J Mol Evol (1999) 49:49–62

JOURNAL OF MOLECULAR EVOLUTION

Two Aspects of DNA Base Composition: G+C Content and Translation-Coupled Deviation from Intra-Strand Rule of *A* = *T* and *G* = *C*

Noboru Sueoka

Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309-0347, USA

Received: 5 November 1998 / Accepted: 1 March 1999

Abstract. The relative contribution of mutation and selection to the G+C content of DNA was analyzed in bacterial species having widely different G+C contents. The analysis used two methods that were developed previously. The first method was to plot the average G+C content of a set of nucleotides against the G+C content of the third codon position for each gene. This method was used to present the G+C distribution of the third codon position and to assess the relative neutrality of a set of nucleotides to that of the G+C content of the third codon position. The second method was to plot the intrastrand bias of the third codon position from Parity Rule 2 (PR2), where $A = T$ and $G = C$. It was found that whereas intragenomic distributions of the DNA G+C content of these bacteria are narrow in the majority of species, in some species the G+C content of the minor class of genes distributes over wider ranges than the major class of genes. On the other hand, ubiquitous PR2 biases are amino acid specific and independent of the G+C content of DNA, so that when averaged over the amino acids, the biases are small and not correlated with the DNA G+C content. Therefore, translation coupled PR2-biases are unlikely to explain the wide range of G+C contents among different species. Considering all data available, it was concluded that the amino acid-specific PR2 bias has only a minor effect, if any, on the average G+C content. In addition, PR2 bias patterns of different species show phylogenetic relationships, and the pattern can be as a taxal fingerprint.

Key words: DNA G+C content — Parity Rule 2 — PR2-bias finger print — Translation-coupled PR2 bias — Bacteria

Introduction

The average G+C content of bacterial DNA varies among species from approximately 25 to 75% (Lee et al. 1959; Belozersky and Spirin 1958), with narrow intraspecific heterogeneity (Sueoka et al. 1959; Rolfe and Messelson 1959; Sueoka 1961a; Schildkraut et al. 1962). It was proposed that mutation, rather than selection and genetic drift, is the major cause of the above-mentioned characteristics of bacterial DNA (Sueoka 1962, 1988). Under this hypothesis, interspecific G+C variation was interpreted as a result of the equilibrium of the mutation rates in two opposing directions, i.e., v ($\alpha \rightarrow \gamma$) and u ($\alpha \leftarrow \gamma$). Here, α represents the A/T or T/A nucleotide pair, and γ represents the G/C or C/G nucleotide pair in the double-stranded DNA. The mutation rates, u and v, are defined per nucleotide per time. The directional mutation pressure was assumed to act directly on both strands of DNA and uniformly throughout the genome, whereas selection was to exert its effect mainly through the function of mutated proteins. It was therefore difficult for selection to affect the DNA G+C content in a directional manner toward a higher or lower G+C content. Random genetic drift was not considered here as a major cause of the interspecific G+C variation, because under bidirectional mutation pressures, random drift does not generate directionality and narrow distributions of the G+C content of the genes of an organism. The range of the average G+C content of bacterial DNA is even wider in the third codon position, covering almost the entire range. For example, the G+C range is 7 to 95% for the species average and 4 to 98% for individual genes in the sample of species used in this study.

Codon usage biases from equimolar compositions of four nucleotides (A, T, G, and C) were noted as soon as the DNA sequencing became a reality. Codon usage bias is a common feature of all organisms examined so far and is specific to species as first noted by Grantham et al. (1980). In 1980s, it became evident that codon usage bias among synonymous codons of each amino acid correlates with the abundance of corresponding tRNA in

Escherichia coli (Ikemura 1981), yeast (Bennetzen and Hall 1982; Ikemura 1992), and *Drosophila* (Moriyama and Powell 1997). It has also been shown that the codon usage bias is correlated with an increased abundance of proteins using preferred codons in *E. coli* (Gouy and Gautier 1982; Ikemura 1982, 1985a; Shields and Sharp 1987; Sharp and Li 1987), in yeast (Bennetzen and Hall 1982), in slime mold (Sharp and Devine 1989), and in *Drosophila* (Shields et al. 1988; Sharp and Li 1989; Moriyama and Powell 1997). Thus, in addition to directional mutation pressure, selection is a significant factor in the codon usage biases. These findings indicated that selection also plays a significant role in the evolution of DNA base composition and raised questions about the validity of the mutational view. For the DNA G+C content, however, the relative contributions of mutation and selection have not been estimated either theoretically or empirically. The situation has also created ambiguity and confusion about the relative effects of mutation and selection on the DNA base composition including the G+C content. To clarify the situation, it was important to recognize two basic aspects of the DNA nucleotide composition, the G+C content and the violation of Parity Rule 2 (PR2) and to assess the relative contributions of mutation and selection to the two aspects. These parameters should be treated separately, rather than working on one parameter for codon usage biases. PR2 is a compelling principle that is a natural consequence of the structure and function of the double helix of DNA. PR2 predicts that when there are no biases between the two DNA strands in mutation and selection, the DNA intrastrand base composition is expected to be $A = T$ and $G = C$ (Sueoka 1995; Lobry 1995; see also Wu and Maeda 1987; Wu 1991; Furusawa and Doi 1992).

In unicellular as well as multicellular organisms, the selectional effect of tRNA abundance may not exert a large directional influence on the G+C content of each gene because of the amino acid-specific nature of the tRNA effect (Sueoka 1995). For selection to change the GC_3 content (G+C content of the third codon position) directionally toward higher or lower G+C, the abundance of specific tRNAs or the proficiency of codon recognition by tRNAs for synonymous codons of 18 amino acids (20 amino acids minus methionine and tryptophan) must change in a concerted way. In fact, not a single selectional agent has been identified for moving the GC_3 together in one direction or another to account for the wide variation and heterogeneity of the DNA G+C content. On the other hand, mutation pressure, for example, the effect of a mutator mutation, has been shown to change the base composition directionally (Cox and Yanofsky 1967, 1969). Thus, a mutational effect is expected to be directional and uniformly effective on the whole genome or at least large regions of the genome, for example, isochores of vertebrates discovered by Bernardi et al. (1985).

The primary purpose of this paper is to evaluate quantitatively the relative contributions of mutation and selection to the evolution of DNA G+C content as well as to the ubiquitous violation of PR2. For this purpose, the G+C content of the third codon position (P_3) and violation of PR2 was analyzed using species with widely differing G+C contents of DNA. The result indicates that the major factor affecting P_3 is directional mutation pressure and that selection through tRNA abundance is likely to cause the bias from the intrastrand PR2 and may have a minor, if any, effect on P_3.

Methods

Definition and Calculation of P *Values*

The third codon position of synonymous codons includes α and γ nucleotide pairs in equal numbers (symmetric) for most amino acids, except for tryptophan (TGG), methionine (ATG), and one (ATA) of the three isoleucine codons. P_3 is the G+C content of the third codon position that is symmetric with regard to α and γ nucleotides in synonymous codons. Thus, P_3 is defined as the G+C content of the total codons minus ATG (Met), TGG (Trp), ATA (Ile), and the termination codon (TAA, TAG, or TGA). These codons are also removed from the calculations of the G+C contents of the first codon position (P_1) and the second codon position (P_2). Subtraction of these codons in this analysis eliminated odd-numbered synonymous codon sets and, therefore, an extra cause of potential asymmetry from the expected rule, PR2, where $A = T$ and $G = C$. In practice, this parameter (P_3) is only slightly different from GC_3 and the G+C content of the third codon position of synonymous codons (P_s).

Plots vs. P_3 *(Relative Neutrality Plots)*

When the G+C content of a set of nucleotides (P) is plotted against P_3, the regression coefficient (slope) represents the neutrality of the nucleotide set relative to the neutrality of P_3 (Sueoka 1988). The relative neutrality to P_3 is designated RN_{P3} here. The cross point of the regression line and the diagonal line is defined as the optimum point (OP). This position in the plot was previously termed the equilibrium point (EP) (Sueoka 1988). The slope was calculated by the least-squares method, and the OP value was calculated as [(ordinate intercept at $P_3 = 0$)/(1 − slope)], as described previously (Sueoka 1988).

PR2-bias Plots

In PR2-bias plots, the abscissa and ordinate are $G/(G + C)$ and $A/(A + T)$, respectively (Sueoka 1995). To avoid asymmetry between α and γ nucleotide pairs at the third codon position, ATG, TGG, ATA, and termination codons were not included in calculations. The center of the plot represents the point where the intrastrand nucleotide composition follows PR2 ($A = T$ and $G = C$), and the distance and direction from the center represent the extent and direction of biases from PR2. In a PR2-bias plot, "$G_3/(G_3) + C_3 \mid 4$" and "$A_3/(A_3 + T_3) \mid 4$" of individual genes are plotted for the abscissa and the ordinate, respectively. Here, "| 4" denotes the four-codon amino acids: alanine, arginine (CGA, CGT, CGG, CGC), glycine, leucine (CTA, CTT, CTG, CTC), proline, serine (TCA, TCT, TCG, TCC), threonine, and valine.

