Role of Premature Stop Codons in Bacterial Evolution

Tit-Yee Wong,1* Sanjit Fernandes,1 Naby Sankhon,1 Patrick P. Leong,2 Jimmy Kuo,3 and Jong-Kang Liu4

Department of Biology, Bioinformatics Program, The University of Memphis, Memphis, Tennessee1; Gigamon Systems, Milpitas, California2; Department of Planning and Research, National Museum of Marine Biology and Aquarium, Pingtung, Taiwan, Republic of China3; and Department of Biological Sciences, National Sun Yat-sen University, Kaohsiung, Taiwan, Republic of China4

When the stop codons TGA, TAA, and TAG are found in the second and third reading frames of a protein-encoding gene, they are considered premature stop codons (PSC). Deinococcus radiodurans disproportionately favored TGA more than the other two triplets as a PSC. The TGA triplet was also found more often in noncoding regions and as a stop codon, though the bias was less pronounced. We investigated this phenomenon in 72 bacterial species with widely differing chromosomal GC contents. Although TGA and TAG were compositionally similar, we found a great variation in use of TGA but a very limited range of use of TAG. The frequency of use of TGA in the gene sequences generally increased with the GC content of the chromosome, while the frequency of use of TAG, like that of TAA, was inversely proportional to the GC content of the chromosome. The patterns of use of TAA, TAG and TGA as real stop codons were less biased and less influenced by the GC content of the chromosome. Bacteria with higher chromosomal GC contents often contained fewer PSC trimers in their genes. Phylogenetically related bacteria often exhibited similar PSC ratios. In addition, metabolically versatile bacteria have significantly fewer PSC trimers in their genes. The bias toward TGA but against TAG as a PSC could not be explained either by the preferential usage of specific codons or by the GC contents of individual chromosomes. We proposed that the quantity and the quality of the PSC in the genome might be important in bacterial evolution.

The universal genetic code categorized TGA, TAA, and TAG as stop codons for the termination of protein translation. The preferential use of a particular stop codon is highly biased among different organisms and different genes (1). The differences among patterns of stop codon usage are likely due to environmental adaptation (1, 14, 16, 23, 25). Most textbooks state that TAA is the preferred stop codon in bacteria (17). Clarke and Miller noticed many TGA, TAA, and TAG trimers in the second and third reading frames of a protein-coding gene (8) and referred to them as premature stop codons (PSC). PSC do not seem to serve any immediate physiological role but could prevent off-frame reading of gene sequences. Clarke suggested that, in Escherichia coli, the occurrence of each type of PSC might be related to the GC content of the bacterial chromosome (7). During a routine nucleotide sequence analysis of a gene from the extremely radioresistant bacterium Deinococcus radiodurans, we noticed that this gene, which contained more than 3,000 nucleotides, had 24 TGA trimers but not a single TAA or TAG trimer in its entire coding region. This interesting phenomenon led us to investigate the PSC profiles of the whole deinococcal genome and, subsequently, other bacterial genomes. In this report, we describe the relationships between PSC frequencies and real stop codon usages in bacteria with different GC contents. We propose that PSC might be important in directing bacterial evolution.

MATERIALS AND METHODS

The bacterial genomic sequences in FASTA Nucleic Acid (FNA) format were downloaded from the Comprehensive Microbial Resource database (http://cmr.sr.org). The results presented in this report are based on 72 bacterial genomes with widely different chromosomal GC contents. Bacterial species, together with their corresponding GenBank accession numbers and GC contents, are listed online at https://umdrive.memphis.edu/tywong/public/Appendix. Bacterial chromosomal DNA sequences were downloaded from the Entrez Genome Project website (http://www.ncbi.nlm.nih.gov/genomes/prokseq.cgi). In this paper, we use the term “gene” to refer to a protein-coding sequence synonymous with an open reading frame (ORF) listed on an FNA file. Non-protein-coding genes and the complementary sequences of protein-coding genes were treated as nongenic sequences. Bacterial codon usage tables were directly downloaded from the Codon Usage Database (http://www.kazusa.or.jp/codon/).

The occurrences of TAA, TAG, and TGA trimers within the protein reading frames in the FNA file were counted by a program written in C language (available upon request) and designated genic counts. Similarly, the occurrences of TGA, TAA, and TAG trimers in the forward and reverse strands of the chromosome were counted and designated total counts. The numbers of TGA, TAA, and TAG trimers of the nongenic fraction were calculated as the differences between the total counts and the numbers of TGA, TAA, and TAG trimers of the genic counts, respectively. The resulting counts, together with the GenBank accession numbers and GC contents of the organism, were transferred to an Excel spreadsheet for subsequent calculations. The types and numbers of real stop codons (which were the last three nucleotides of an ORF) in a genome were also recorded.

