program finds out the abundance value of a sequence in a given proteome or genome. Amino acid frequency in a proteome was calculated by dividing the amino acid abundance value by the total proteome size. Codon abundance was calculated using the computer program ‘Trisearch’ (developed for this study in our department). This program finds out the abundance of different trinucleotides occurring consecutively in a given DNA sequence. So if a DNA sequence starts with a reading frame, then this program gives information about the abundance value of different codons. Comparative analysis of the abundance values for a given dinucleotide or trinucleotide or amino acid was done by comparing the coefficient of variation of its abundance frequency, which is calculated by dividing the standard deviation value with the average value (percentage value was not taken for convenience in calculation).

Results

Dinucleotide abundance frequency

Abundance values of dinucleotides at the first, second and third positions of codons in all ORFs (protein coding sequences) of 13 bacterial genomes having a wide range of genome G + C content (Table 1) are presented in Figure 1a–c. In Figure 1a, dinucleotides such as AA, AT, AG, TA, TC (mostly AT-rich) are more abundant in organisms with lesser G + C content (Bacillus subtilis, Clostridium acetobutylicum, Helicobacter pylori, Listeria monocytogenes, Lactobacillus acidophilus, Yersinia pestis) and less abundant in organisms having higher genomic G + C content (Agrobacterium tumefaciens, Mycobacterium leprae, Pseudomonas syringae, Xanthomonas campestris). An inverse correlation was observed in the case of GG,
Figure 1.  

(a-c) Dinucleotide abundance frequencies at the first, second and third positions of codon in all open reading frames of different bacterial genomes. Dinucleotide abundance frequencies at the first position of codons suggest that certain dinucleotides such as AA, AT, AG, TA and TC are more abundant in genomes having lower G+C content and dinucleotides such as GG, GC, CT and CG and CC are more abundant in genomes having higher G+C content. The TA dinucleotide is more abundant at the second and third positions in organisms having lower G+C content. The common observation in (b) and (c) is that codons having G+C at the third nucleotide position are more abundant in organisms having higher G+C genome content. At, Agrobacterium tumefaciens; Bs, Bacillus subtilis; Ca, Clostridium acetobutylicum; Ec, Erwinia caratovora; El, Escherichia coli; Hp, Helicobacter pylori; La, Lactobacillus acidophilus; Lm, Listeria monocytogenes; Mi, Mycobacterium leprae; Nm, Neisseria meningitidis; Ne, Nitrosomonas europaea; Ps, Pseudomonas syringae; Xc, Xanthomonas campestris and Yp, Yersinia pestis.  

(d) Coefficient of variation of dinucleotide abundance at the three positions of codons in bacterial genomes. The coefficient of variation (standard deviation divided by average values) of 16 dinucleotide abundance values at three positions of codons in 14 bacterial genomes is presented. The std/avg values at the first position of codons for all dinucleotides are lesser in comparison to those for the second or third positions, except for CT, which is higher than the third position and AG which is higher than the second position. This suggests dinucleotide abundance value is less diverse at the first position than at the second or the third positions of codons in bacterial genomes.
GC, CT, CG and CC (mostly GC-rich) dinucleotides. This suggests that amino acids encoded by AT-rich codons are more abundant in organisms whose genome G + C content is less and that amino acids encoded by GC-rich codons are more abundant in organisms whose genome G + C content is higher. The coefficient of variation values suggest that dinucleotides such as AC, TC, GA, GT and GG are less diverse than other dinucleotides among the bacterial genomes at the first position of codons (Figure 1d). The common observation in Figure 1b and c is that codons having G/C at the third nucleotide position are more abundant in organisms having higher genome G + C content. GA dinucleotide abundance was found to be low at the second position of codons in all genomes. This is attributed to the lower frequency of occurrence of the TGA codon, which is a nonsense codon. Contrary to our observation, the TA dinucleotide has earlier been reported to have a low abundance in bacteria. Though the low abundance of TA dinucleotides is restricted to bacteria having greater than or equal to 50% G + C content in the genome as in C. acetobutylicum, L. monocytogenes, L. acidophilum and H. pylori where genomes are A + T-rich, these bacteria have a higher abundance of TA dinucleotides in both the second and third positions in the codons (Figure 1b–c).