Results

Ten bacterial species were selected based on their representation of a wide range of genomic G+C contents (25–75%) as well as average P_3 values (7–95%), where

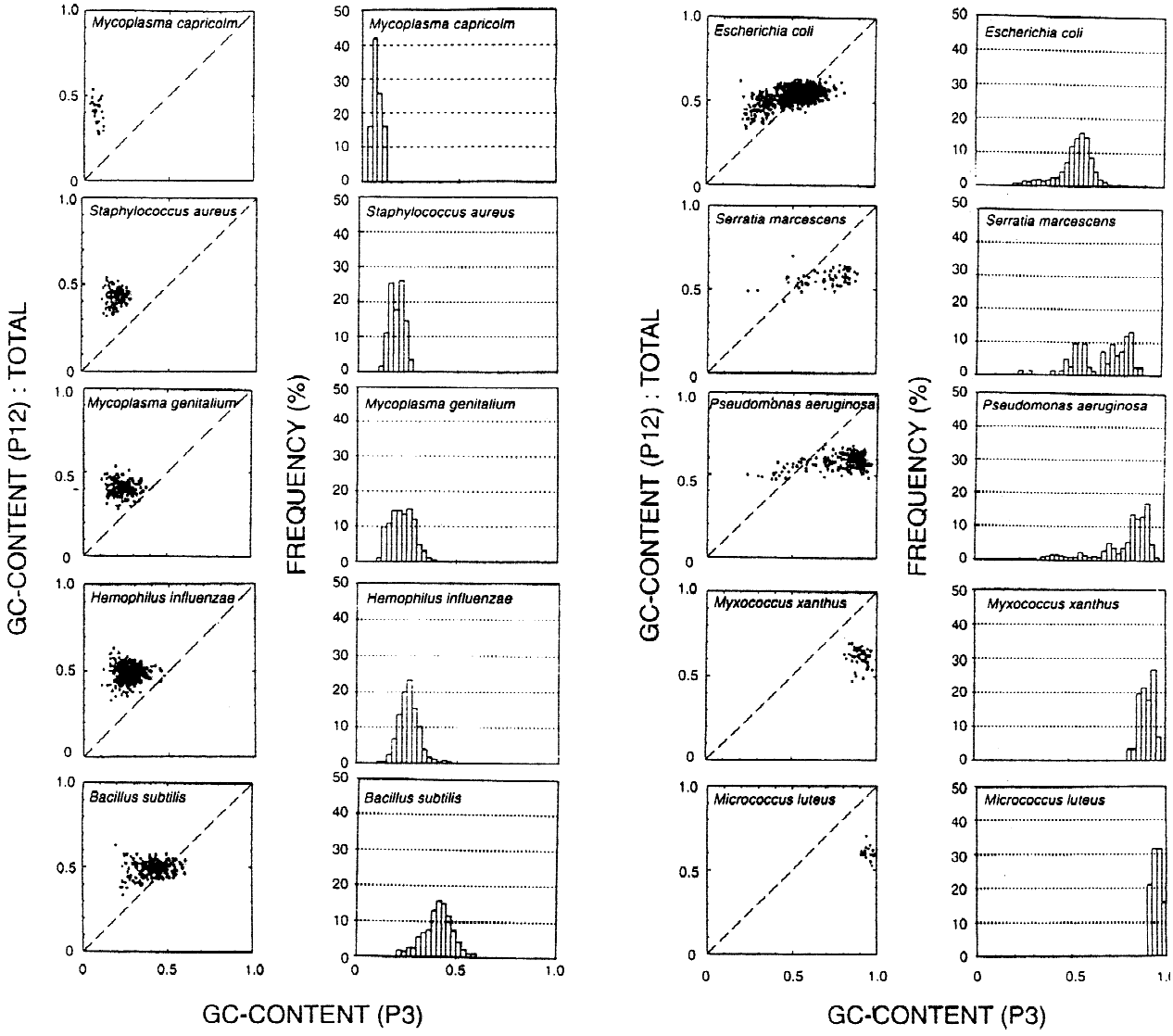

Fig. 1. Plots of P_{12} vs. P_3 and the frequency distribution of P_3 for 10 bacterial species. The average G+C content (P_{12}) of the first codon position (P_1) and the second codon position (P_2) of individual genes in each bacterial species is plotted against the G+C content of the third codon position (P_3). The frequency distribution of P_3 of individual genes for each species is also presented as a histogram. Definition of *P* values and specific information for individual species are presented in the Appendix.

a relatively large number of genes of each species is sequenced. The average of P_1 and P_2 (P_{12}) was plotted against P_3 for each bacterial species, and the average bias from Parity Rule 2 (PR2) was examined for the four-codon amino acids.

Regression Analysis of P_{12} *and* GC_{igs} *vs.* P_3 *(*RN_{P3} *Plot Analysis)*

P_{12} vs. P_3 Plot. Figure 1 presents the regression of P_{12} to P_3 for genes of 10 bacterial species. The figure also shows the distribution of P_3 as a histogram for each species. Genes of six bacteria classified as Type I (*Mycoplasma capricolum, Staphylococcus aureus, Mycoplasma genitalium, Haemophilus influenzae, Myxococcus xanthus,* and *Micrococcus luteus*) have P_3 with a unimodal and narrow distribution, whereas genes of four bacterial species classified as Type II (*Escherichia coli, Pseudomonas aeruginosa, Serratia marcescens,* and *Bacillus subtilis*) have wide and skewed intragenomic heterogeneity in P_3. The results suggest that in Type II bacteria, there are at least two classes of genes, major and minor. The major class consists of 70 to 90% of the genes, and these genes have narrow P_3 distributions. The minor class consists of the remaining 10 to 30% of the genes analyzed, and these genes have wide P_3 distributions. (The percentages are approximage estimates from histograms in Fig. 1). The classification of bacterial species into Type I and Type II is purely a descriptive convention, and the G+C range of minor class genes is variable among bacterial species. Specific information on the

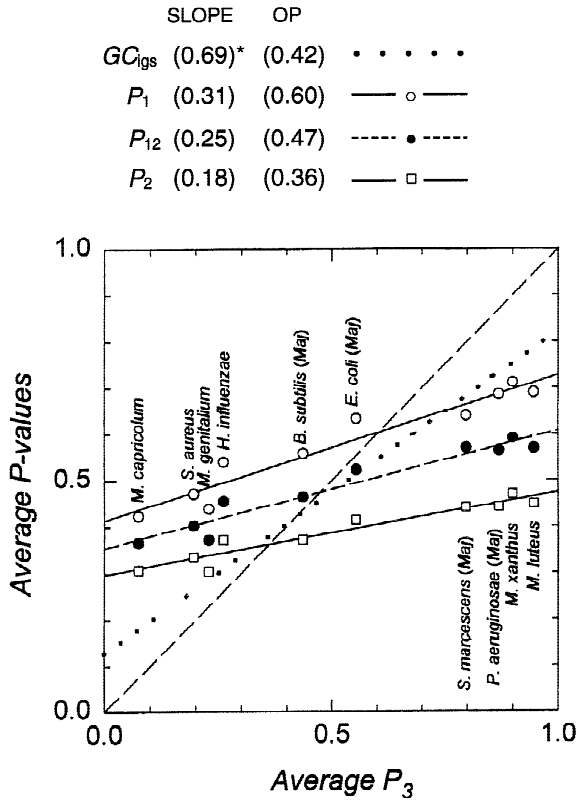

Fig. 2. Relative neutrality poltos of P_1, P_2, P_{12}, and GC_{igs} vs. P_3 (RN_{P3} plots). The P values were calculated as the average P values of genes of Type I bacteria and major class genes of Type II bacteria. In the latter cases, the names of Type II bacteria are shown with "(Maj)." The *dotted line* represents the slope of intergenic spaces (GC_{igs}) that was calculated from the data on intergenic spaces previously reported by Muto and Osawa (1987). The regression coefficient (slope) represents the neutrality of each set of nucleotides relative to the neutrality of P_3 (Sueoka 1988, 1992). The slope of the *diagonal dashed line* represents the neutrality of P_3, which is defined as 1. The values of the slope and OP (see Methods) of each regression line are presented at the *top*.

G+C heterogeneity of each bacterial species is presented in the Appendix.

In Fig. 2, averages of P_1, P_2, and P_{12} for individual bacteria are plotted against P_3. The regression coefficients (slopes) and corresponding OP values are shown at the top in the figure. In this RN_{P3} plot (see Methods), for example, the slope for P_{12} values is an estimate of the neutrality of P_{12} relative to the neutrality of P_3 (RN_{P3}), and the value (1 − slope) is an estimate of the selectional constraint of P_{12} relative to that of P_3 that is close to 0. Thus, the slope of P_{12} (0.25) in Fig. 2 signifies that an estimate of the neutrality of P_{12} is 25%, relative to the neutrality of P_3 in this set of bacterial species. The relative neutrality of P_{12}, 25%, is equivalent to stating that an estimate of the selectional constraint of P_{12} is 75%. Similarly, RN_{P3}'s of P_1 and P_2 are 31 and 18%, respectively, and the corresponding relative constraints are 69 and 82%, respectively. In these estimations, the neutrality of and selectional constraint against P_3 are approximated to be 1 and 0, respectively.