We classified codons into four groups, based on their GC nucleotide contents. Codons without any G or C (such as AAA) were designated GC-free codons (zero-G/C codons). Codons with one G or C (such as GAA) were designated low-G/C codons (one-G/C codons). Codons with two G or C nucleotides (such as TGG) were designated moderate-G/C codon (two-G/C codons), while codons composed entirely of G and/or C (such as GCC) were designated high-G/C codons (three-G/C codons). SigmaPlot 9.0 (Systat Software) was used for statistical calculations and plotting.

* Corresponding author. Mailing address: LS 523, Biology Department, The University of Memphis, Memphis, TN 38152. Phone: (901) 678-4462. Fax: (901) 678-4457. E-mail: twong@memphis.edu.

† Published ahead of print on 15 August 2008.
RESULTS

Distributions of TAA, TAG, and TGA in D. radiodurans. The genome of *Deinococcus radiodurans* contains two chromosomes, a megaplasmid (MP1), and a small plasmid (CP1). Table 1 summarizes the frequencies of TGA, TAA, and TAG trimers in each piece of DNA. TGA is much more frequent as a PSC (86% overall) than TAA (7.3% overall) and TAG (6.5% overall). In fact, in chromosome 1 alone, there are more than 900 genes containing only TGA as PSC. The order of abundance of the trimers is TGA > TAG > TAA. This does not reflect the relatively high GC content (66%) of the deinococcal chromosomes: the trimer TAG, which contains the same percentage of G as TGA, is the least frequent PSC.

We also calculated the frequencies of these trimers in the nongenic sequences and in real stop codons in *D. radiodurans*. Distributions of TGA, TAA, and TAG in the nongenic portion of the chromosome are more random, but the occurrence of TGA is still very high. The overall order of abundance (TGA [75.8%] > TAG [15.7%] > TAA [8.4%]) in the nongenic fragments is more in line with the high GC content of the deinococcal chromosomes. Interestingly, for real stop codon usage, while TGA is the most commonly used codon for protein termination (83%), the use of TAA for a stop codon (12%) is almost double the frequency of the use of TAA as a PSC (7.3%). TAG is the least favored stop codon (5%) in *D. radiodurans*. This unusually high frequency of TGA and low frequency of TAG as PSC in the deinococcal genome led us to examine the PSC distributions in other bacterial genomes.

Distributions of TGA, TAA, and TAG trimers in other bacterial genomes. To assess the relative abundances of the three PSC more broadly, and to evaluate the effect of chromosomal GC content in forming these trimers, we analyzed 72 bacterial genomes with widely differing chromosomal GC contents for their frequencies of TGA, TAA, and TAG as PSC. Figure 1 shows a three-dimensional plot displaying the relative percentages of TGA, TAA, and TAG as PSC in each genome. If the use of the three types of PSC was determined by random events, one would expect that the data points should be distributed near the center of the triangular graph, with each type of PSC equal to roughly one-third (33%) of the total PSC. We found instead that the occurrence of each type of PSC is far from random. The frequency of TGA is the most variable and

