ELSEVIER

Gene 261 (2000) 53–62

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

DNA G + C content of the third codon position and codon usage biases of human genes

Noboru Sueoka[a,*], Yuichi Kawanishi[b]

[a]*University of Colorado, Department of Molecular, Cellular, and Developmental Biology, Boulder, CO 80309-0347, USA*
[b]*Computer Chemistry Systems Department, Science Systems Division, Fujitsu Limited and Center for Information Biology, National Institute of Genetics, Mishima, Japan*

Received 21 June 2000; received in revised form 28 September 2000; accepted 5 October 2000
Received by G. Bernardi

Abstract

The human genome, as in other eukaryotes, has a wide heterogeneity in the DNA base composition. The evolutionary basis for this heterogeneity has been unknown. A previous study of the human genome (846 genes analyzed) has shown that, in the major range of the G + C content in the third codon position (0.25–0.75), biases from the Parity Rule 2 (PR2) among the synonymous codons of the four-codon amino acids are similar except in the highest G + C range (Sueoka, N., 1999. Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G + C content of third codon position. Gene 238, 53–58.). PR2 is an intra-strand rule where $A = T$ and $G = C$ are expected when there are no biases between the two complementary strands of DNA in mutation and selection rates (substitution rates). In this study, 14,026 human genes were analyzed. In addition, the third codon positions of two-codon amino acids were analyzed. New results show the following: (a) The G + C contents of the third codon position of human genes are scattered in the G + C range of 0.22–0.96 in the third codon position. (b) The PR2 biases are similar in the range of 0.25–0.75, whereas, in the high G + C range (0.75–0.96; 13% of the genes), the PR2-bias fingerprints are different from those of the major range. (c) Unlike the PR2 biases, the G + C contents of the third codon position for both four-codon and two-codon amino acids are all correlated almost perfectly with the G + C content of the third codon position over the total G + C ranges. These results support the notion that the directional mutation pressure, rather than the directional selection pressure, is mainly responsible for the heterogeneity of the G + C content of the third codon position.

Keywords: Directional mutation and selection pressures; DNA G + C content; Biases from Parity Rule 2; Human genes

1. Introduction

In bacteria, the intra-genomic heterogeneity of the G + C content is small, whereas intergenic variation of the G + C content is extremely large (covering 0.25–0.75) (Rolfe and Meselson, 1959; Sueoka et al., 1959). This feature of bacterial DNA base composition was interpreted to mean that the bidirectional mutation pressures between G/C and A/T pairs are the major source of variation (Sueoka, 1962). The variation is even larger (0.05–0.97) when the range of the third codon positions is compared (Muto and Osawa, 1987; Sueoka, 1988). In higher eukaryotes, intra-genomic heterogeneity in higher eukaryotes is generally large and inter-specific variation of the average G + C content is small (Sueoka, 1964). Later, more importantly, Bernardi et al. (1985) revealed discontinuous regional differences in the DNA G + C content within each chromosome (isochores) in mammalian genomes. The isochore evolution has been explained by functional selections for the adaptive values of G + C content of DNA and by structural stability at higher temperature of corresponding RNA and proteins (Bernardi, 2000).

The Parity Rule 2 (PR2) expects the intra-strand frequency relationships, $A = T$ and $G = C$. The violation of intra-strand rule, Parity Rule 2 (PR2), is a ubiquitous compositional feature of DNA (Sueoka, 1995; Lobry, 1995). (Here, A, T, G, and C represent fractional frequencies of the corresponding nucleotides.) This rule has been experimentally discovered first by Chargaff and his collaborators in long stretches of contiguous single strands of *Bacillus subtilis* DNA (Karkas et al., 1968; Rudner et al.,

Abbreviations: P_1, P_2 and P_3, symmetric G + C content of the first, second and third codon positions; P_{12}, average of P_1 and P_2; PR2, Parity Rule 2; AT-bias, intra-strand bias from $A = T$; GC-bias, intra-strand bias from $G = C$

* Corresponding author. Tel.: +1-303-492-8244; fax: +1-303-492-0388.
E-mail address: sueoka@stripe.colorado.edu (N. Sueoka).

0378-1119/00/$ - see front matter
PII: S0378-1119(00)00480-7

1968; Rudner et al., 1969). This rule was logically proven, in the case where mutation and selection are equally effective (or random) in both strands of DNA (Sueoka, 1995; also see Wu and Maeda, 1987; Wu, 1991; Furusawa and Doi, 1992; Wada et al., 1993; Lobry, 1995). Currently, three types of biases from PR2 have been recognized. One is the translation-coupled biases that are species- and amino acid-specific (Sueoka, 1995). The translation-coupled PR2 biases are most likely due to selection through tRNA abundance as proposed by Ikemura (1981) who explains the correlation between biases of synonymous codon usage and relative frequencies of tRNA. The second is the DNA replication-coupled biases (Lobry, 1996; Grigoriev et al., 1998; Murazek and Karlin, 1998; McInerney, 1998; Mackievicz et al., 1999; Morton, 1999), and the third is transcription-coupled biases (Francino et al., 1996; McInerney, 1998; Kano-Sueoka et al., 1999). None of the three types of PR2 biases is, however, able to explain either inter-specific variation of the G + C content of prokaryotic DNA nor the intra-genomic heterogeneity of the G + C content that is found in higher eukaryotes.

The objective of this work is to help understand the major factor (mutation or selection pressures) that is critical for the evolution of the intra-genomic heterogeneity of the mammalian genome. In a previous study of this project, 846 unique human genes were divided into six groups according to their G + C content of the third-codon position (the G + C content: 0.2–0.4, 100 genes; 0.4–0.5, 137 genes; 0.5–0.6, 141 genes; 0.6–0.7, 155 genes; 0.7–0.8, 210 genes; 0.8–1.0, 103 genes) (Sueoka, 1999a). The bias from PR2 was then examined separately for each group. The results showed: (a) In the major contiguous region (the G + C content of the third codon position, 0.2–0.8), no fundamental differences in the PR2 biases were found for four-codon amino acids. This result eliminates the effect of translation-coupled PR2 biases on G + C contents. (b) The highest G + C group (0.8–1.0) showed a conspicuous change in four-codon amino acid fingerprint (PR2-bias plot, see Section 2.4) and suggested an abrupt change in the mode of synonymous codon usage in this region.