At the OP (Fig. 2), the values of P_3 and P are the same, and there is no more constraint on P than on P_3. When P_3 is found on either side of the OP, P is likely to be pressured toward the OP value by the vertical distance from OP, that is, $(P - P_{OP})$ multiplied by a constant factor of (1 − slope). The factor (1 − slope) is the selectional constraint, and P_{OP} is the P value at the OP. This principle generates a linear regression with a coefficient less than 1 (Sueoka 1988, 1992). For example, under this model, when $P_{12} = 0.47$ (P_{OP} value of P_{12}), P_{12} is optimal and not susceptible to selectional constraint. In other words, P_{OP} may represent the functionally optimal G+C content for a particular set of nucleotides. This interpretation may be reasonable in light of the existing correlation of the DNA G+C content with relative contents of amino acids of cytosolic proteins (Sueoka 1961b) and with both cytosolic and membrane-bound proteins (Lobry 1997). Therefore, it is likely that the OP value for P_{12} may correspond to an optimal amino acid composition for proteins in the bacterial cytoplasm. Note that when there is little negative selection, average P_{12} values of different species will be on the diagonal line (slope = 1), whereas when selection against directional mutations is complete, the slope will be zero. Therefore, in bacteria, estimates of relative contributions of directional mutation and selection to the evolutionary changes of P_{12} are approximately 25 and 75%, respectively. It is reminded that P_{12} represents the G+C content of the nucleotides that are most susceptible to amino acid replacements by nucleotide substitutions.

G+C Content of Intergenic Spaces (GC_{igs}). Since intergenic spaces and introns are presumably free from the influence of transcription and translation, a detailed analysis of intergenic spaces should yield valuable information regarding the relative contribution of mutation vs. selection to the evolution of the DNA G+C content (Muto and Osawa 1987). When GC_{igs} was calculated in Fig. 2 using the data of Muto and Osawa (1987), the slope of GC_{igs} vs. P_3 is 0.69 in bacteria, indicating that, relative to the neutrality of P_3, the G+C content of intergenic spaces is about 70% neutral against selection. In comparison, slopes of introns vs. P_3 are 0.62 in vertebrates combined (Aota and Ikemura 1986) and 0.46 in human genes (Mouchiroud et al. 1991). A positive regression coefficient of GC_{igs} to the DNA G+C content has been regarded as evidence of directional mutation pressure acting on the set of nucleotides in question (Muto and Osawa 1987).

PR2 Bias Analysis (Strand Bias Analysis)

Intrastrand parity rules of DNA predict that under strand-independent mutation and selection between the two strands of DNA, the intrastrand nucleotide composition

should closely follow PR2, where $A = T$ and $G = C$ within the strand. Violation of PR2 is ubiquitous in the coded regions in the organisms examined so far. As is evident in Fig. 3, PR2 bias is amino acid specific, which indicates that the major ubiquitous cause for the violation must be strand-specific selection (Sueoka 1995). Therefore, to distinguish the effect of directional mutation pressure from that of selection on the evolutionary changes of P_3, the PR2 biases were studied further.

General Features. Figure 3 shows PR2-bias plots of genes for the 10 bacterial species. In these plots, only the average points for individual four-codon amino acids are shown, and the diameter of bubbles reflects the relative frequency of the individual four-codon amino acids. In the PR2-bias plot, the center of the graph represents the absence of PR2 biases in both AT and GC, and the farther from the center, the stronger the bias. Thus, the distance from the center is a vector including both the magnitude and the direction of PR2 bias. The quantitative features of the parity plot for *E. coli* are expected to be consistent with the result of the positive correlation between codon usage frequencies and tRNA abundance (Ikemura 1981; Sharp and Li 1987a). To confirm this correlation, further studies on the PR2 biases and relative expression of individual genes are necessary.

The phylogenetic kinship of three bacteria, *E. coli, S. marcescens,* and *P. aeruginosa,* is clear in PR2-bias plots (amino acid specific PR2-bias plots; Figs. 3 and 4). Note that despite the significant difference in P_3, the PR2 bias patterns have a clear similarity among the three bacterial species, whereas the patterns of all other bacteria are uniquely different from each other. These three bacteria have been classified as phylogenetically related by sequence analysis of 16S rRNA (Woese 1987) and 5S rRNA (Hori and Osawa 1987) and by amino acid sequence analysis of proteins (e.g., Lloyd and Sharp 1993).

The codon usage bias in the genes of *B. subtilis* has been reported to be lower compared to that in other bacteria (Ogasawara 1985) and less correlated with the level of gene expression (Shields and Sharp 1987). Figure 3 includes PR2-bias plots for four-codon amino acids of the major and minor class genes of *B. subtilis.* The results show a unique PR2-bias pattern as is also true in other species, but their patterns of both classes of genes are not particularly compressed compared to those of other species. It is possible that the correlation between the abundance of protein and codon preference could be enhanced in some bacteria but not in others. However, it is clear that PR2 biases exist in all the organisms this author has examined. However, this result does not exclude the possibility that a correlation between tRNA abundance and codon usage bias always exists even if the correlation between abundance of protein and the bias of codon usage is not observed in the organism analyzed.

Genes of Type I Species. Figures 5A and B show the pattern of the average PR2 bias of the third codon positions of four synonymous codons and that of the third codon position of all amino acids, respectively. It is noted that there is no apparent relationship between the average PR2 bias and the P_3 of each species. As an exception, *M. luteus* has a large average PR2 bias. The anomaly of *M. luteus* should be analyzed further when more gene sequences become available.

Genes of Type II Species. Figure 6 shows the PR2-bias patterns of major and minor class genes in Type II bacteria. In *B. subtilis* and *P. aeruginosa,* the PR2-bias patterns of the two classes of genes are similar. However, there is a clear tendency for minor class genes to have a more compressed pattern of PR2 bias than major class genes, indicating that minor class genes have less PR2 bias. This result may indicate that in Type II bacteria, minor class genes are expressed less abundantly than major class genes, if we assume that the extent of PR2 bias reflects the abundance of gene expression. In *E. coli,* the PR2-bias pattern of minor class genes is much more compressed than that of major class genes, and the relative positions of serine and glycine in the plot also differ. This finding is consistent with the fact that minor class genes of *E. coli* belong to a unique class of genes (see the Appendix). In *S. marcescens,* the PR2-bias patterns of the two classes are slightly different, although that of the minor class may be somewhat compressed.

Discussion

P_3 as the Standard for the Neutral G+C Scale

Whether it is best to use P_3 as the neutrality standard has not been clear, because there has been no method available to estimate the influence of selection through tRNA on the DNA G+C content. There has been no answer to the question which of the two, directional mutation pressure or selection, is the major factor that influences the G+C content in bacteria and mammalian genomes. In human, no significant regressions were found between PR2 biases and P_3 (Sueoka 1995).

P_3 has the following properties that are suitable for the relative neutrality standard. (a) P_3 covers the widest range of G+C content in bacteria and reveals a remarkable linearity of significant regression for P_{12} over the entire range (e.g., Fig. 2). (b) The definition of P_3 is applicable to any organism and is based on a straightforward, well-defined set of nucleotides. (c) There is some evidence that directional mutation pressure affects P_3. For example, a strong positive regression to P_3 exists in the G+C content of intergenic spaces (GC_{igs}) among bacterial species (0.69; Fig. 2) and in introns (GC_{int}) of vertebrates combined (Aota and Ikemura 1986) and those of mammals (Bernardi et al. 1988). Moreover, in human, the third codon position of practically every