A comparative analysis of the Figure 1a–c indicated that the dinucleotide abundance frequency at the first position follows a more uniform pattern than those of the third and second positions. To find out the extent of variation of abundance frequency of a dinucleotide at a given position among different genomes, the coefficient of variation values for the dinucleotides at different positions of codons were compared (Figure 1d). It is evident from Figure 1d that the values for the first nucleotide position are lower than those of the second and third positions in 15 out of 16 cases, and the values at the third nucleotide position are lower than those of second position in 10 out of 16 cases. More uniformity at the first position might be due to a correlation of amino acid abundance frequency among different bacteria. Greater variation in the pattern at the second and third positions is likely due to the involvement of the third nucleotide of codons and it is known that the occurrence of the nucleotides at this (third) position is correlated with genome G + C content. In fact, the analysis of G + C-rich and A + T-rich bacterial genomes separately for the dinucleotide abundance in the second and third positions of codons suggested that the values are less diverse for organisms having similar G + C% in the genome (data not presented). A related observation by analysing the ORFs of several bacterial genomes has been reported recently. However, Figure 1b looks more crisscross in relation to Figure 1c though both involve the third nucleotide of codons. This observed difference is mainly because the order of the dinucleotides plotted is same in both the cases.

Relative abundance frequency of amino acids is similar among the proteomes

Abundance frequency values for 20 amino acids in 13 (seven of which are part of the dinucleotide analysis) bacterial proteomes are presented in Figure 2a. It is evident from Figure 2a that the relative abundance of different amino acids follows a similar pattern in the encoded proteomes of different bacteria. Amino acids like C, H, and W are low in abundance, and L and A are the more abundant amino acids in all proteomes. Expectedly, amino acids F, I, K, N and Y, which are encoded by AT-rich codons are found more in organisms with lower G + C content. Similarly, A, G, P and R which are encoded by GC-rich codons are found to be more abundant in organisms having higher genomic G + C content. From Figure 2b it is evident that abundance frequencies of amino acids such as D, E, G, S, T and V are less diverse among the bacterial proteomes, which correlate with the abundance studies of dinucleotides at the first position of codons. It should be pointed out that among the 13 proteomes analysed here, only seven are from the genomes that were analysed for the dinucleotides. Considering the fact that it is the extent of the G + C content of the organisms on the basis of which the dinucleotide and the proteome patterns have emerged, it is reasonable to expect that the pattern might have more widespread occurrence among other genomes. From Figure 2a it is evident that the amino acids encoded by four and six codons are more abundant in the proteomes than amino acids encoded by one or two codons. This suggests that there is a positive correlation between the redundancy of the codons and their abundance in the genome.

Similarities in the preference for a nucleotide at the third position of synonymous codons

It is known that the abundance values of tRNA molecules in the cytosol are correlated with the codon usage pattern of the organism. For analysing codon usage pattern, we have considered codons of the eight family boxes of genetic code. This is because, presence of any nucleotide (A, T, G, C) at the third position of codons in a family box will code for the same amino acid. Thus by analysing the family box codons it can be found out whether there is any preferential occurrence of a particular nucleotide over other nucleotides at the third position. The relative occurrence of synonymous codons in the family boxes of the genetic code for 15 organisms was analysed. The relative abundance values of synonymous codons in family boxes for different amino acids are presented in Table 2. G and/or C was commonly observed to be more abundant at the third position of codons in genomes having higher G + C content and less abundant in genomes having lower G + C content. This is in conformity with the already known observations. However, the preponderance of G over other
nucleotides in the proline family box codons of \textit{B. subtilis} (with a lower G + C genome content) and the preponderance of C and G over other nucleotides in the arginine and valine family box codons respectively, in \textit{H. pylori} (which has a significantly lower G + C content) genomes, are interesting departures from our general observations in other bacterial genomes. Also, in both \textit{H. pylori} and \textit{B. subtilis}, G and/or C appear to be preferred to A/T in the glycine family box codons. In addition, in \textit{H. pylori} G + C have similar preference as A + T have in the threonine codon.