Using 14,026 gene sequences, the present study has confirmed the characteristic shift of PR2 biases in the high G + C region (0.75–0.95) and further characterized the anomalous nucleotide compositional feature that is unique in the high G + C ranges.

2. Material and methods

2.1. Parameters of human genes

The codon usage table of 18,509 human genes were extracted through DDBJ at the National Institute of Genetics, Mishima, Japan by Y.K. Redundant genes were eliminated to the remaining 14,027 genes by N.S. using the following procedure: (a) More than one entry of the same size with same nucleotide frequencies in all three codon positions were discarded. (b) Among the remaining entries with same number of codons, entries with only one pair of difference (e.g. A is less by one but G is more by one within any one of the three codon positions) were also removed. The numerical analyses and plotting of the data were carried out by N.S. using MS Excel 2000 and SigmaPlot 2000.

2.2. Calculation of P_1, P_2, P_{12} and P_3

In the genetic code, the third codon position of synonymous codons includes A/T (W) and G/C (S) nucleotides in equal number (symmetric) for most amino acids except for tryptophan (TGG), methionine (ATG) and one (ATA) of the three isoleucine codons. P_3 is a measure of the G + C content of the third codon position of individual genes, and is defined as the G + C content of the third codon position for total codons, from which ATG, TGG, ATA, and the termination codons (TAA, TAG, or TGA) have been removed. As in the previous report (Sueoka, 1999a), these six codons were also removed from the calculation of the G + C contents of the first codon position (P_1) and the second codon position (P_2). P_{12} is the average of P_1 and P_2. Removal of these six codons from analysis eliminates odd numbered synonymous codon sets and, therefore, avoids an extra cause of the potential bias from PR2. In practice, the parameter (P_3) is only slightly but significantly different from the G + C content of the third codon position (GC_3).

2.3. P against P_3

In the plot of P (P_1, P_2, and P_{12}) against P_3, each point represents a gene (Fig. 1). If a P-value were as neutral as P_3 against selection, data points should have distributed along the diagonal line. Thus, the slope smaller than 1 indicates that the extent of neutrality is less than that of P_3, and, accordingly, the slope provides a measure of relative neutrality of P to that of P_3. The rationale to use P_3 as the neutrality standard, where the neutrality of P_3 is assumed to be 1, has been discussed in detail (Sueoka, 1988, 1999b). The merit of this plotting is three folds: (a) P_3-P plot shows intra-genomic or inter-species distribution of a P-value as well as P_3. (b) Regression coefficient (slope) provides an estimate of the degree of neutrality of a P relative to P_3. (c) The cross point (OP) of the regression line of a P with diagonal line represents the G + C content that is optimal for the P (Sueoka, 1988). In this study, P_3 ranges were divided by every 0.05 of P_3 values. The highest P_3 range (0.95–1.00) has only one gene (GenBank Y13436, human *sox*1 gene, $P_3 = 0.960$) (Fig. 1). The *sox*1 gene is not included in the analyses of this article because it is the only gene in the highest P_3 range (0.95–1.00).

2.4. PR2-bias plot

The PR2-bias is dectected by the value of AT-bias [$A/(A + T)$] as the ordinate and GC-bias [$G/(G + C)$] as the

abscissa (Sueoka, 1995). In this plot, the center of the plot, where both coordinates are 0.5, is the place where $A = T$ and $G = C$ (PR2). A vector from the center represents the extent and direction of biases from PR2. PR2 bias plots are particularly informative when PR2 biases at the third codon position of the four codon amino acids of individual genes are plotted. In this case, '$A_3/(A_3 + T_3)$ 4' and '$G_3/(G_3 + C_3)$ 4' are plotted as the ordinate and abscissa, respectively. Here, '4' denotes the four-codon amino acids, and A_3, T_3, G_3 and C_3 are fractions of the corresponding nucleotides at the third codon position, where $A_3 + T_3 + G_3 + C_3 = 1$. The four-codon amino acids are alanine, arginine4 (CGA, CGT, CGG, CGC), glycine, leucine4 (CTA, CTT, CTG, CTC), proline, serine4 (TCA, TCT, TCG, TCC), threonine, and valine. In some cases (Fig. 2A), PR2 biases at the third codon position are also presented for all symmetric synonymous codons, including two- as well as four-codon amino acids. In this case, the ordinate and abscissa are shown as '$A_3/(A_3 + T_3)$ 2&4' and '$G_3/(G_3 + C_3)$ 2&4,' respectively.

2.5. Extreme P_3 ranges having anomalous PR2-bias fingerprints

The relative frequencies of normal vs. anomalous genes were calculated by classifying individual genes into two classes by the relative x-coordinates of glycine and alanine bubbles; one class (normal class) having bubbles at larger x-coordinates in glycine than those in alanine and the other class (anomalous class) having opposite relations (for detail, see Section 3.3).

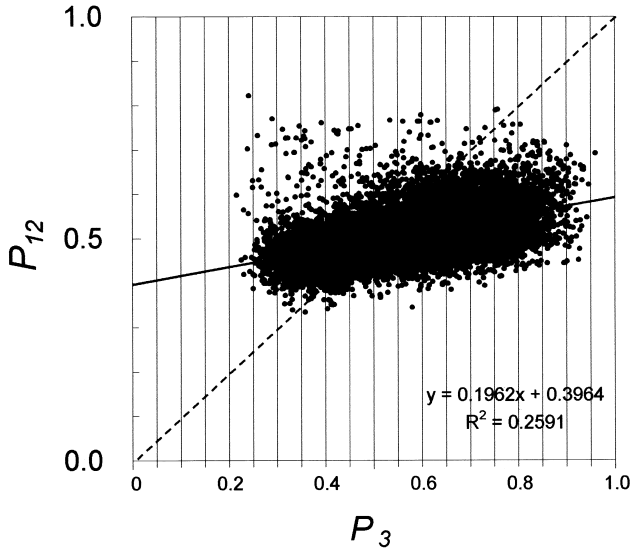

Fig. 1. P_{12} plots against P_3. 14,026 human genes were used for analysis. The solid line represents regression line. The regression equation where x represents P_3 and y represents P_{12}, includes the slope (regression coefficient) and y-intercept at $x = 0$. Squared correlation coefficient (R^2) are also shown. The OP value (the cross points of the regression lines and the diagonal lines) is 0.49 for P_{12}.