---

**TABLE 1. Occurrences of TAA, TAG, and TGA trimers in *Deinococcus radiodurans***

<table>
<thead>
<tr>
<th>Chromosome*</th>
<th>Count type*</th>
<th>Total genes*</th>
<th>TAA (%)</th>
<th>TAG (%)</th>
<th>TGA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr1</td>
<td>Genic</td>
<td>2,634</td>
<td>2,633 (7)</td>
<td>2,251 (6)</td>
<td>32,839 (87)</td>
</tr>
<tr>
<td></td>
<td>Nongenic</td>
<td>NA</td>
<td>8,144 (13)</td>
<td>16,021 (20)</td>
<td>36,446 (60)</td>
</tr>
<tr>
<td></td>
<td>1st frame stop</td>
<td>NA</td>
<td>704 (27)</td>
<td>284 (10)</td>
<td>1,669 (62)</td>
</tr>
<tr>
<td>Chr2</td>
<td>Genic</td>
<td>369</td>
<td>481 (8)</td>
<td>443 (7)</td>
<td>5,198 (85)</td>
</tr>
<tr>
<td></td>
<td>Nongenic</td>
<td>NA</td>
<td>1,378 (15)</td>
<td>2,390 (25)</td>
<td>56,309 (60)</td>
</tr>
<tr>
<td></td>
<td>1st frame stop</td>
<td>NA</td>
<td>86 (23)</td>
<td>31 (8)</td>
<td>259 (69)</td>
</tr>
<tr>
<td>MP1</td>
<td>Genic</td>
<td>145</td>
<td>278 (9)</td>
<td>258 (8)</td>
<td>2,707 (83)</td>
</tr>
<tr>
<td></td>
<td>Nongenic</td>
<td>NA</td>
<td>849 (9)</td>
<td>1,133 (8)</td>
<td>3,140 (83)</td>
</tr>
<tr>
<td></td>
<td>1st frame stop</td>
<td>NA</td>
<td>29 (14)</td>
<td>16 (7)</td>
<td>161 (78)</td>
</tr>
<tr>
<td>CP1 plasmid</td>
<td>Genic</td>
<td>40</td>
<td>164 (15)</td>
<td>188 (18)</td>
<td>692 (66)</td>
</tr>
<tr>
<td></td>
<td>Nongenic</td>
<td>NA</td>
<td>366 (19)</td>
<td>525 (29)</td>
<td>943 (51)</td>
</tr>
<tr>
<td></td>
<td>1st frame stop</td>
<td>NA</td>
<td>11 (12)</td>
<td>9 (10)</td>
<td>73 (78)</td>
</tr>
<tr>
<td>Overall</td>
<td>Genic</td>
<td>3,556 (7.3)</td>
<td>3,140 (6.5)</td>
<td>41,436 (75.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nongenic</td>
<td>10,738 (8.4)</td>
<td>20,069 (15.7)</td>
<td>96,868 (75.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1st frame stop</td>
<td>830 (12)</td>
<td>340 (5)</td>
<td>5,691 (83)</td>
<td></td>
</tr>
</tbody>
</table>

* Chr1, chromosome 1; Chr2, chromosome 2; MP1, megaplasmid; CP1, plasmid. Genic counts were calculated from the FNA files.
  
  * Nongenic counts are the total chromosomal counts minus the genic counts. The first frame stop count represents the count for the real stop codon of the genes.
  
  * NA, not applicable.
has the widest span across these species. In *Nocardia farcinica*, 93% of the PSC are TGA triplets, higher than any other organism investigated. In contrast, only 25% of all PSC are in the form of TGA in the genome of *Fusobacterium nucleatum*. The occurrence of TAA as a PSC is more restricted. The genome of *Borrelia garinii* has the highest percentage of TAA trimers (about 51%), while as little as 4% of all PSC trimers are TAA in *Mycobacterium avium* subsp. *paratuberculosis*. The use of TAG as a PSC is the least frequent and most restricted, its use ranging from 32% in *Chlamydia trachomatis* to 4% in *Silicibacter pomeroyi*.

More importantly, the triplets as PSC fall roughly on a line. Bacterial species are not distributed evenly along this line. Along the line, there are many clusters of genomes with similar PSC ratios (TAA/TAG/TGA). Species within each cluster are often phylogenetically related. For example, all the *Chlamydia/Chlamyphila* species, including *Chlamydia abortus* (PSC ratio = 34.8:30.9:34.3), *Chlamydia muridarum* (PSC ratio = 35.0:30.8:34.2), *Chlamydia trachomatis* (PSC ratio = 33.2:32.1:34.8), and *Chlamyphila pneumoniae* (PSC ratio = 33.9:30.7:35.4) are clustered near the center of the graph (Fig. 1, circle A). Members of the genus *Rickettsia*, which includes *R. typhi*, *R. conorii*, and *R. prowazekii*, form a cluster and have a PSC ratio of about 50:23:26 (Fig. 1, circle B). “Candidatus Pelagibacter ubique” is a free-living phototrophic bacterium. This organism has the smallest genome of any cell known to replicate independently in nature (10, 24, 27). This bacterium is placed in the order *Rickettsiales* based on the similarity of the ribosomal gene sequence to those of members of that order (20). The PSC ratio of the free-living “Candidatus Pelagibacter” (50:22:28) closely resembles the PSC ratios of its parasitic counterparts, despite the different lifestyles of these organisms. The enteric clade also forms a cluster with a PSC ratio of about 31:11:58 (Fig. 1, circle C). Included in this cluster are *Erwinia* species (one species), *Shigella* species (one species), *Yersinia* species (three species), *Escherichia coli* (two strains), and *Salmonella* species (two species).