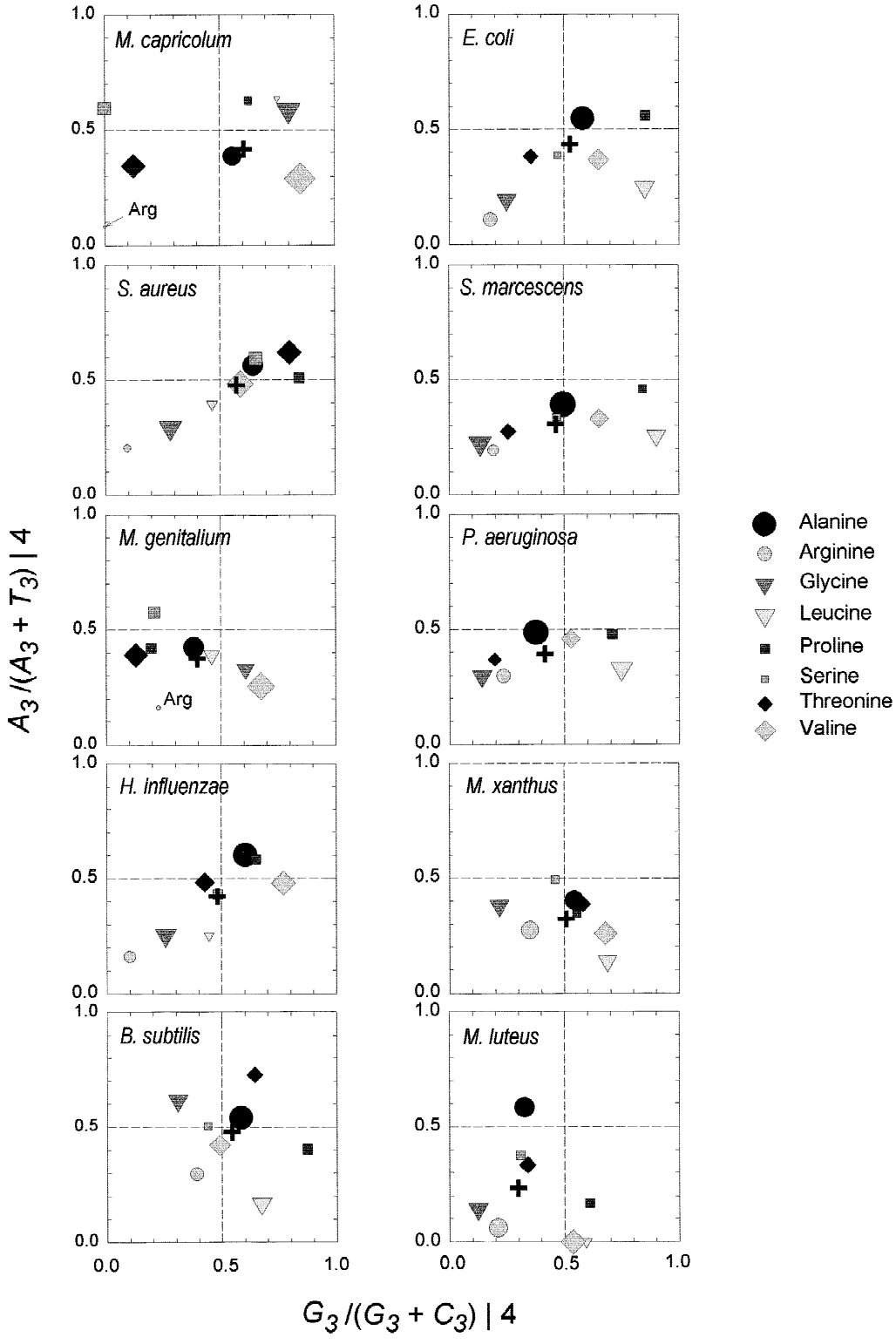

Fig. 3. PR2-bias plots of bacterial genes. The *symbols* represent the average of "$G_3/(G_3 + C_3) \mid 4$" for the abscissa and "$A_3/(A_3 + T_3) \mid 4$" for the ordinate. A_3, T_3, G_3, and C_3 represent the content of the four nucleotides of DNA at the third codon position and "| 4" indicates that the calculation was made for the codons of the four-codon amino acids. For arginine, leucine, and serine, four of the six synonymous codons were used; the first and second codon letters are the same in each case. In the PR2-bias plot, Parity Rule 2 ($A = T$ and $G = C$) is held at the center and the distance from the center represents the extent of bias (both magnitude and direction) from the rule. The size of the symbols indicates the relative frequency of the amino acid, and the "+" symbol represents the weighted mean for all of the four-codon amino acids. The configuration of symbols for each species may be regarded as the PR2-bias fingerprint.

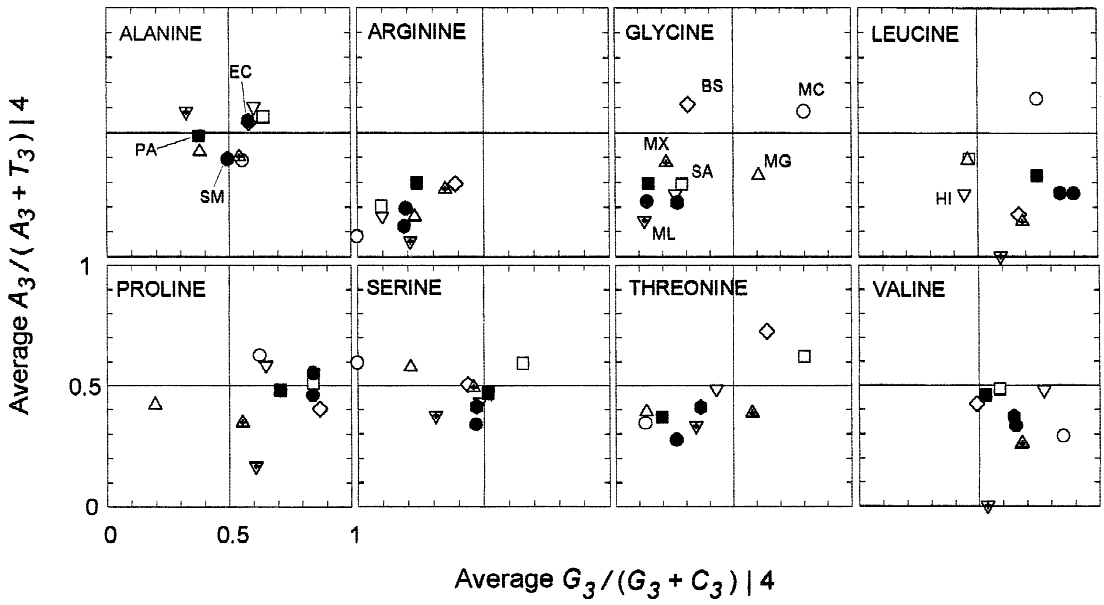

Fig. 4. Average PR2-bias plot of 10 bacterial species for the four-codon amino acids. The data presented in Fig. 3 were plotted for each amino acid. MC, *M. capricolum* (○); SA, *S. aureus* (□); MG, *M. genitalium* (△); HI, *H. influenzae* (▽); BS, *B. subtilis* (◇); EC, *E. coli* (●); SM, *S. marcescens* (●); PA, *P. aeruginosa* (■); MX, *M. xanthus* (▲); ML, *M. luteus* (▼). *Filled symbols* indicate the trio of EC, SM, and PA.

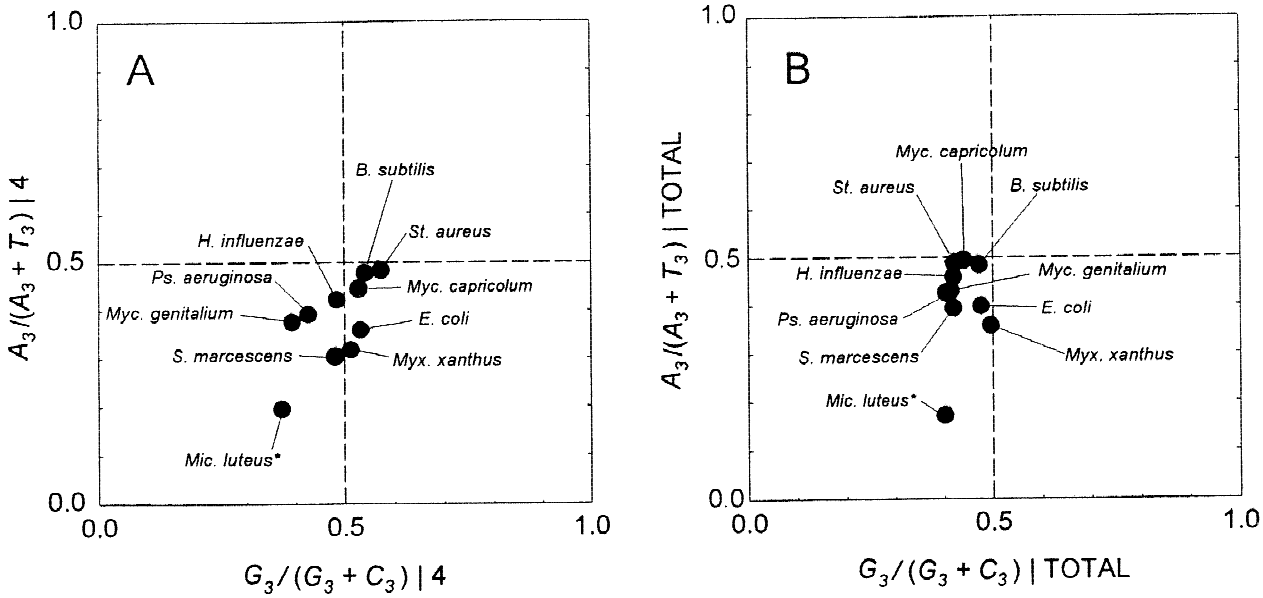

Fig. 5. Average PR2-bias plot for the 10 bacterial species. **A** The average PR2-bias plot for the third codon position of four-codon amino acids of the 10 bacterial species. **B** The average plot for the third codon position of the amino acids except the codons for methionine and tryptophan and the ATA codon of isoleucine. The average values of the 10 bacterial species are plotted. Note that there is a general tendency for $T > A$ in A and B and $C > G$ in A.

codon has a significant regression against P_3 when the usage frequency of the codon is plotted against the P_3 of individual genes (Sueoka, 1992). (d) The effect of tRNA abundance on P_3 cannot be large because the tRNA effect is amino acid specific and not strongly directional as a whole (Sueoka 1995) (Fig. 5). In addition to amino acid-specific PR2 biases, DNA replication-coupled PR2 biases were observed in some bacteria (Lobry 1996a; Mclean et al. 1998). These biases are apparently caused by mutational differences between leading and lagging strands of DNA in replication. Recently, we have examined the effect of the PR2 biases of individual genes of 10 completely sequenced bacterial genomes on G+C contents (Lobry and Sueoka, in preparation). It was found that whereas the PR2 biases are different between the two groups of genes, with the sense sequences of one group on the leading strand and those of the other group on the lagging strands, there is little difference in P_3