3. Results

3.1. P_{12} vs. P_3 and PR2 biases at the third codon position of human genes

In Fig. 1, 14,026 human genes having more than 100 codons were plotted for P_{12} (average of P_1 and P_2) against

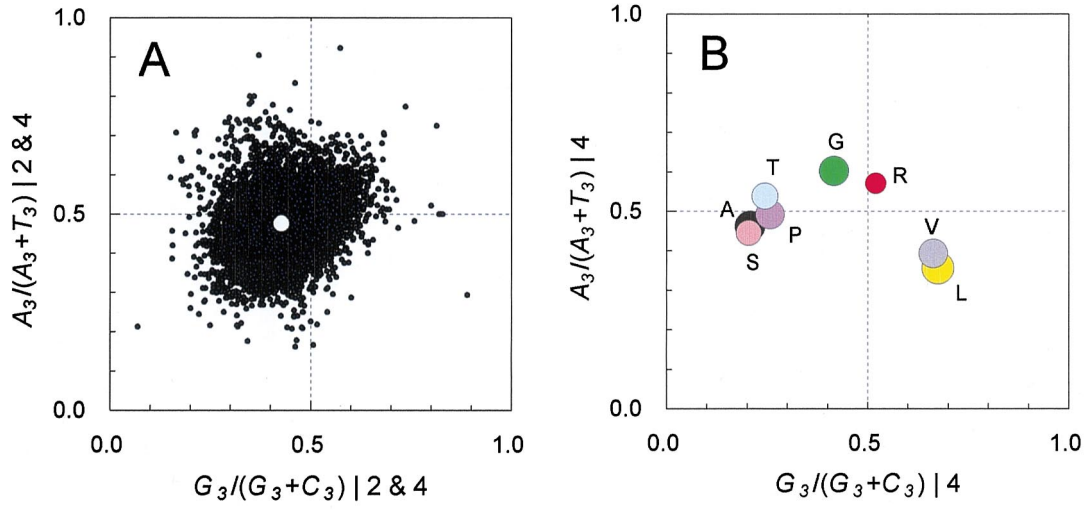

Fig. 2. PR2-bias plots (A) and PR2-bisd bubble plot of the average four-codon amino acids (B) of 14,026 human genes. (A) PR2-bias plot of individual genes where the third codon position of two- and four-codon amino acids are combined. The ordinate, '$A_3/(A_3 + T_3)$| 2 & 4' and the abscissa, ' $G_3/(G_3 + C_3)$| 2 & 4' represent A- and T-nucleotides and G- and C-nucleotides of the third codon position, respectively, which were calculated by combining the codons of both two- and four-codon amino acids. The average position [AT-bias: 0.455 ± 0.071 (SD)] and [GC-bias: 0.475 ± 0.062 (SD)] is shown as an empty circle. (B) PR2 bias-fingerprints. The average values of PR2-biases of the third codon position are plotted for individual four-codon amino acids as indicated by '| 4.' The size (diameter) of each data-point symbol shows a relative frequency of each amino acid. The color coding of bubbles are same as in Fig. 3. The pattern is used as a fingerprint of PR2 biases for comparative purposes.

P_3. P_3's of individual genes cover the G + C range of 0.22–0.96. The regression coefficient of P_{12} to P_3 is 0.196 ± 0.011, indicating that a relative neutrality is 20% or a relative constraint is 80% to P_3 (100% neutrality or 0% constraint). These values are within a typical range for a wide variety of organisms (Sueoka, 1988, 1992). P_3 ranges are shown with vertical lines at every 0.05 of the P_3 value. The P_3 range covers from 0.2 to 1.0.

Fig. 2A presents the PR2-bias plot for the third codon position where TGG, ATG, ATA and stop codons have been excluded (Section 2.2). PR2 biases of individual genes were calculated from their average A_3, T_3, G_3 and C_3 values including two-codon as well as four-codon amino acids. Fig. 2B shows the PR2 bias plot for the four-codon amino acids. In this plot, there are eight symbols representing the PR2-bias values of the eight four-codon amino acids. The PR2 bias pattern acts as a fingerprint of PR2-bias that is unique to taxa reflecting phylogenetic relationships (Sueoka, 1995).

3.2. PR2 biases for genes in different P_3 ranges

Fig. 3 shows the PR2-bias fingerprints of the individual P_3 ranges. The number of genes in each P_3 range is shown in parenthesis. It is noted that the fingerprints of P_3 ranges from 0.25 to 0.70 are similar, whereas the both extreme ranges show clearly different fingerprints. Since the lowest P_3 range (0.20–0.25) registers only 12 genes, no further analysis of this range will be made at this point. In the range of 0.70–0.95, the alteration of the fingerprints is gradual, suggesting that, in addition to the normal fingerprint genes, a new class of genes with a different fingerprint are overlapped in increasing proportions as P_3 increases, or alternatively that all genes uniformly changes the PR2 bias pattern as P_3 increases. In this article, the former possibility will be explored below.