The PSC profiles and the size of the genomes are not related. For example, the genome of the pathogenic *E. coli* O157:H7 EDL933 (5.52 Mb, with 5,679 ORFs) is considerably larger than the genome of the laboratory strain *E. coli* K-12 (4.63 Mb, with 4,289 ORFs). However, their corresponding PSC ratios (31.94:10.39:57.67 and 31.50:10.62:57.87) are almost identical. Thus, the evolutionary history seems to be a major factor affecting the PSC profile of individual organism.

The results shown in Fig. 1 suggested that our initial observation of the high frequency of TGA trimers in the *Deinococcus* genome was not an isolated incident. In fact, many bacterial genes have TGA as 100% of their PSC. In one extreme case, the PSC trimers in 3,561 genes (out of 5,686 genes) in the *Nocardia farcinica* genome are all TGA.

Bacterial use of TGA, TAA, and TAG as real stop codons is more random (Fig. 1). However, the frequency of TAG is still noticeably less than the frequency of TAA or TGA stop codons. The variations in usage of TGA, TAA, and TAG as stop codons among these species range from 8% to 81%, 9% to 76%, and 8% to 50%, respectively. Unlike the linear relationships of PSC, the real stop codon usages among these bacteria appear to cluster. The largest cluster shows a high-TAA preference, centered on 23% TGA, 20% TAG, and 58% TAA. Another cluster shows a high-TGA preference, centered on 68% TGA, 15% TAG, and 16% TAA. There are other bacteria with elevated TAG. However, their populations are too small to be analyzed.

**Relationship between PSC and chromosomal GC content.**

The G nucleotide present in TGA and TAG triplets suggests that their frequencies should increase relative to that of the TAA triplet as the GC content of the chromosome increases. As predicted, the proportion of PSC represented by TGA (Fig. 2) increases linearly with the GC content of the bacterial chromosome, while that represented by TAA drops (Fig. 2). However, much to our surprise, we found that the percentage of TAG in the genome, like TAA, was inversely proportional to the GC content of the chromosome (Fig. 2). Furthermore, the ratios of the percentages of TAG and TAA in a genome follow the equations: $%TGA = (98 - %TAA) \times 0.34$ and $%TAA = (98 - %TGA) \times 0.66$.

Applying data from the 72 bacterial genomes to the equations above showed a standard derivation of less than 3.6. Figure 2 also suggests that the percentage of TGA reaches a lower limit of approximately 25%. This lower limit of 25% TGA may be linked to the GC content of a chromosome (about 27%). However, the TGA upper limit of 94% is substantially higher than the highest chromosomal GC content, approximately 70%.

**Average numbers of PSC trimers per gene in different bacteria.**

The numbers of PSC trimers in the genes, even within the same organism, ranged from zero to more than 150. We found that bacteria with higher GC content tend to contain significantly fewer PSC in their genes than those bacteria with lower GC content. Data on the average numbers of PSC per gene of the 72 bacterial species stratified according to their GC content are presented in Fig. 3 (the insert shows the pairwise comparison of GC content and the average number of PSC per gene). Among these organisms, *Fusobacterium nucleatum* has...
the lowest GC content, 27.1%, while *Thermus thermophiles* has the highest GC content, 69.4%. Although the GC contents of these two organisms differ by only 40%, the number of PSC per gene in *F. nucleatum* (76.4 PSC/gene) is more than sevenfold higher than the number of PSC per gene in *T. thermophiles* (10.4 PSC/gene). Our results further revealed that, with few exceptions (such as that concerning *Staphylococcus aureus*; see below), the average number of PSC per gene in a genome is inversely proportional to the GC content of that genome. This inverse relationship between PSC number and GC content holds true for all three types of PSC (data not shown). Interestingly, most metabolically versatile bacteria have fewer PSC in their genes. For example, the genes of the *Deinococcus, Pseudomonas, Salinibacter, Azoarcus*, and *Klebsiella* spp., as well as those of the low-GC-content bacterium *Staphylococcus*, all contain relatively few PSC (less than 25 PSC/gene) in their genes. On the other hand, bacteria that have a very high number of PSC in their genes often are symbionts. For example, *Fusobacterium* is associated with human oral cavities. *Rickettsia* is an obligate parasite of ticks and humans, and *Borrelia* forms a symbiosis with ticks. There is no indication to suggest that the genes of those symbionts are proportionally longer than the genes of the free-living bacteria. Therefore, gene length cannot be used to explain the disparity between PSC numbers of the high- and low-GC-content groups.