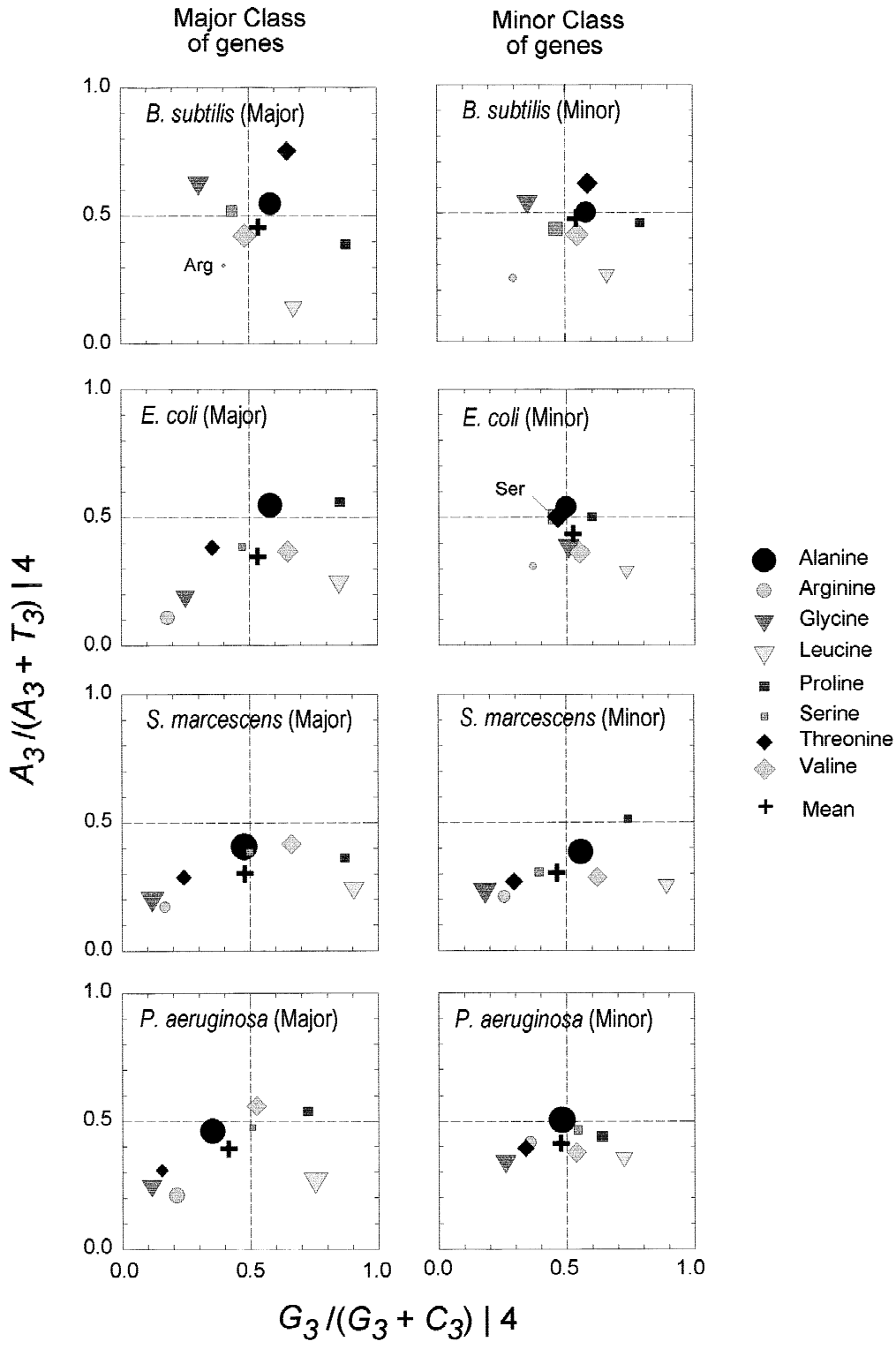

Fig. 6. PR2-bias fingerprints for major and minor class genes of Type II bacteria. The same conventions as in Fig. 3 are used. The method used to define the boundary for separating major and minor classes is described in Appendix.

between the two groups of genes. In addition, directional mutation pressure must also affect tRNA as well as codon usage frequency in the same direction ($\alpha \rightarrow \gamma$ or $\alpha \leftarrow \gamma$) (Osawa 1992). However, the mutational effect on the whole set of tRNA may be expected to be much slower than the mutational effect on changing P_3 directionally because of selection against change in tRNA. In extreme cases, the directional mutation pressure may ex-

plain the absence of tRNAs corresponding to NNA codons in *Micrococcus luteus* ($P_3 = 0.95$) and to NNG codons in *Mycoplasma capricolum* ($P_3 = 0.074$) (Osawa et al. 1992).

The potential drawback of using P_3 as the relative neutrality standard has been the possibility that codon usage bias may change GC_3 substantially by a selectional effect through tRNA abundance. However, there have been no cases to support that the tRNA abundance influences GC_3. As shown in this study, there is no correlation between PR2 bias and P_3, and it is not possible to explain the magnitude of existing P_3 variation by amino acid-specific PR2 biases. The use of P_3 for the neutrality standard assumes that P_3 is nearly neutral against directional mutation pressure in most genes and at near-equilibrium with mutation rates u and v. At equilibrium, P_3 is expected to be close to its equilibrium value, $v/(u + v)$. The assumption of near-equilibrium seems reasonable in most cases because of the asymptotic nature of the transition toward equilibrium of P_3 (Sueoka 1962, 1992). At present, for assessing the extent of neutrality of a set of nucleotides, P_3 seems to be the most appropriate available parameter to represent the neutrality standard.

For the sake of argument, one could use the G+C content of intergenic spaces (GC_{igs}) as an alternative standard scale for neutrality. The abundance of tRNA is known to bias the codon usage, presumably through selection for efficient translation (Ikemura 1981; Bennetzen and Hall 1982), and accordingly, intergenic spaces may be another logical standard for the measurement of relative neutrality. The relative neutrality of GC_{igs} to the neutrality of P_3 is 0.69 (Fig. 2). When we use the GC_{igs} of various bacteria as the neutrality standard (adscissa), the regression coefficient of P_3 vs. GC_{igs} becomes 1.45. This fact may be interpreted that P_3 has a selection positive by 45% relative to the intergenic spaces (Fig. 2). GC_{igs} may conceptually be an ideal parameter to use as the neutrality standard because of its separation from the effects of transcription and translation processes, except for the existence of regulatory elements and putative elements for chromatin structures in intergenic spaces. Despite these advantages of GC_{igs}, P_3 was adopted as the relative neutrality standard in this study because (a) the intergenic spaces are likely to contain sequence elements that have nonneutral 5′-upstream and 3′-downstream regulatory elements as well as sequences whose functions are currently unknown; (b) intergenic spaces in bacteria are generally short, probably because of selection to reduce the chromosome doubling time, which may partially eliminate neutral sequences; and (c) in eukaryotes, the use of intergenic spaces or introns for the neutrality standard presents a problem because they contain various control elements as well as a large amount of "junk" DNA (Ohno 1972), where neutrality of the G+C content is difficult to assess.

Directional Mutation vs. Directional Selection

The nature of mutational and selectional effects on P_3 is fundamentally different. The most obvious difference between mutation and selection affecting DNA base composition is the fact that mutation pressure works equally throughout the genome or, at least, within large regions (e.g., isochores). Bidirectional mutation pressures (u and v) and equilibrium of P_3 seem to explain the wide interspecific variation and narrow intragenomic heterogeneity of P_3, as seen for Type I bacteria and for major class genes of Type II bacteria (Fig. 1). A common mutation pressure can also explain the high regression coefficient of intergenic spaces to P_3 (Fig. 2). To explain these compositional features by the classical unidirectional mutation rate and unidirectional selection coefficient, widely variable thresholds had to be introduced (Li 1987). In general, selection may have effects on a specific strand and on small units such as codons for individual amino acids, except in some bacteria in which DNA replication-coupled PR2 bias is strong. For example, for selection to have directional effects on P_3, the majority of synonymous codons must be selected in a unidirectional fashion toward an AT-rich or a GC-rich composition at the third codon position in a genomewide fashion. For this process to occur first, the tRNA has to change directionally by mutation and, then, mutated tRNA that fits to preferred codons is likely to be selected. Such a process is unlikely to occur, which is supported by the conservative nature of PR2 biases (Fig. 3).

Properties likely to be involved in directional selection include the heat stability of high-G+C DNA, the rigidity of the double helix of high-G+C DNA, and decreases in both of these characteristics in high-A+T DNA. However, none of these effects has been established as a credible evolutionary factor to explain the wide interspecific variation of the DNA G+C content in bacteria. For example, no correlation was found between the optimum growth temperature and the G+C content of thermophilic bacteria (Galtier and Lobry 1997).

Intragenomic Heterogeneity of the G+C Content in Bacteria

The results shown in Fig. 1 indicate that Type II bacteria have wide intragenomic heterogeneity in P_3, which indicates that the intragenomic heterogeneity of P_3 is not confined to multicellular eukaryotes, It is not yet clear whether some common principles operate for the heterogeneity in both bacteria and higher eukaryotes. It is interesting to note that among the bacteria analyzed in this article, intragenomic G+C heterogeneity exists particularly in minor class genes of Type II bacteria whose DNA G+C contents are not extremely high or low. Five possibilities have been considered as general causes for the G+C heterogeneity: (a) *directional selection hypoth-*

esis—DNA stability at high-G+C regions or instability at high-A+T regions may provide some advantage for a high cell body temperature; (b) *multiple-source-of-mutation hypothesis*—major sources of mutation (e.g., replication errors and repair errors) with different directionality act on the entire genome; (c) chromatin structure-effect hypothesis—local chromatin structure may affect directional substitution rates; (d) *tRNA-effect hypothesis*—genes encoding abundant proteins may temporarily resist directional mutation pressure in their change of GC_3 by selection through tRNA abundance; and (e) *horizontal transfer hypothesis*—exogenous DNA with a different G+C content may integrate into the host genome.