3.3. Anomalous genes in the high G + C ranges

The anomaly of PR2 bias fingerprints in the high P_3 ranges covering 0.75–0.95 (1883 genes or 13%) is more extensive than that of the lowest P_3 range. An exact estimation of the number of the anomalous high-P_3 genes is not simple because of their obvious overlapping with the major type. A key to distinguish the anomalous from the major type is the positional changes of several amino acids in PR2-bias plots (Fig. 3). To estimate the relative numbers of anomalous and the major-type genes, the average abscissa values of glycine and alanine bubbles in individual P_3 ranges were measured. The glycine bubble moves gradually toward smaller $G_3/(G_3 + C_3)$ values in high P_3 ranges (0.70–0.95), whereas the alanine-bubble moves toward opposite direction. The average PR2 GC-biases of glycine and alanine, $G_3/(G_3 + C_3)_{Gly}$ and $G_3/(G_3 + C_3)_{Ala}$, for the genes in the individual P_3 ranges were used for calculation of the relative frequency of two types of genes. The difference (Δ) between the two values of PR2 GC-biases (Gly minus Ala) was calculated as $\Delta = [G_3/(G_3 + C_3)_{Gly} - G_3/(G_3 + C_3)_{Ala}]$. Subsequently, genes in each range were split into two classes; one class for those with positive Δ's and the other for those with negative Δ's. The two classes of genes in each P_3 range with positive and negative Δ are presented separately as frequencies of genes (Fig. 4A) and as relative frequencies (Fig. 4B). Some background negative values of Δ exist in all P_3 ranges including the major range of P_3 (0.35–0.70). The fractional average of anomalous genes with negative Δ in the major range is 0.103 representing the background values and shown as the horizontal broken line in Fig. 4B. These background negative Δ's are likely to be generated from statistical errors of PR2 biases due to small numbers of amino acid per gene in some genes. These values were, therefore, subtracted to estimate frequencies of anomalous type genes. The result shows that the estimated number of anomalous genes above background in the high P_3 range is 604 genes or 4.3% of the total genes analyzed.

3.4. Two-codon analyses

To examine whether or not deviations from PR2 exist between the two types of two-codon amino acids (CT-type and GA-type), $(G_3 + C_3) \mid 4$ and $(G_3 + C_3) \mid 2$ were plotted against P_3 (Fig. 5A,B). The result shows that, in both cases, only small deviations from PR2 (diagonal line). Here, when PR2 (an intra-strand rule, $A = T$ and $G = C$) holds, $G/(G + A) = C/(C + T)$ should also be true because $G + A = C + T$ is a corollary of $A = T$ and $G = C$.

Fig. 6 presents plots similar to those of Fig. 5 but for individual amino acids of TC-type and GA-type. The relative content of C_3 for the CT-type amino acids $[C_3/(C_3 + T_3)]$ as well as G_3 content for the GA-type amino acids $[G_3/(G_3 + A_3)]$ show that both types are almost linearly correlated with the slope close to 1 against P_3. It is also noted that the regressions of GA-type for glutamine and leucine2 to P_3 are uniformly higher than the diagonal line, indicating these two amino acids prefer the G_3-codons to A_3-codons throughout the P_3 ranges except in the highest G + C regions.

4. Discussion

4.1. PR2-biases vs. P_3

The result of Fig. 2 confirms the previous result obtained with a smaller sample (846 genes) and shows that PR2-bias fingerprints are virtually identical over a wide range of P_3 (0.25–0.75). The result supports the conclusion that the large intra-genomic heterogeneity of P_3 is not generated by amino acid-specific, translation-coupled selection in human. Here, the amino acid-specific PR2 biases, but not the G + C content, have been interpreted as the result of selection through tRNA (Sueoka, 1995, 1999a,b). The fingerprint analysis of PR2 bias reveals detailed features