**Relationship between chromosomal GC content and real stop codon usage.** A somewhat different pattern appears when the relative use of real stop codons is compared with the bacterial chromosomal GC content (Fig. 4). As for PSC frequency, increased chromosomal GC content is associated with an increase in the frequency of use of TGA and a decrease in the use of TAA. However, organisms with similar GC contents show substantially more variation in the relative frequency of use of TGA and TAA as real stop codons than in their use of the same triplets as PSC (compare Fig. 2 and 4). Furthermore, the relationship between GC content and stop codon usage for TAA and TGA does not appear to be strictly linear, with the mean frequencies changing little as GC content varied from 24 to 40%, and this then apparently switched abruptly at around 50 to 60%. Even more dramatically, the relative frequency of

---

**FIG. 3.** Relationships between the average number of PSC per gene in a genome and chromosomal GC content on 72 bacterial genomes. These bacteria are placed according to the ranking of their chromosomal GC contents, from low to high. The insert shows the pairwise comparison of GC content versus PSC per gene and the linear regression line.
TAG as a stop codon is nearly independent of GC content, remaining close to 20%. In contrast, its use as a PSC decreases steadily from 25% to 5% as GC content increases (Fig. 2). Although TAG is often the least used stop codon in most bacteria, more than 50% of the genes in Anaplasma marginale St. Maries, a relatively low-GC-content bacterium, used TAG as a stop codon. The use of TAG as a PSC never rose above 32% (Fig. 1 and 2). This result suggests that those factors influencing the selection of real stop codons and the formation of PSC are different.

**Correlations between codon usages and PSC formations.**

The Pareto principle of statistical analysis (12) was employed to evaluate the relationships between the codon usages and PSC frequency. We considered the GC content and the numbers of TAA, TAG, and TGA PSC to be signals for a genome. Eight bacterial genomes having the strongest signals were selected. Two genomes (those of *Thermus thermophiles* and *Deinococcus radiodurans*) were selected to represent the very-high-GC-content group. Two genomes (those of *F. nucleatum* and *Borreliia afzelii*) were selected to represent the very-low-GC-content group. Additionally, the genomes of *Borreliia garinii*, *Chlamydia trachomatis*, and *Nocardia farcinica* were included because they had the highest PSC counts of TAA, TAG, and TGA, respectively. The low-GC-content bacterium *Staphylococcus aureus*, which has relatively few PSC in its genes, was also included in this study. Since the mechanisms leading to the formation of PSC in the second and third reading frames are quite different, formations of PSC in each reading frame were therefore analyzed separately.

The formation of a TAA PSC in the second reading frame results from the juxtaposition of two codons characterized by the sequence [X]TA-A[X][X], where [X] represents any nucleotide. Likewise, formation of a TAG PSC in the second read-

---

**Table 2** shows the frequencies of [X]TA and [X]TG codons in the eight bacterial genomes. As predicted, bacterial genomes with high frequencies of PSC TAA/TAG use the [X]TA codons frequently. This group of bacteria prefers ATA for isoleucine, TTA for leucine, and GTA for valine. They use the [X]TG codons sparingly. Bacterial genomes with high frequency of PSC TGA are just the reverse. This group of bacteria prefers CTG for leucine and GTG for valine and use the [X]TA codons sparingly. While this result suggested a strong correlation between codon usages and the formation of PSC TAA and TGA in the second reading frame, these correlations between codon usage and [X]TA/[X]TG codon frequencies are more likely due to the GC content of the chromosome. When all the synonymous codons are considered, the preferential use of a particular set of synonymous codons is always related to the relative GC contents of the codons and the chromosomal GC content of that particular species. We found several examples of this for the low-GC-content group of bacteria. (i) There are four synonymous codons for isoleucine; the two zero-G/C codons (ATA and ATT) are the most frequently used codons, while the one-G/C codon (ATC) is used sparingly. (ii) Even though both TTA and CTG are [X]TA codons for leucine, the zero-G/C codon TTA is the preferred codon and not but not the one-G/C codon CTG. (iii) There are four synonymous codons for valine, and the one-G/C codons GTA and GTT are used more frequently than the two-G/C codon GTC. (iv) The [X]TG codons, such as CTG (for leucine) and GTG (for valine), are used sparingly, likely because synonymous codons with lower GC contents are available. The opposite is true when the synonymous codons of the high-GC-content group of bacteria are considered. Bacteria with high GC contents often prefer codons rich in G and C.