In this article, these possibilities are discussed for bacteria, although some of the underlying principles might be shared in both prokaryotes and eukaryotes.

Directional Selection Hypothesis. The G+C content of DNA is linearly correlated with thermal stability (Marmar and Doty 1959). Therefore, bacteria with a high G+C content may have an advantage in a high-temperature environment. This hypothesis, however, has been rejected in bacteria (Galtier and Lobry 1997).

Multiple-Source-of-Mutation Hypothesis. Sueoka (1988) raised the possibility that the wide range of P_3 heterogeneity in some organisms may result from at least two sources of mutations (e.g., replication errors and repair errors) with different values of mutational bias μ_D, defined as $v/(u + v)$. The values of μ_D1 (e.g., replication) and μ_D2 (e.g., repair) may differ in different organisms, and μ_D2 may be influenced by the extent of expression in the germ line (Filipski 1987; Sueoka 1988, 1993). Under this scenario, the range of P_3 values of an organism may be determined by μ_D1 and μ_D2, both of which can be subjected to mutator mutations. The extent of transcription in the germ line in multicellular organisms and essential genes for growth in unicellular organisms may influence P_3 values and cause the spread. There has been no experimental evidence for this hypothesis.

Chromatin Structure-Effect Hypothesis. The structural constraints of regional G+C content may also be required for a highly condensed state of chromatin in nucleoids and spores. It is an unsettled question whether the structural requirement is a significant cause for generating different domains with specific G+C contents (isochores) by modulation of directional mutation pressure or by selection of different G+C for functional advantages for the genes.

tRNA-Effect Hypothesis. The contribution of tRNA abundance to the bias of codon usage among the synonymous codons has been widely quoted because the abundance of tRNA has been shown to correlate with codon usage biases in *E. coli* and *Salmonella typhimurium* (Post and Nomura 1979; Ikemura 1981), yeast (Bennetzen and Hall 1982; Ikemura 1982), and *Drosophila* (Moriyama and Powell 1997). It seems reasonable to assign selection through tRNA for violations of PR2. Selection through tRNA abundance that causes violation of PR2 may explain the codon usage biases in general. The possibility that selection through tRNA changes significantly the G+C content is unlikely because of the results of PR2 bias analyses in this article. In the human genome, regression analysis of PR2 biases vs. P_3 does not show a significant correlation (Sueoka 1995). If the directional mutation rate for α to γ suddenly increases due to a mutator mutation, the tRNA abundance that was equilibrated with the previous mutation rates may not change as quickly as the P_3 of genes for abundant proteins as proposed previously for *S. marcescens* (Sharp 1990; Shields 1990). Unless a new mutator mutation occurs, substitutions in tRNA molecules may eventually accumulate through mutation and selection so that the overall tRNA complement may become adjusted to the new values of P_3 that are close to the equilibrium determined by the mutational bias, μ_D or $v/(u + v)$.

Horizontal Transfer Hypothesis. Numerous examples of transfer of genes from one bacterial species to another are known (see the review article by Ochman and Lawrence 1996). This hypothesis is a definite possibility in some cases where the G+C content of a specific genomic region is different from neighboring regions (e.g., see the relative neutrality plot data in Fig. 1 and the PR2 bias plot in Fig. 6 for *E. coli*). However, it is doubtful that the hypothesis applies to general features of variation and heterogeneity of the DNA G+C content, namely, wide interspecies variation and narrow intragenomic heterogeneity.

Parity Violation Due to Strand-Specific Bias in Selection

Detailed analyses of PR2 bias should be helpful in separating the effects of mutation and selection in the analyses of the interspecific variation of P_3. In the analysis of evolution of the DNA nucleotide composition, the codon usage bias between synonymous codons has usually been defined as the deviation from equimolar usage of the four nucleotides. However, mutation generally occurs equally in both strands affecting the G+C content, except in some bacteria during DNA replication, where leading and lagging strands are subject to different directional mutation pressures more or less uniformly (Wu 1991; Lobry 1996a). Conversely, selection for the translation efficiency of nucleotide sequence works on the sense strand (codons) in an amino acid- and species-specific manner. The effect of directionality of DNA replication on PR2 violation is highly variable among bacterial species. The

replication-coupled mutation affects the bias from PR2 more or less uniformly throughout the genome by as much as 10% on the G+C scale between leading and lagging strands in some bacteria (Lobry 1997; Mclean et al. 1998). Consequently, for the study of directional mutation and selection on the DNA nucleotide composition, analyses of interspecific variation and intragenomic heterogeneity of PR2 biases [$A_3/(A_3 + T_3)$ and $G_3/(G_3 + C_3)$] as well as those of P_3 seem important. However, it recently became clear that DNA replication-coupled PR2 biases are found in some bacteria but not in other bacteria and that even *Borrelia burgdorferi*, with the largest PR2 biases observed so far, does not show a difference in P_3 between the genes on the leading strand and those on the lagging strand (Mclean et al. 1998; Lobry and Sueoka, in preparation).

Genes of Type I Bacteria and Major Class Genes of Type II Bacteria. Biases in the usage frequency of synonymous codons have been well documented (see Li 1997). There are different views on the factors that contribute to the biases. (a) Mutation may be mainly responsible for the DNA nucleotide composition of the third codon position (namely, A_3, T_3, G_3, and C_3). (b) Selection may be mainly responsible for the DNA base composition of the third codon position. (c) Mutation may be mainly responsible for the G+C content of the third codon position, whereas strand-specific selection and mutation are responsible for the bias from PR2 in the third codon position. This article proposes the last view for the interspecific variation of major class genes. Three plausible causes for the usage bias of synonymous codons from equal frequency are (a) the directional mutation pressure that is mainly responsible for P_3 variation among bacterial species; (b) amino acid-specific selection due to abundance, wobbling, and accuracy of codon recognition of tRNA that is responsible for violation of the PR2; and (c) strand-specific violation of PR2 by DNA replication-coupled mutation. The last cause is highly variable among bacterial species and is not ubiquitous (Lobry 1997; Lobry and Sueoka, in preparation).

Phylogenetic Conservation of PR2 Biases. The results of PR2 plotting of the four-codon amino acids indicate that the pattern of PR2 bias is a conserved characteristic in evolution and may serve as a fingerprint of different classes of organisms including bacteria and vertebrates (Sueoka 1995; this article). PR2-bias plots presented in Fig. 4 indicate a conservative nature of PR2 biases among *E. coli*, *S. marcescens*, and *P. aeruginosa* despite their differences in P_3. This feature of the PR2-bias plot, therefore, may be useful for phylogenetic analysis, providing new information based on selection most likely through tRNA abundance, DNA replication-coupled mutations, and possibly transcription.

The PR2-bias plot of *Micrococcus luteus* is distinct from those of other bacterial species (Figs. 5A and B). Unfortunately, sequence data for only a very small number of *M. luteus* genes (19 genes) are currently available. In view of the several unused codons, extreme P_3 (95%), and peculiar feature of PR2 bias of this organism, further studies of the *M. luteus* sequence should be highly instructive.

Major and Minor Class Genes of Type II Bacteria. Major class genes of Type II bacteria show wider PR2-bias patterns than minor class genes (Fig. 6), except in *S. marcescens*, where both patterns are almost equally spread. It is reasonable to expect that abundantly expressed genes have greater PR2 biases because they take advantage of the relative tRNA abundance to gain a higher efficiency of translation (Ikemura 1981). Thus, the result indicates that the major class of Type II bacteria includes both highly and moderately expressed genes. In this sense, minor class genes of *B. subtilis*, *E. coli*, and *P. aeruginosa* may comprise more moderately expressed genes than major class genes. In *S. marcescens*, on the other hand, both major and minor classes may include highly expressed genes as well as moderately expressed genes. The fact that the PR2 bias pattern of the minor class genes of *S. marcescens* is not compressed, unlike any other Type II bacteria, does not contradict the result of Sharp (1990), where codon adaptation indices (CAI) of eight homologous genes in the minor class range of GC_3 correlate negatively with low-GC_3 genes (minor class). However, the CAI analysis may not be appropriate because the CAI uses equimolarity of the four nucleotides as the null hypothesis, thus there is no way to determine the relative contributions of selection or mutation to CAI. In *B. subtilis* and *P. aeruginosa*, the compressed PR2 patterns suggest that minor class genes are expressed less abundantly than major class genes. The minor class genes of *E. coli* may be treated as an exogenous class of genes. Current analyses suggest that the cause of the G+C heterogeneity of minor class genes in bacteria is species specific, and no generalized mechanism can be proposed.