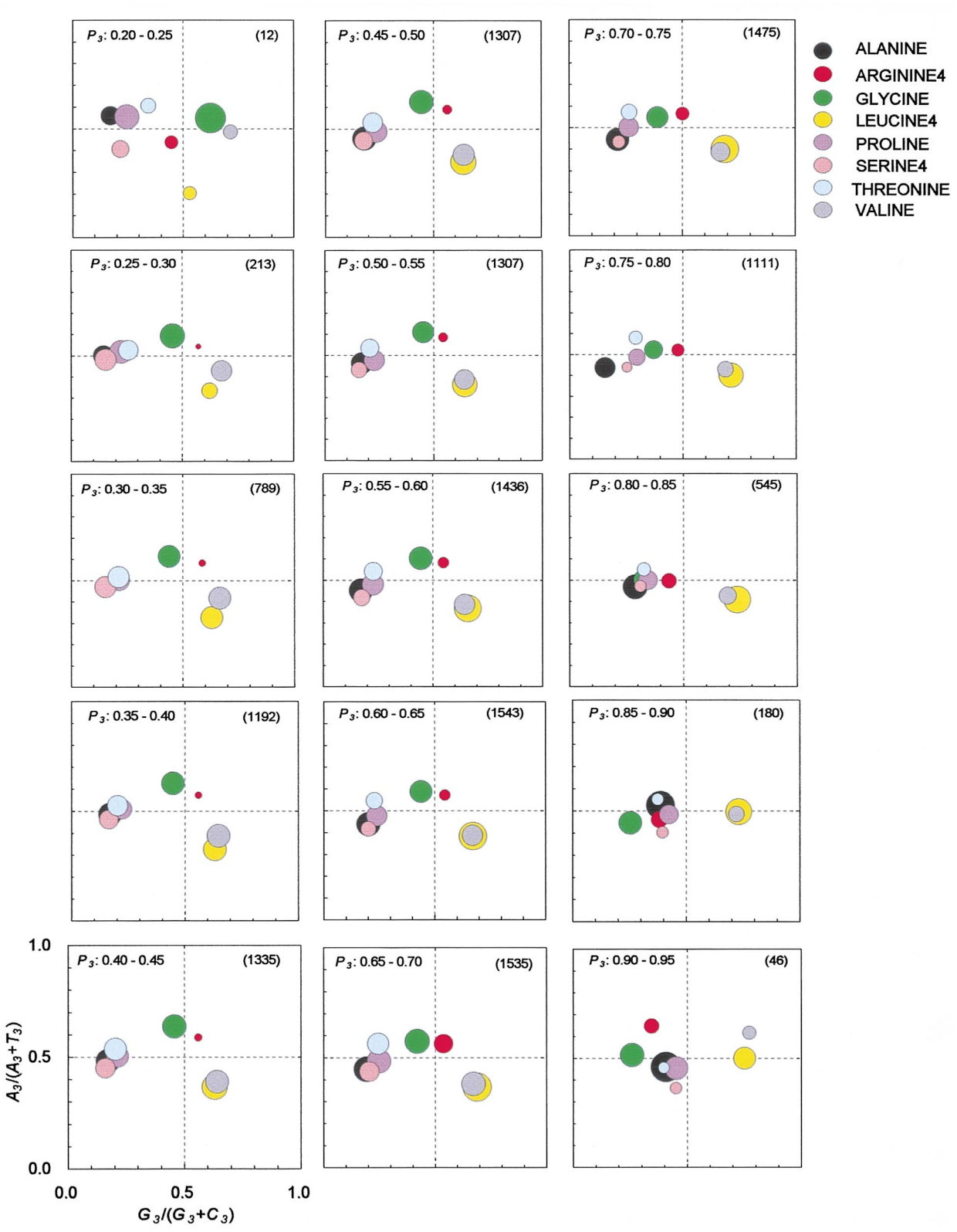

Fig. 3. PR2-bias fingerprints of human genes in increasing P_3-ranges. The genes in each P_3-range were analyzed for PR2 biases. The bubble size (diameter) is proportional to the relative frequency of each four-codon amino acid. The number in parentheses represents the number of genes in each P_3 range. The color code for each bubble is shown on the right margin of the panel. The highest P_3 range (0.95–1.00) is not shown because only one gene exists in the current data (see Section 2.3).

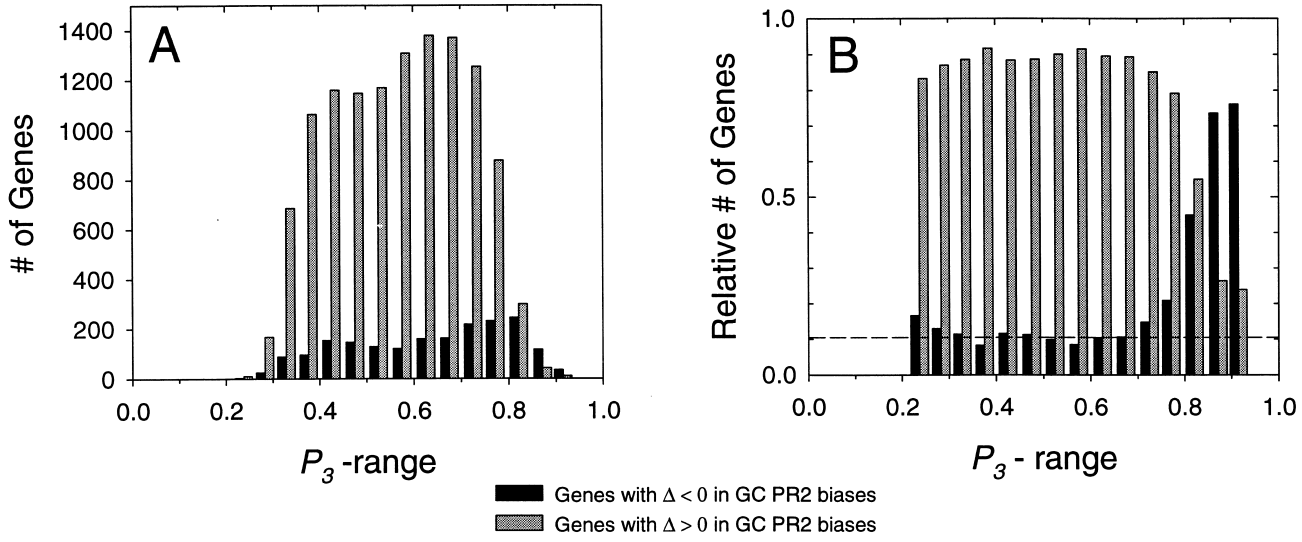

Fig. 4. Analysis of the high-P_3 anomalous genes. The detailed description of the method of calculating gene frequencies of the major- and anomalous-PR2 bias types in individual P_3 ranges are presented in Section 4.3. (A) Frequencies of genes with higher GC PR2-bias of glycine than that of alanine (major type, gray column) and genes with lower GC PR2-bias of glycine than that of alanine (anomalous type, dark column) for each P_3 range. (B) The same data shown in the A-panel were re-plotted as relative fractions of the two types for individual P_3 range.

of PR2 biases in four-codon amino acids. The four-codon amino acids represent 49.5% of the total codons in the present data.

The directional mutation pressures of A/T (W) $\rightleftharpoons$ G/C (S), with two opposing directional mutation pressures that leads to an equilibrium in the G + C content, may be the main cause for the inter-genomic heterogeneity of the DNA G + C content in bacteria (Sueoka, 1962). Under this theory where v for W $\rightarrow$ S and u for W $\leftarrow$ S, the P_3 reaches the value of $v/(u + v)$ at equilibrium. This Equilibrium Theory was based originally on the narrow intra-genomic heterogeneity and wide inter-specific variation of DNA G + C content in bacteria. Therefore, there is no *a priori* reason to apply the theory as such for the wide intra-genomic variations that are generally found in eukaryotes. It has been known that vertebrate chromosomes consist of a large number of isochores, and within an isochore, the G + C content is more or less homogeneous (Bernardi et al., 1985). These authors favor the view that the major factor for the formation of isochores are selection due to functional advantages of the higher G + C content of DNA-RNA being thermally more stable. Recently, a small positive correlation between GC_3 and frequency of codons for hydrophobic amino acids was observed in mammals and birds. The correlation was interpreted to mean that selection for the high G + C content existed in warm blooded vertebrates