Formation of TAA, TAG, or TGA trimers in the third reading frame requires a juxtaposition of an [X][X]T codon followed by a AA[X], AG[X], or GA[X] codon, respectively. Careful examination of the codon usage table revealed that the preferential codon usage could not affect the formation of TAA or TGA on the third reading frame. This is because the amino acids encoded by the AA[X] or GA[X] codons have only one synonymous codon and the distinction between the paired synonymous codons is the nucleotide in the third base. For example, the two codons coding for asparagine (AAG and AAC) and glutamine (AAU and AAU) are both AA[X] codons. Thus, either codon would potentially become TAA. Similarly, the two codons coding for lysine (AAG and AAA) are both AA[X] codons. Likewise, the glutamate (GAA and GAG) and aspartate (GAC and GAT) codons are all GA[X] codons. Therefore, any codon coding for asparagine, lysine, glutamate, or aspartate could potentially

---

**FIG. 4.** Correlation between stop codon usages and the GC contents of 72 bacterial chromosomes. TAA as a real stop codon decreases as a function of increasing chromosomal GC content of the organism. TGA as real stop codon increases as a function of increasing GC content of the organism. The use of TAG as real stop codon remains at about 20% regardless to changes in chromosomal GC content. The regression lines are second-order polynomials.
create TAA and TGA trimers on the third reading frame, regardless of the codon preference of the individual organism.

A juxtaposition of [X][X]TAG[X] codons would lead to the formation of TAG in the third reading frame. The AG[X] codons are used by two amino acids, namely, arginine and serine. Both of these amino acids have six synonymous codons. Again, the GC content of the chromosome seems to dictate codon usage in these bacteria. The arginine codon AGA (a one-G/C codon) is more commonly used by the low-GC-content group, while the high-GC-content group (which has relatively more PSC TAG) generally prefers one-G/C codons (AGT, TCA, and TCT), while the low-GC-content group (which has less PSC TAG) generally prefers the two-G/C codons (AGG, TGG, and CTG).

Some bacteria show a very strong preference for a particular type of PSC. However, there is no evidence to support that such a strong preference for a particular type of PSC is affected by the codon usage. For example, Nocardia farcinica has the greatest percentage of PSC TAG in its genes. However, the frequencies of use of the [X]TA codons of these two bacteria are generally comparable to those of other low-GC-content bacteria. Staphylococcus aureus is a low-GC-content bacterium. It has significantly fewer PSC on its genes, but the usages of [X]TA and [X]TG codons by Staphylococcus aureus are not significantly less frequent than the usages of these codons by the rest of the low-GC-content group.

**DISCUSSION**

In this study, we showed that there is a great disparity of PSC in different bacterial genomes (Fig. 1 to 3). There were several important findings. (i) PSC formations were not totally correlated with the chromosomal GC content of a species: the higher the GC content in a chromosome, the fewer the TAG triplets that could be found in the genome. (ii) The PSC profile of a species might be related to the natural history of the species. (iii) The formation of a particular type of PSC was not necessarily related to the codon preferences of the organism. (iv) Organisms with higher chromosomal GC contents often contained significantly fewer PSC in their genes. (v) Unlike the real stop codon usages, the percentages of PSC TGA formed a functional with the bacterial chromosomal GC content. (vi) Finally, metabolically versatile bacteria generally contained fewer PSC in their genes. These observations suggested that the various types of PSC in a genome were the result of