Genetic Drift and G+C Content

Genetic drift is an important concept for the study of polymorphism and the process of fixation-in-population as well as for the study of neutral or near-neutral alleles and individual nucleotide pairs (Wright 1937; Kimura 1964, 1983; Ohta 1973). The effective population size (N_e) is an important parameter for the analysis of polymorphism and gene fixation of each gene or nucleotide. However, the probability of fixation of alternative states in the population (e.g., an allele or a nucleotide at a fixed site) is likely to be a function of bidirectional substitution rates. For a near-neutral set of nucleotides, the fixation is mainly a function of mutation rates. Moreover, when a

large number of nucleotides are involved, such as average values of sets of nucleotides (P_1, P_2, P_3, GC_3, and GC_{igs}), or parameters for PR2 biases such as $A_3/(A_3 + T_3)$ and $G_3/(G_3 + C_3)$, the effect of fluctuation by random drift is expected to diminish as the number of nucleotides in the set increases. Likely main factors involved in determining these parameters are (a) nonrandom factors such as strand-independent (symmetric) directional mutation pressure for the G+C content and (b) strand-specific (asymmetric) directional selections and mutations for PR2 biases. Therefore, the importance of N_e in these averages varying beyond the expected statistical error is likely to diminish as the size of the nucleotide set (the number of nucleotides) increases. Under this directional mutation theory, the error distribution of the G+C content (P) of a set of nucleotides is binomial at equilibrium, so that the standard deviation of P is $\sqrt{P}(1 - P)/b$, where b is the number of nucleotide pairs in the set (Sueoka 1962). Here, the diffusion kinetics assumed for the calculation of allelic fixation in the population by Kimura (1964) does not apply as such to the behavior of the average parameters such as P values and other GC values. Instead, the directionality of strand-independent mutations defined by $v/(u + v)$ between α and γ nucleotides and the G+C content at equilibrium are likely to determine the DNA G+C content, whereas amino acid-specific, strand-specific selections as well as DNA replication-coupled, strand-specific mutations are likely to determine PR2 biases.

A Model for the G+C Content and PR2 Biases in Bacteria

A model based on the result of present analysis is presented schematically in Fig. 7. For the evolutionary change of the DNA G+C content, directional mutation pressure (DMP) is likely to play the major role. The major player responsible for DMP is likely to be a mutator mutation that enhances the directional mutation pressure. The ubiquitous effect of amino acid-specific PR2 biases can be measured as the deviation from the center point of the PR2 plot ($A = T$ and $G = C$), which is shown as the vertical double-headed arrow in Fig. 7. Currently, there is no direct way to measure the effect of codon usage biases on P_3. If we assume that the selectional effect due to codon usage biases affects P_3 to the same extent as the overall PR2 biases, the effect is less than 15% on the G+C scale from P_3 in both the plus and the minus directions (the horizontal, small, dotted, double-headed arrow in Fig. 7). This plus–minus feature is likely to make the effect even smaller.

Conclusions

Mutation and selection influence two different phases of the DNA nucleotide composition in bacteria: one is the G+C content and the other is the violation of PR2. The present analysis indicates that the major factor for the interspecific variation of the G+C content of the third codon position is strand-independent directional mutation pressure, whereas the PR2 biases comes mainly from the ubiquitous, amino acid- and strand-specific, translation-coupled selection and, in some cases, DNA replication-coupled mutation. This model (Fig. 7) may be applicable to the genes of Type I bacteria and the major class genes of Type II bacteria. Distinct mechanisms in different species of Type II bacteria are likely to be responsible for the intragenomic G+C heterogeneity of minor class genes.

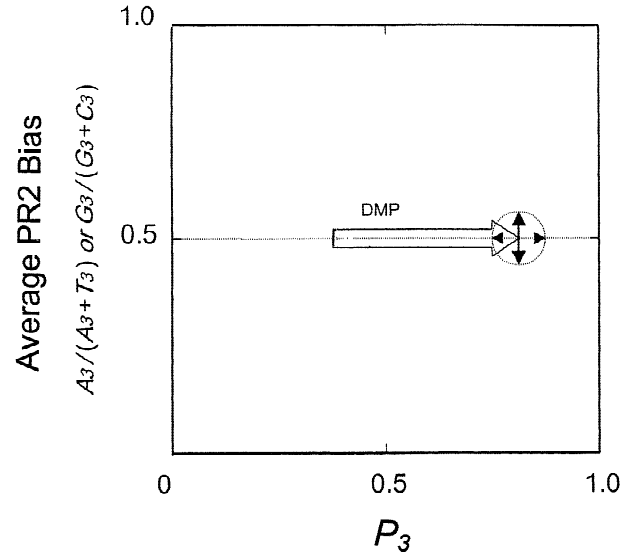

Fig. 7. Schematic diagram of the change in G+C content due to directional mutation pressure and PR2 bias. The *large open arrow* represents the effect of directional mutation pressure (DMP). The *vertical, double-headed filled arrow* represents the PR2 bias at the third codon position, which is measurable from the frequency of codon usage. The *horizontal, double-headed dotted arrow* represents the hypothetical contribution of PR2 bias to P_3.

Acknowledgments. The author is grateful to Drs. Joel Heilig and Taminko Kano-Sueoka for improving the manuscript. This work was supported by NSF Grant DIR8820806.

References

Aota S, Ikemura T (1986) Diversity in G+C content at the third codon position of codons in vertebrate genes and its cause. Nucleic Acids Res 14:6345–6355

Belozersky AN, Spirin AS (1958) A correlation between the compositions of deoxyribonucleic and ribonucleic acids. Nature 182:111–112

Bennetzen JL, Hall B (1982) Codon selection in yeast. J Biol Chem 257:3026–3031

Bernardi G, Olfsson B, Filipski J, et al. (1985) The mosaic genome of the vertebrates. Science 228:953–958

Bonamy C, Labarre J, Reyes O, Leblon G (1994) Identification of IS1206, a *Corynebacterium glutamicum* IS3-related insertion sequence and phylogenetic analysis. Mol Microbiol 14:571–581

Braun G, Cole ST (1984) DNA sequence analysis of the Serratia marcescens ompA gene: Implications for the organization of an enterobacterial outer membrane protein. Mol Gen Genet 195:321–328

Cox EC, Yanofsky C (1967) Altered base ratios in the DNA of an Escherichia coli mutator strain. Proc Natl Acad Sci USA 58:1895–1902

Cox EC, Yanofsky C (1969) Mutator gene studies in *Escherichia coli*. J Bacteriol 100:390–397

Filipski J (1987) Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in the germline cells. FEBS Lett 217:184–186

Fraser CM, et al. (1995) The minimal gene complement of *Mycoplasma genitalium.* Science 270:397–403

Furusawa M, Doi H (1992) Promotion of evolution: Disparity in the frequency of strand-specific misreading between the lagging and leading DNA strands enhances disproportionate accumulation of mutations. Proc Natl Acad Sci USA 157:127–133

Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structure and optimal growth temperature in prokaryotes. J Mol Evol 44:632–636

Gouy M, Gautier C (1982) Codon usage in bacteria: Correlation with gene expressivity. Nucleic Acids Res 10:7055–7074

Grantham R, Gautier C, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acid Res 8:r49–r62

Goutierrez G, Casadesus J, Oliver JL, Marin A (1994) Compositional heterogeneity of the *Escherichia coli* genome: A role for VSP repair? J Mol Evol 39:340–346

Ikemura T (1981) Correlation between the abundance of *E. coli* t-RNA and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151:389–404

Ikemura T (1982) Correlation between the abundance of yeast transfer RANs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoccepting tRNAs. J Mol Biol 158:573–597

Ikemura T (1985a) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Ikemura T (1985b) Codon usage, tRNA content, and rate of synonymous substitution. In: Ohta T, Aoki K (eds) Population genetics and molecular evolution, Japan Sci Soc Press, Tokyo/Springer-Verlag, pp 385–406

Kano A, Adachi Y, Ohama T, Osawa S (1991) Novel anticodon composition of transfer RNAs in *Micrococcus luteus*, a bacterium with a high genomic G+C-content: Correlation with codon usage. J Mol Biol 221:387–401

Kimura M (1964) Diffusion models in population genetics. Jap Appl Probab 1:177–232

Kimura M (1983) The neutral theory of molecular evolution. Cambridge, University Press Cambridge

Kunst K, et al. (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis.* Nature 390:249–256

Lee KY, Wahl R, Barbu E (1956) Contenu en bases puriques et pyrimidiques des acids desoxyribonucleiques des bacteries. Ann Inst Pasteur 91:212–224

Li WH (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol 24:337–345

Li WH (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Lloyd AT, Sharp PM (1992) Evolution of codon usage patterns: The extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. Nucleic Acids Res 20:5289–5295

Lobry JR (1995) Properties of a general model of DNA evolution under no-strand-bias conditions. J Mol Evol 40:326–330

Lobry JR (1996a) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660–665

Lobry JR (1996b) Origin of replication of *Mycoplasma genitalium.* Science 272:745–746

Lobry JR (1997) Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205: 309–316

Lobry JR, Gautier C (1994) Hydrophobicity, expressibility, and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acid Res 22: 3174–3180

Marmur J, Doty P (1959) Heterogeneity in deoxyribonucleic acids. I. Dependence on composition of the configurational stability of deoxyribonucleic acids. Nature 183:1427–1429

Mclean MJ, Wolfe KH, Divine KM (1998) base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Biol 47:691–696

Medique C, Rouxel T, Vigier P, Henaut A, Danchin A (1991) Evidence for horizontal gene transfer in Esherichia coli speciation. J Mol Biol 222:851–856

Moriyama EN, Powell JR (1997) Codon usage bias and tRNA abundance in *Drosophila.* J Mol Evol 45:514–523

Mouchiroud D, D'Onofrio G, Aissani B, Macaya G, Gautier C, Bernardi G (1991) The distribution of genes in the human genome. Gene 100:181–187

Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84:1166–1169

Nakamura K, Inouye M (1980) DNA sequence of the Serratia marcescens lipoprotein gene. Proc Natl Acad Sci USA 77:1369–1373

Nichols BP, Miozzari GF, van Cleemput M, Bennett GN, Yanofsky C (1980) Nucleotide sequences of the *trpG* regions of *Escherichia coli, Shigella dysenteriae, Salmonella typhimurium* and *Serratia marcescens.* J Mol Biol 142:503–517

Nomura M, Sor F, Yamagishi M, Lawson M (1987) Heterogeneity of GC content within a single bacterial genome and its implications for evolution. Cold Spring Harbor Symp Quant Biol 52:658–663

Ochman H, Lawrence JG (1996) Phylogenetics and the amelioration of bacterial genomes. In: *Escherichia coli* and *Salmonella* Neighthard FC (ed) ASM Press, Washington DC, pp 2627–2637

Ogasawara N (1985) Markedly unbiased codon usage in *Bacillus subtilis.* Gene 40:145–150

Ohama T, Muto A, Osawa S (1990) Role of GC-biased mutation pressure on synonymous codon choice in *Micrococcus luteus*, a bacterium with a high genomic GC-content. Nucleic Acid Res 18:1565–1569

Ohno S (1972) So much "junk" DNA in our genome. In: Smith HH (ed) Brookhaven Symp Biol 23:366–379

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246:96–98

Osawa S (1995) Evolution of the genetic code. Oxford University Press, Oxford

Osawa S, Jukes TH, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. Microbiol Rev 56:229–264

Post LE, Nomura M (1980) DNA sequences from the *str* operon of *Escherichia coli*. J Biol Chem 255:4660–4666

Rolfe R, Messelson M (1959) The relative homogeneity of microbial DNA. Proc Natl Acad Sci USA 45:1039–1043

Schildkraut CL, Marmur J, Doty P (1962) The formation of hybrid DNA molecules and their use in studies of DNA homologies. J Mol Biol 4:430–443

Sharp PM (1990) Processes of genome evolution reflected by base frequency differences among Serratia marcescens genes. Mol Microbiol 4:119–122

Sharp PM, Devine KM (1989) Codon usage and gene expression level in *Dictyostelium discoideum:* Highly expressed genes do "prefer" optimal codons. Nucleic Acids Res 17:5029–5038

Sharp PM, Li WH (1987a) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230

Sharp PM, Li WH (1987b) The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Sharp PM, Li WH (1989) On the rate of DNA sequence evolution in Drosophila. J Mol Evol 28:398–402

Shields DC (1990) Switches in species-specific codon preferences: The influence of mutation biases. J Mol Evol 31:71–80

Shields DC, Sharp PM (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutation biases. Nucleic Acids Res 15:8023–8040

Shields DC, Sharp PM, Higgins DG, Right F (1988) "Silent" sites in Drosophila genes are not neutral: Evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Sueoka N (1961a) Variation and heterogeneity of base composition of deoxyribonucleic acids: A compilation of old and new data. J Mol Biol 3:31–40

Sueoka N (1961b) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. Proc Natl Acad Sci USA 47:1141–1149

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582–592

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657

Sueoka N (1992) Directional mutation pressure, selective constraints, and genetic equilibria. J Mol Evol 34:2653–2657

Sueoka N (1993) Directional mutation pressure, mutator mutations, and dynamics of molecular evolution. J Mol Evol 37:137–153

Sueoka N (1995) Intra-strand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325; *Erratum* (1996) J Mol Evol 42:323

Sueoka N, Marmur J, Doty P (1959) Heterogeneity in deoxyribonucleic acids. II. Dependence of the density of deoxyribonucleic acids on guanine-cytosine. Nature 183:1427–1431

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Wright S (1937) The distribution of gene frequencies in populations. Proc Natl Acad Sci USA 23:307–320

Wu, C-I (1991) DNA strand asymmetry. Nature 352:114

Wu C-I, Maeda N (1987) Inequality in mutation rates of the two strands of DNA. Nature 327:169–170

Appendix

Mycoplasma capricolum—Thirty-one genes having 100 or more codons were analyzed. The mean P_3 is 0.074, and the P_3 distribution around the mean is narrow (Type I). The tRNA complement allows all possible uses of synonymous codons (Kano et al. 1991), although the P_3 value is extremely low.

Staphylococcus aureus—One hundred eighteen genes having 100 or more codons were analyzed. The mean P_3 is 0.194. The P_3 distribution around the mean is narrow (Type I).

Mycoplasma genitalium—All 194 genes having 300 or more codons were analyzed using the complete genome sequence (Fraser et al. 1995). The mean P_3 is 0.226. The P_3 distribution around the mean is narrow (Type I). Interestingly, although *M. genitalium* has been reported to be related to *B. subtilis* by sequence identity (see Kunst et al. 1997), the PR2 bias plots of the two bacteria do not show similarity. Analyzing PR2 bias along the DNA sequence map, the location of the replication origin has been proposed (Lobry 1996b).

Haemophylus influenzae—Five hundred two genes of 100 or more codons were analyzed. The mean P_3 is 0.262. The P_3 distribution around the mean is narrow (Type I).

Bacillus subtilis—Two hundred eighty-six genes having 100 or more codons were analyzed. The P_3 distribution around the mean is moderately wide (Type II). The mean P_3 of major class genes is 0.428. In the set of genes presented here, minor class genes comprise approximately 16% (the chosen class boundary of P_3 is 0.34), which most likely corresponds to Class 3 genes ("high-AT genes") of the complete genome (13%) by a correspondence analysis (Kunst et al. 1997).

Escherichia coli—One thousand two hundred ninety genes having 100 or more codons were analyzed. The P_3 distribution around the mean is moderately wide (Type II). The mean P_3 of major class genes is 0.565. In this set of genes, minor class genes include approximately 11% (the chosen class boundary of P_3 is 0.40). The distribution of minor class genes deviates from the slope that is similar among interspecific as well as intragenomic distributions as shown in Fig. 1. Minor class genes consist of a set of unique genes that include genes of pili, exotoxins, endotoxins, and some membrane proteins. A very similar feature is found in *Salmonella typhimurium* and *Vibrio chorelae* (data not shown). Minor class genes correspond to the low-expressivity class in a correspondence analysis of *E. coli* by Gouy and Gautier (1982) and Class 3 by Medique et al. (1991). Membrane protein genes and cytoplasmic protein genes have been differentiated by correspondence analysis (Lobry and Gautier 1994). The possibility of horizontal transfer of these genes between enteric bacteria has been suggested by Medique et al. (1991), discussed by Gutierrez et al. (1994) and Bonamy et al. (1994), and reviewed by Ochman et al. (1996).

Serratia marcescens—Eighty-three genes having 100 or more codons were analyzed. The P_3 distribution around the mean is wide (Type II). The mean P_3 of major class genes is 0.797. Minor class genes comprise approximately 29% of the analyzed gene set (the chosen class boundary of P_3 is 0.60). The intragenomic heterogeneity of this species was noted and analyzed previously. The high GC_3 of the *trp*(*G*)*D* gene expected from the average DNA base composition of *S. marcescens* was first noted by Nichols et al. (1980). Similar situations were reported in the *Ipp* gene (Nakamura and Inouye 1980) and the *ompA* gene (Braun and Cole 1984). In comparison with the homologous gene in *E. coli*, the *trp*(*G*)*D* gene of *S. marcescens* had more substitutions toward GC than toward AT. Using a two-dimension chromatographic separation of tRNAs, Ikemura (1985b) observed that the quantitative tRNA fingerprint of *S. marcescens* resembles that of *E. coli*. Nomura et al. (1987) noted that the small ribosomal protein genes, *rplK* and *rplA*, of *S. marcescens* had GC_3 values similar to those of *E. coli* homologues. They interpreted the difference between the two types of genes [*trp*(*G*)*D* vs. *rplK–rplA, Ipp,* and *ompA*] as the result of a unique difference in directional mutation pressure in different domains in the chromosome. Sharp (1990) explained the heterogeneity of the base composition of the third codon position in *S. marcescens* by both selection and mutation. There was a steep negative correlation ($r = -0.83$, $p < 0.002$) between the difference in G+C content (*S. marcescens–E. coli*) and the CAI [codon adaptation index (Sharp and Li 1987a)] of *E. coli* in eight homologous genes (Sharp 1990). Shields (1990) pointed out that the heterogeneity of GC_3 in *S. marcescens* may stem from the fact that the third codon letters of abundant protein genes may resist directional mutation pressure by selection through tRNA abundance.

Pseudomonoas aeruginosa—Two hundred fifty-four genes having 100 or more codons were analyzed. The mean of major class genes is 0.879, and the P_3 distribution of minor class genes is wide (Type II). Minor class genes in Fig. 1 comprise 28% (the chosen class boundary of P_3 is 0.75).

Myxococcus xanthus—Fifty-six genes having 100 or more codons were analyzed. The mean P_3 is 0.906, and the P_3 distribution is narrow (Type I).

Micrococcus luteus—Nineteen genes having 100 or more codons were analyzed. The mean P_3 is 0.950, and the P_3 distribution is narrow (Type I). A unique feature of this high-G+C bacterium is selective usage of tRNAs whose anticodons are efficient in using high-G+C codons at the third codon position (Ohama et al. 1990). For a detailed description, see Osawa (1995).