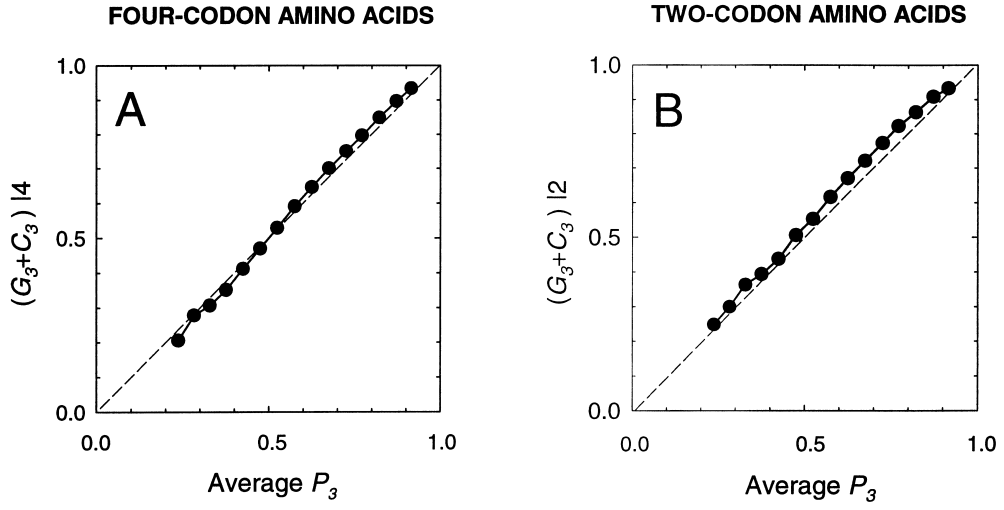

Fig. 5. Average $G_3 + C_3$ content of the synonymous codons in the four- and two-codon amino acids in P_3 ranges. The four-codon amino acids (A) and two-codon amino acids (B) were plotted against individual P_3 ranges.

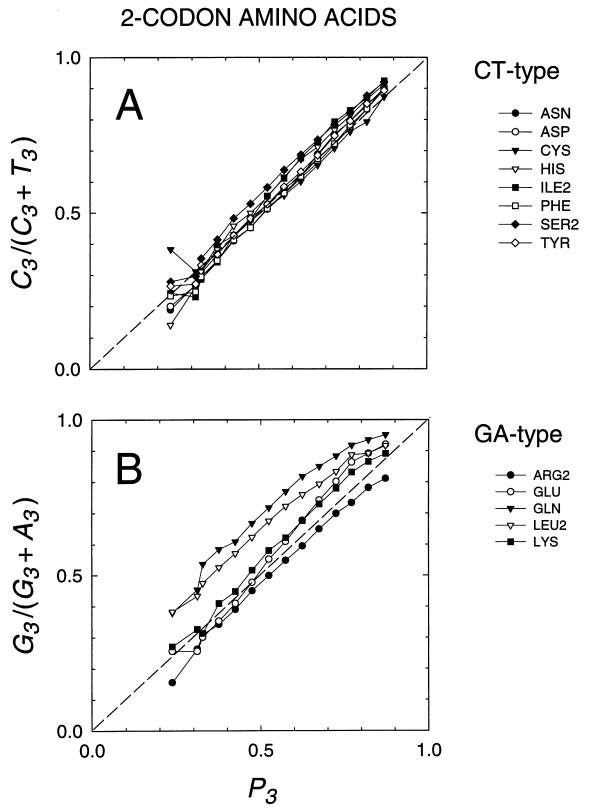

Fig. 6. Regression of S nucleotides with P_3 in two-codon amino acids. (A) CT-type amino acids. (B) GA-type amino acids.

(D'Onofrio et al., 1999). The present results tend to support an alternative model that the difference in the G + C content among isochores is likely to be effected by the isochore-specific directional mutation pressure, and not by the isochore-specific selection through translation-coupled, amino acid-specific biases from the PR2 due to the codon usage preference (Sueoka, 1988, 1995, 1999a). On this point, Francino and Ochman (1999) have recently reported that pseudogenes Ψα1 and Ψη derived from the globin α and β genes of primates that are integrated into different positions within the genome show divergence in the G + C contents toward those of the isochores where they integrate. The new G + C contents of the pseudogenes are almost identical to the local G + C contents of the integrated locations, indicating that the G + C content of the pseudogenes reached near equilibrium. Their results are a strong support for the mutation-equilibrium model (Sueoka, 1962, 1992) of the isochore formation (Filipski, 1987; Sueoka, 1988). The isochore evolution of mammalian genomes may be due to locally different directional mutation pressures, but how the local differences evolved is not clear. It is also clear that the DNA replication-coupled PR2 biases (skews) between leading and lagging strands are not the cause of the G + C heterogeneity, because the PR2-bias distribution (Fig. 2A) does not show two-claster or elongated distribution (Lobry and Sueoka, 2000).

4.2. Additional implications of PR2-biases vs. P_3

The fact that PR2-bias fingerprints of human genes are similar in the wide range of P_3 (Fig. 3) indicates that the amino acid-specific selection pressure is more-or-less similar in the major portion of P_3 range (0.25–0.75) covering 86% of genes. The result is consistent with the notion that for the majority of genes, the codon usage pattern is similar in different cell types. Translation efficiency through the tRNA abundance spectrum may be responsible for the biases of synonymous codon usage. This tRNA model was first proposed by the fact that there are unique amounts of individual tRNAs in *E. coli* (Ikemura, 1981), yeast (Bennetzen and Hall, 1982; Ikemura, 1982). Later development of this model can be found by Ikemura (1985); Sharp et al. (1988); Moriyama and Powell (1997); Akashi (1999) that emphasized tRNA mediated selection plays the major role on the evolution of the DNA base composition including the G + C content. Recently, correlations between the tRNA abundance and the codon usage biases in various bacterial species were further supported by Kanaya et al. (1999). The present result as well as previous results in human and in bacteria (Sueoka, 1988, 1993, 1995, 1999a,b) support also the selection model for amino acid-specific PR2 biases through tRNA abundance, but does not support the selection model for the intra-genomic heterogeneity of the DNA G + C content as the primary cause. Rather, the primary cause to change the DNA G + C content is likely to be directional mutation pressure, and subsequently mutations and selection slowly affect abundance of individual tRNAs, thus adapting to the change of the DNA G + C content.