---

**TABLE 2. Bacterial codon usages**

<table>
<thead>
<tr>
<th>Codon type</th>
<th>Amino acid</th>
<th>Codon</th>
<th>Low GC (high TAA and TAG)</th>
<th>High GC (high TGA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[X]TA</td>
<td>Ile</td>
<td>ATA</td>
<td>48.79 39.81 39.33 11.45 18.49</td>
<td>0.38 1.2 0.61</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>TTA</td>
<td>56.59 43.62 43.02 32.43 54.2</td>
<td>0.15 1.15 0.41</td>
</tr>
<tr>
<td></td>
<td>CTA</td>
<td>7.79  9.6 10.01 14.91 8.63</td>
<td>0.79 3.25 1.72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GTC</td>
<td>22.43 15.21 15.28 16.89 22.81</td>
<td>1.53 1.38 2.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CTG</td>
<td>1.1   2.18 2.27 9.9 2.41</td>
<td>60.65 41.31 59.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GTG</td>
<td>4.27  5.1 4.89 12.95 9.48</td>
<td>47.05 49.82 39.95</td>
<td></td>
</tr>
<tr>
<td>AG[X]</td>
<td>Arg</td>
<td>AGG</td>
<td>2.32 6.08 6.17 2.6 1.46</td>
<td>3.01 16.13 3.14</td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>AGA</td>
<td>27.82 20.38 20.55 11.29 11.62</td>
<td>0.59 0.85 2.03</td>
</tr>
<tr>
<td></td>
<td>Ser</td>
<td>AGT</td>
<td>16.78 16.11 16.21 10.37 16.56</td>
<td>1.75 0.39 3.55</td>
</tr>
<tr>
<td></td>
<td>Ser</td>
<td>AGC</td>
<td>2.61 9.38 9.96 8.45 5.07</td>
<td>12.76 13.62 23.42</td>
</tr>
<tr>
<td>Related synonymous codons</td>
<td>Arg</td>
<td>CCG</td>
<td>0.04 0.45 0.45 4.22 0.48</td>
<td>27.29 36.35 18.76</td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>CGA</td>
<td>0.23 1.8 1.88 9.52 4.99</td>
<td>3.58 1.27 3.14</td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>CGT</td>
<td>1.37 1.94 1.94 13.04 13.58</td>
<td>5.04 1.64 10.34</td>
</tr>
<tr>
<td></td>
<td>Arg</td>
<td>CGC</td>
<td>0.11 0.86 0.94 7.65 3.22</td>
<td>42.48 29.01 46.34</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>ATT</td>
<td>44.53 58.26 58.1 34.81 52.71</td>
<td>1.31 2.45 9.63</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>ATC</td>
<td>5.86 7.96 7.9 19.83 14.76</td>
<td>35.5 22.45 24.94</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>CTT</td>
<td>18.14 29.04 29.39 22.77 10.88</td>
<td>1.04 16.52 4.36</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>CTC</td>
<td>0.65 2.29 2.37 12.15 2.1</td>
<td>31.42 73.19 33.87</td>
</tr>
<tr>
<td></td>
<td>Ser</td>
<td>TCG</td>
<td>0.6 2.58 2.57 6.81 3.86</td>
<td>17.22 3.77 11.86</td>
</tr>
<tr>
<td></td>
<td>Ser</td>
<td>TCA</td>
<td>19.19 17.12 17.17 9.59 20.06</td>
<td>0.94 0.32 2.33</td>
</tr>
<tr>
<td></td>
<td>Ser</td>
<td>TCT</td>
<td>19.31 25.07 24.95 31.56 12.2</td>
<td>0.51 0.81 2.43</td>
</tr>
<tr>
<td></td>
<td>Ser</td>
<td>TCC</td>
<td>1.09 3.25 3.32 12.27 1.71</td>
<td>13.3 15.25 8.01</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>GTT</td>
<td>32.1 28.8 28.99 24 27.16</td>
<td>1.78 2.54 2.74</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>GTC</td>
<td>2.11 2.54 2.76 10.28 7.36</td>
<td>37.39 27.21 22.2</td>
</tr>
</tbody>
</table>

* Codon usages (per 1,000 nucleotides) of [X]TA, [X]TG, and AG[X] codons of *Thermus thermophiles* (Tt), *Deinococcus radiodurans* (Dr), *Fusobacterium nucleatum* (Fn), *Borrelia afzelii* (Ba), *Borrelia garinii* (Bg), *Chlamydia trachomatis* (Ct), *Nocardia farcinica* (Nf), and *Staphylococcus aureus* (Sa). Also shown are the usages of their corresponding synonymous codons.
certain selective pressure imposed on the bacterial genomes. Herein, we propose that both the quantity and quality of PSC may be important for bacterial genome evolution.

The Ohno theory of evolution by genome expansion is widely accepted (4, 15, 18). Genome expansion by DNA recombinations would create a repertoire of redundant genes. The extra copies of duplicated genes are free substrates for the evolution of proteins through base substitutions (6). Many details regarding the mechanism of genome evolution remain unclear. The environment of some bacteria may change rapidly and, the species would become extinct if they failed to respond in time. However, genomic change through base substitutions is a slow process (11). Fusion of the surplus redundant genes could allow some bacteria to leap forward rapidly. Gene recombination has two patterns: in-frame recombination and off-frame recombination. Obviously, in-frame concatenation of two gene fragments would allow a gene to elongate, but because the amino acids encoded by the newly formed gene are essentially identical to the original parental sequences, the topology (and thus the function) of a newly formed fusion gene product might not differ significantly from that of its parental proteins. This new gene would require a long time of continual modifications to evolve into a functionally different protein. Therefore, in-frame recombination genes might not change fast enough to allow adaptation to a rapidly changing environment. The most effective way to create new protein significantly different from the parental proteins is by off-frame recombination. Because of the frameshift effect, off-frame recombination would instantaneously create a protein with a different topology (and possibly new function). If this new protein could enhance the survival of the cell, the species would survive. Thus, the success of off-frame recombinations in creating a new functional gene would be influenced by the quantity of PSC trimers in the DNA; the fewer the number of PSC trimers in the gene fragment, the more likely a longer DNA fragment could be inserted into an existing gene without truncation.