In the case of the leucine codons, a keto nucleotide (G/T) was found to be more abundant at the third position over an amino nucleotide (A/C). The reason for the preferential occurrence of a keto nucleotide over an amino nucleotide is not known. Similarly, in the arginine codons, a pyrimidine nucleotide (C/T) was found to be more abundant over a purine nucleotide (G/A) in all the 15 genomes and T is preferred to A, and C preferred to G in all organisms. There are several other observations in this study which suggest the non-randomness of nucleotide occurrence at the third position of codons. For example,
in the case of the alanine family box codons, C + G is maximally preferred in ten organisms and in these genomes, G + C% is either 50 or more. However, it is interesting to note that except Escherichia coli and Erwinia carotovora, all organisms have C maximally preferred at the third nucleotide position. Amongst the codons for glycine, C is maximally preferred in all genomes having genome GC% either 50 or more. In the serine codons, G + C is maximally preferred for five organisms having higher G + C%, of which four organisms (Ralstonia solanacearum, X. campesiris, P. syringae and M. lepae) maximally preferred G and only in one case (A. tumefaciens) G and C have similar preferences. In the threonine codons, C + G is maximally preferred where genome G + C% is either 50 or more, which also includes Y. pestis whose genome G + C% is 47. Nine out of the ten organisms have C maximally preferred, except E. carotovora where there is a similar preference for G and C. In the proline codons, G is maximally preferred in organisms having genome GC% either 50 or more along with B. subtilis, whose genome is AT-rich (43%). In the valine codons, T is preferred to A in all 14 genomes. In addition to the above, a common observation in all the family box codons in H. pylori is that there is a general preference for T to A.

Discussions

In-frame dinucleotide analysis of bacterial genomes has revealed that genome G + C content and amino acid abundance are the prime factors that determine the dinucleotide abundance frequency in the genomes. A low abundance of TA dinucleotide was observed for several prokaryotic and eukaryotic genomes, suggesting a universal avoidance of TA dinucleotides in genomes. However, in this study a high abundance of TA dinucleotide was observed at the second and third positions of codons in genomes having a low G + C content, suggesting that TA
dinucleotide is not avoided in these genomes. Lesser diversity in the dinucleotide abundance values at the first position of codons across the genomes is correlated with the observed similarity in the abundance pattern of amino acids in these genomes. In fact, an observation of regular pattern of the dinucleotide abundance at the third position also supports the similarity in the amino acid abundance pattern in different proteomes. Amino acid abundance and its relation with the number of synonymous codons in the genetic code have already been reported.\(^\text{8,10}\) The reason behind the abundance value of amino acids and the number of synonymous codons is not known. It has been speculated that by random mutation, a change of a codon to any other codon would lead to accumulation of more leucine, serine and arginine codons than tryptophan or methionine codons.\(^\text{11}\) The altered codon is fixed in the gene if it does not confer any deleterious effect. Occurrences of leucine and alanine in greater abundance might have come about due to this effect. If this is true, the amino acid abundance pattern is expected to be similar in all organisms as the genetic code is almost universal.