The extent of amino acid-specific PR2 biases (fingerprints) in human and in most other eukaryotes is as wide as in bacteria although the fingerprints are quite different (Sueoka, 1995). In addition, the result of the present study seems to agree with the possibility that the relative abundance of various tRNA molecules is similar among various cell types in human. Unlike in bacteria and yeast, there has been no convincing evidence of correlations between codon usage frequencies and abundance of proteins. Nevertheless, in higher eukaryotes, biases of the synonymous codon usage may still be correlated with the tRNA frequencies by selection due to the advantage for the translation efficiency.

4.3. Anomalous PR2 biases

The result on the four-codon amino acids presents a remarkably homogeneous fingerprint of PR2 biases in the major range of P_3 (0.25–0.75) (Fig. 3). However, exceptions found in the two extreme P_3 ranges is important to understand the cause of the P_3 heterogeneity in the human genome. The anomaly of the low P_3 region is based on a small number of genes (12) in the lowest P_3 range (0.20–0.25), although some spillover to the next range is likely. The anomaly seems mainly due to an unusually high content

of glycine and threonine, whereas relative positions of the four-codon amino acids in the fingerprint are not very much different from the standard human fingerprint (Figs. 2B and 3). Further study of the low G + C anomaly case is not possible until a larger sample becomes available.

The anomaly of PR2-bias fingerprints of the genes in the high-P_3 ranges (0.75–0.95; 1883 genes or 13%) is more extensive than that of the lowest P_3 range. An exact estimation of the number of the high-P_3 genes is not simple because of their obvious overlapping with the major type. The method of estimating two types of genes described in Section 3.3 provides a way to separate the two types and removing a background.

The PR2-bias fingerprints (four-codon amino acids) of the high P_3 ranges show systematic changes in bubble position. Thus, shifts of bubbles are seen toward right (toward $G_3 > C_3$) in serine4, proline, threonine and alanine (Type 1 shift), toward left (toward $C_3 > G_3$) in arginine4 and glycine (Type 2 shift) and no or a slight shift toward right in leucine4 and valine (Type 3 shift). It is interesting to note that these shifts (Fig. 3) are correlated with the bases of the second codon positions for the four-codon amino acids in question; namely, C_2 for Type 1 shift (alanine, proline, serine4, and threonine), G_2 for Type 2 shift (in arginine4 and in glycine); and T_2 for type 3 shift (in leucine4 and valine). The transition of C to T through CpG methylation and deamination may explain Type 1- and Type 3- but not Type 2-shift.

A serious alternative to the above picture of the anomalous PR2 biases is the possibility that the anomaly is due to gradual and uniform change of Δ of all genes in the high G + C ranges rather than the mixture of a major and anomalous-type genes. This alternative possibility should be investigated further with more genes in high G + C ranges. Further studies of the anomalous patterns of PR2 biases are important to understand the relative roles of mutation and selection in the evolution of codon usage biases. Currently, the causes for the anomalies are not clear. The anomalies, however, do not change the overall conclusion that the heterogeneity of P_3 in the major P_3 range does not influence the PR2 biases substantially.

4.4. Two-codon amino acids

When all data are combined, the frequency of codons with C or G at the third codon position shows almost perfect correspondence with P_3 in both two-codon amino acids and four-codon amino acids (Fig. 5), indicating that the GC-mutation pressure affects the codon frequencies more-or-less uniformly within individual isochores. This intragenic uniformity is best interpreted as the mutational effect than the selective effect at the DNA base composition level (Sueoka, 1992). Average values of $C_3/(C_3 + T_3)$ for individual CT-type amino acids (the third codon position is T or C) and those of $G_3/(G_3 + A_3)$ for AG-type amino acids (the third codon position is A or G) in each P_3-range were separately plotted against P_3 as Fig. 6A,B, respectively. The relative usage of S nucleotides (C or G) at the third codon position correlates almost perfectly with P_3 in both TC-type and AG-type. Uniformly higher values of $G_3/(G_3 + A_3)$ for glutamine and leucine2 above the diagonal line (Fig. 6B) indicate that these amino acid-specific PR2 biases are independent of P_3. Since this effect covers the whole P_3 ranges, the most likely cause is a preference of G_3 codon to A_3 in glutamine and leucine2 in the tRNA-related portion of the translation system. To explain this point, further studies with different organisms are necessary.

4.5. Do DNA replication- and transcription-coupled PR2 biases exist in the human genome?

As described in introduction, PR2 biases are generated by at least three strand-specific (asymmetric) mutations and/or selections; (a) the amino acid-specific, translation-coupled selections, (b) the DNA replication-coupled directional mutation pressure, and (c) the transcription-coupled directional mutation pressures. Among these three factors, the amino acid specific PR2 biases is ubiquitous among the organisms so far examined. As shown in this article, it is unlikely that the translation-coupled PR2 is responsible for the intra-genomic heterogeneity of the DNA G + C content in human. In contrast, the DNA replication- and translation-coupled PR2 biases can cover large regions and, therefore, they are potentially capable of changing the G + C contents. However, as observed in bacteria, the DNA replication-coupled PR2 biases are likely to be similar throughout the genome, unless locally different chromatin structures (e.g. in isochores) influence the PR2 biases. It is known in bacteria that the DNA replication- and transcription-coupled PR2 biases are not ubiquitous among bacterial species, and the extent of the biases is widely variable even among those organisms that show the biases (Lobry, 1996; Mclean et al., 1998; Grigoriev, 1999; Lobry and Sueoka, 2000).