In addition to the number of PSC trimers in the genes, the quality of each type of PSC trimers might also be important for the success of an off-frame gene recombination event. The efficiencies of TGA, TAA, and TAG as translational stop codons are quite different. In Salmonella enterica serovar Typhimurium, the leakiness of the TGA codon occurs at a frequency of at least $10^{-2}$ to $10^{-3}$ (21), while TGA and TAA are more error-proof, at about $7 \times 10^{-3}$ to $1.1 \times 10^{-4}$ (5) and $9 \times 10^{-4}$ to $1 \times 10^{-3}$ (22), respectively. The mechanism of termination at the TGA site is complex. In many cases, TGA is the site of a programmed frameshift. A programmed frameshift allows the ribosome to pass over the UGA stop signal and continue reading the mRNA. Programmed frameshift is a regulatory mechanism for controlling the gene expression of many proteins (13, 19). Furthermore, protein termination at the bacterial UGA site is not straightforward. Successful termination at the UGA site is influenced not only by the concentration of the releasing protein RF2 (25) but also by the ribonucleotide sequences before and after the UGA site on the mRNA (26). In fact, the expression of the bacterial RF2 gene is itself governed by a programmed frameshift (25). TGA also serves as the code for selenocysteine (16), the 21st amino acid, in a wide range of organisms but tryptophan in other organisms and organelles (14). Interestingly, mutations at the TGA site in various genes are very rare (21). All these evidences suggested that TGA might not serve solely as a terminator for protein translation. Perhaps, in the event of an off-frame gene fusion, the presence of these leaky TGAAs might allow some off-frame reading, at least temporarily, until better DNA repair (such as base substitution) could occur. Thus, the bias toward a leaky TGA might provide an immediate and provisional relief to a bacterium facing a rapidly changing environment. The use of TGA as a “switching point” might be analogous to the use of ATA codon in many proteins. For example, Barrai et al. (2, 3) showed that ATA could stimulate polymerase attachment inside coding sequences. In fibrinogen, the relative abundance of the ATA codons might permit the production of pieces of the fibrinogen molecule still functional in a stoppage network and benefit the organism.

The strategy of creating new genes by off-frame recombination might benefit only some organisms and might be harmful to others, depending on the ecophysiological status of the individual organism. The metabolically versatility of a bacterium is often correlated to its environment (9). This might explain why many bacteria, such as Deinococcus, Pseudomonas, Salmonibacter, Azoarcus, Klebsiella spp., and the skin-associated Staphylococcus, all known for their metabolic versatilities, contain very few PSC trimers in their genes (Fig. 3). Symbionts, such as Fusobacterium, Rickettsia, and Borrelia, often contain large number of PSC trimers in their genes. Symbionts live in a stable environment. A sudden change by frameshift recombination would create an undesirable product and disrupt their long-established relationship with the host, leading to the collapse of the symbiotic relationship. This might explain why there are many more PSC serving as check marks in the genes and the use of the more error-proof TAA and TAG as PSC in their genes. We should emphasize that, although the number of PSC trimers in the gene and GC content of a chromosome are generally inversely related, and although the intrinsic nature of the stop codons are AT rich, the notion that TA-rich genomes should contain more PSC is not true. For example, the genome of Staphylococcus aureus is AT rich (GC content = 32.8%), but its PSC in the gene is also quite low (less than 25 per gene). This versatile pathogen, which is well known for its resistance to antibiotics, is commonly found on the skin. Unlike other intracellular parasites, the environment of the skin changed rapidly. Having fewer PSC in its genome might enhance survival of S. aureus.

The inverse relationship between the GC content and the number of TAG codons remains unexplained. Perhaps bacteria use a yet-unknown mechanism to remove TAG sequences from their genes. Systematically replacing those AG[X] and [X]TA codons with other codons in the gene would eventually reduce the number of TAG triplets in the genome. Therefore, the higher the GC content, the fewer TAG codons that could be found in the gene.

ACKNOWLEDGMENTS

Part of this research was supported by the “ Aim for the Top University Plan” grant by National Sun Yat-sen University to T.-Y. Wong and research grant NSC95-2621-B110-004, National Science Council, Taiwan, Republic of China, to J. K. Liu.

We kindly acknowledge K. Gartner, M. Beck, and S. Schwartzback for their critical reviews and discussions.
REFERENCES