Genome G + C content and tRNA abundance in the cytosol are considered as the major determinants for codon usage in an organism.\(^\text{7,8,12}\) However, identification of altered G + C region and codon usage region within genomes that are not acquired recently through horizontal gene transfer has indicated that mutation and selective forces also play an important role in determining codon usage in an organism.\(^\text{14-18}\) Our observation relating to the preference of a keto nucleotide at the third nucleotide position in leucine codons of different bacterial genomes suggests the probable influence of the first and second nucleotides in a codon in determining the occurrence of a nucleotide at the third position of the codon. Preference for keto bases over amino bases has been reported for genes that are present in the leading DNA strands.\(^\text{19}\) This possibility is ruled out in our study because leucine codons have been analysed for all ORFs in organisms which are present in the leading as well as in the lagging strands. To find out if the keto bases at the third position give a translational advantage to the codons, we analysed the abundance of the leucine family box triplets in both the non-reading frames. It was observed that the keto to amino ratio at the third position for these triplets is higher than unity, which suggests that there is a general preference for a keto base after the dinucleotide CT occurring in the genomes (data not shown). However, a translational advantage cannot be ruled out in this case as the keto to amino ratio values in the reading frames were found to be higher than those in both the non-reading frames. In the present study, the arginine family box codons were observed to have a high preference for pyrimidine than purine nucleotide at the third position of codons. Since all the 15 bacterial genomes analysed exhibited this preference, it can be strongly argued in favour of the preference of a pyrimidine nucleotide at the third nucleotide position of the arginine family box codons. To find out if the presence of a pyrimidine nucleotide at the third position gives any translational advantage, we analysed the triplets of the arginine family box in both the non-reading frames of the genomes. The ratio of pyrimidine nucleotides to purine nucleotides in almost all cases has been found to be less than unity, which suggests a low preference of a pyrimidine nucleotide after the dinucleotide CG in genomes (data not shown). Thus it can be inferred that the preferential occurrence of a pyrimidine nucleotide at the third position of the arginine family box codons could be an advantage to the translational process. By whole genome analysis, it has been found that the first two nucleotides of a codon influence the occurrence a nucleotide at the third position. In fact, a similar observation was reported earlier\(^\text{20}\) in a different context, which suggests that anti-codon-codon pairing might follow a standard energy rule for efficient translation. According to this rule binding energy should not be too high or too low for the translational efficiency. For example in E. coli genes, NNC codon is preferred to NNT codon, where N is A and/or T. Similarly, NNT was found to be preferred to NNC, where N is G and/or C. Our observations of the preference of keto and pyrimidine nucleotides in leucine and arginine family boxes respectively, cannot be explained by this rule. Further studies on this aspect will reveal the mechanism of influence of the first two nucleotides on the occurrence of a nucleotide at the third position and also highlight the evolutionary significance of such occurrence.

RESEARCH ARTICLES


ACKNOWLEDGEMENTS. We thank Prof. Bolin K. Koowar, Head, Department of Molecular Biology and Biotechnology, Tezpur University for support and Dr. A. Prakash for comments on the manuscript. We also thank Sidhartha S. Satapathy for help with the computer programing, and Debojyoti Das, and Sujit K. Verma for useful discussions.

Received 30 August 2006; revised accepted 9 October 2007

MEETINGS/SYMPOSIA/SEMINARS

National Workshop on Application of Radioisotopes in Biology (ARB-2008)

Date: 7–10 March 2008
Place: Mangalore

Structure and topics to be covered (more stress on practical classes and hands-on experience):

Theory: (i) Application of radioisotopes and radiotracers in biology, (ii) Radioimmunoassay (RIA) and its applications, (iii) Storage, and stability of radiolabelled compounds, (iv) Safe handling of radioisotopes and labelled compounds and (v) Activities of Board of Radiation and Isotope Technology (BRIT), Mumbai in the area of radioisotope and radiation technology application.

Practicals: (i) Experiments on labelling of DNA/RNA, (ii) Detection, measurement and estimating radioactivity and (iii) Radio-Immunoassay (RIA) experiments.

Contact: Dr H. M. Somashekarappa
Convener, ARB-2008
University Science Instrumentation Centre
Mangalore University
Mangalagangothri 574 199
Tel: 0824 2287671
Fax: 0824 2287367
E-mail: maarrb2008@yahoo.com
Web: www.mangalareuniversity.ac.in

Sixth All India Summer Research Training Programme on Molecular Techniques

Date: 16–30 May 2008
Place: Tiruchengode

Summer school on: Genomics, Proteomics, Enzymology and enzyme technology, Electrophoretic techniques, Immunotechnology, Polymerase Chain Reaction (PCR), Animal Biotechnology, Plant Biotechnology, Methods of Virus Cultivation, Bioinformatics, Plant Tissue Culture, Bioautography, Gel Documentation.

Contact: Dr A. Mohankumar
School of Biological Sciences
Sengunthar Arts and Science College (Campus)
Tiruchengode 637 205
Phone: 04288-283545 (extn 218)
Fax: 04288-284344 (Campus)
Mobile: (0) 98427-91195
E-mail: moniver@satyam.net.in