Assessing the effect of DNA replication- and transcription-coupled PR2 biases on the intra-genomic heterogeneity of the DNA G + C content in higher eukaryotes is currently difficult because of the paucity of data on the relationship between the local orientations of DNA replication. However, the tight distribution of the PR2 biases of human genes (Fig. 2A) strongly suggests that, in human, DNA-replication coupled PR2 biases exist generally to a small extent, if any. The analysis of bacterial genome suggests that DNA replication-coupled PR2-biases generate oblong gene distributions of various degrees or, in extreme cases, two completely separate distributions of genes among species on the PR2 bias plot (Lobry and Sueoka, submitted).

5. Conclusions

The present result of P_3-PR2 bias analyses favors the view that intra-genomic heterogeneity of human genome is mainly due to directional mutations specific to isochores.

It is not yet certain whether the same conclusion is applicable to the anomalous PR2 biases in high G + C ranges of P_3 (0.75–9.5). Assuming that the relative abundance of tRNAs is similar in different cell types, two alternative mechanisms may be proposed. One is that high GC-mutation pressures in high G + C isochores force the mutated codons of the genes to use minor isoaccepting tRNAs for some amino acids in sacrifice of the translation efficiency (mutation model), and the other is that high G + C isochores were formed by selection in which genes are more functionally adapted to high G + C isochores (selection model).

Acknowledgements

This work was supported originally by NSF (DIR8820806) and Yamamoto Foundation for Promotion of Genetic Research to N.S. We are particularly grateful to Dr Takashi Gojobori for his generous gift of the human data through DDBJ.

References

Akashi, H., 1999. Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene 238, 39–51.

Bennetzen, J.L., Hall, B., 1982. Codon selection in yeast. J. Biol. Chem. 257, 3026–3031.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241, 3–17.

Bernardi, G., Olfsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of the vertebrates. Science 228, 953–958.

D'Onofrio, G., Jabbari, K., Musto, H., Bernardi, G., 1999. The correlation of protein hydropathy with the base composition of coding sequences. Gene 238, 3–14.

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

Francino, M.P., Ochman, H., 1999. Isochores result from mutation not selection. Nature 400, 30–31.

Francino, M.P., Chao, L., Riley, M.A., Ochman, H., 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science 272, 107–109.

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. J. Theor. Biol. 157, 127–133.

Grigoriev, A., 1999. Strand-specific compositional asymmetries in double stranded DNA viruses. Virus Res. 60, 1–19.

Grigoriev, A., Freeman, J.M., Plasterer, T.N., Smith, T.F., Mohr, S.C., 1998. Genome arithmetic. Science 281, 1923–1924.

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 RNAs 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 isoaccepting tRNAs. J. Mol. Biol. 158, 573–597.

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

Kanaya, S., Yamada, Y., Kudo, Y., Ikemura, T., 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238, 143–155.

Kano-Sueoka, T., Lobry, J.R., Sueoka, N., 1999. Intra-strand biases in bacteriophage T4. Gene 238, 59–64.

Karkas, J.D., Rudner, R., Chargaff, E., 1968. Separation of *B. subtilis* DNA into complementary strands, II. Template functions and composition as determined by transcription with RNA polymerase. Proc. Natl. Acad. Sci. USA 60, 915–920.

Lobry, J.R., 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. 40, 326–330.

Lobry, J.R., 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13, 660–665.

Lobry, J.R., Sueoka, N., 2000. Asymmetric directional mutation pressure in ten eubacterial genomes. Mol. Biol. Evol, (submitted).

Mackievicz, P., Gierlik, A., Lowalczuk, M., Dudek, M.R., Cebrat, S., 1999. Asymmetry of nucleotide composition of prokaryotic chromosomes. J. App. Genetics 40, 1–41.

McInerney, J.O., 1998. Replication and transcriptional selection on codon usage in *Borrelia burgdoferi*. Proc. Natl. Acad. Sci. USA 95, 10698–10703.

Mclean, M.J., Wolfe, K.H., Devine, K., 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. 47, 691–696.

Moriyama, E.N., Powell, J.R., 1997. Codon usage bias and tRNA abundance in *Drosophila*. J. Mol. Evol. 45, 514–523.

Morton, B.R., 1999. Strand asymmetry and codon usage bias in the chloroplast genome of *Euglena gacilis*. Proc. Natl. Acad. Sci. USA 96, 5123–5128.

Murazek, J., Karlin, S., 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA 95, 3720–3725.

Muto, A., Osawa, S., 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA 84, 166–169.

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

Rudner, R., Karkas, J.D., Chargaff, E., 1968. Separation of B. subtilis DNA into complementary strands, III. Direct analysis. Proc. Natl. Acad. Sci. USA 60, 921–922.

Rudner, R., Karkas, J.D., Chargaff, E., 1969. Separation of microbial deoxyribonucleic acids into complementary strands. Proc. Natl. Acad. Sci. USA 63, 152–159.

Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., Wright, F., 1988. Codon usage patterns in *Escherichia coli Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster*, and *Homo sapiens*: a review of the considerable within-species diversity. Nucleic Acids Res. 16, 8207–8211.

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., 1964. On the evolution of informational macromolecules. In: Bryson, V., Vogel, H.J. (Eds.). Evolving Genes and Proteins. Academic Press, New York, pp. 479–496.

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, 95–114.

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

Sueoka, N., 1995. ntra-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., 1999a. Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G + C content of third codon position. Gene 238, 53–58.

Sueoka, N., 1999b. Two aspects of DNA base composition: G + C content

and translation-coupled deviation from intra-strand rule of A = T and G = C. J. Mol. Evol. 49, 49–62.

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.

Wada, K.N., Doi, H., Tanaka, S., Wada, Y., Furusawa, M., neo-Darwinian, A., 1993. algorithm: asymmetrical mutations due to semiconservative DNA-type replication promote evolution. Proc. Natl. Acad. Sci. USA 90, 11934–11938.

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

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