J Mol Evol (2001) 53:456–464
DOI: 10.1007/s002390010235

JOURNAL OF MOLECULAR EVOLUTION

Intragenic Variation of Synonymous Substitution Rates Is Caused by Nonrandom Mutations at Methylated CpG

Kazuhisa Tsunoyama,[1,3] Matthew I. Bellgard,[2,3] Takashi Gojobori[1,3]

[1] Department of Genetics, Graduate University for Advanced Studies, Mishima, Japan.
[2] School of Information Technology, Murdoch University, Murdoch, 6150, WA, Australia.
[3] Center for Information Biology, National Institute of Genetics, Mishima, Japan.

Received: 12 January 2001 / Accepted: 28 February 2001

Abstract. It has been observed that synonymous substitution rates vary among genes in various organisms, although the cause of the variation is unresolved. At the intragenic level, however, the variation of synonymous substitutions is somewhat controversial. By developing a rigorous statistical test and applying the test to 418 homologous gene pairs between mouse and rat, we found that more than 90% of gene pairs showed a statistical significance in intragenic variation of synonymous substitution rates. Moreover, by examining all conceivable possibilities for the cause of the variation, we successfully found that intragenic variation of synonymous substitutions in mammalian genes is caused mainly by a nonrandom mutation due to the methylation of CpG dinucleotides rather than by functional constraints.

Key words: Intragenic variation — Synonymous substitution rates — Methylation — Molecular evolution

Introduction

As synonymous substitutions of nucleotide do not affect the primary structure of a protein, it has been commonly thought that functional constraints for synonymous changes are either very weak or nonexistent (Kimura 1968; King and Jukes 1969). Thus, synonymous substitutions had been supposed to directly reflect spontaneous mutation. As a result, as long as the spontaneous mutation rate is constant, the rates of synonymous substitutions are expected to be fairly uniform among different genes as well as within a gene (Kimura 1983).

In reality, the intergenic variation of synonymous substitution rates has been observed for most organisms and many researchers have discussed the cause of intergenic variation for more than a decade (Graur 1985; Li et al. 1985; Britten 1986; Wolfe et al. 1989; Bernardi et al. 1993; Wolfe and Sharp 1993; Mouchiroud et al. 1995; Ohta and Ina 1995; Comeron and Kreitman 1998). However, it has been unclear whether the cause of the variation is due to functional constraints or a nonrandom mutation.

Since the underlying spontaneous mutation rate is considered to be more constant within a gene rather than among genes, the rate of synonymous substitution is thought to be constant within a gene if synonymous substitutions are exempted from the functional constraints at the DNA or mRNA levels. However, several studies using the window analysis have suggested that the synonymous rate is not uniform within a gene of viruses, bacteria, and Drosophila (Lawrence et al. 1991; Eyre-Walker and Bulmer 1993; Ina et al. 1994; Cacciò et al. 1995; Zoubak 1995; Comeron and Aguadé 1996). As for mammalian genes, Cacciò et al. (1995) and Zoubak et al. used 69 homologous genes of four mammalian orders and they suggested that the synonymous substitution pro-

Correspondence to: T. Gojobori; *present address:* Center for Information Biology, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan; *email:* tgojobor@genes.nig.ac.jp cess is nonrandom even within the genes. However, when they statistically examined homologous gene pairs of the same mammalian orders, they could not find significant nonrandomness of synonymous substitutions. Thus, in mammals, the intragenic variation of synonymous substitution rates is somewhat unclear.

In order to solve this issue, we developed a rigorous statistical test to examine whether the rates of synonymous substitutions vary within a gene. We then applied the test to 418 homologous gene pairs between mouse and rat. This comparison was made because these two species provide the largest number of mammalian homologous gene pairs currently available in the public databases. Moreover, the divergence between these two species was not too large to confront the saturation effect of synonymous substitutions but not too small that we would suffer from a shortage of substitutions.

Materials and Methods

Data Extraction. The gene pairs between *R. norvegicus* and *M. musculus* were extracted from the SODHO database (Tateno et al. 1997) which was constructed with the DDBJ database release 30. The following three criteria were used to further select gene pairs. (1) We extracted gene pairs sharing the same function to ensure that the pair was orthologous. (2) In order to avoid statistical fluctuation due to the window size, we eliminated gene pairs from the analysis whose gene lengths were less than twice the window size. (3) We used gene pairs that had no gaps in the pairwise alignment, guaranteeing that gene pairs were of the same length. We are fully aware that two serious problems on the window analysis can arise if we use pairwise alignments with gaps: (1) If we omit the gapped regions in a given gene sequence and then conduct the window analysis, some windows will contain consecutive regions that are artificially connected to each other and thereby have no biological meaning, (2) If one ignores gaps in the calculation for a window, then the estimated values will depend on the number of codons in the window. These problems are more serious when the window size is relatively small.

To conduct the window analysis, a window was set on the first codon of the pairwise alignment and shifted one codon at a time. This process was repeated by shifting the window codon by codon until it reached the last codon of the alignment. The window size was chosen to be 60 bases (20 codons) unless mentioned otherwise. We then estimated the proportion (Ps) of synonymous differences for each window. Use of the Ps value is sufficient as it is free from the "saturation" effect of synonymous substitutions. The modified Nei and Gojobori method (Zhang et al. 1998) was used for this estimation. In the present study, we did not use the original Nei and Gojobori (1986) method as it has been shown that it may underestimate the Ps value when there is a strong transition/transversion bias (Ina 1995). The other estimation methods published were not used as they sometimes return inapplicable values when closely related sequences are used (Zhang 1998).

Furthermore, we eliminated gene pairs that had implausibly high Ps values (Ps > 1) due to statistical fluctuation from the sampling errors.

In this way, we finally obtained 418 gene pairs from rat and mouse.

Statistical Test for Intragenic Variation of Synonymous Substitutions. For each gene pair, we estimated the proportion (Ps) of synonymous differences within a gene by the window analysis. In order to examine, with statistical validity, whether Ps values vary within the gene, we generated random nucleotide sequences reflecting codon usage of the gene pair and compared them with actual sequences.

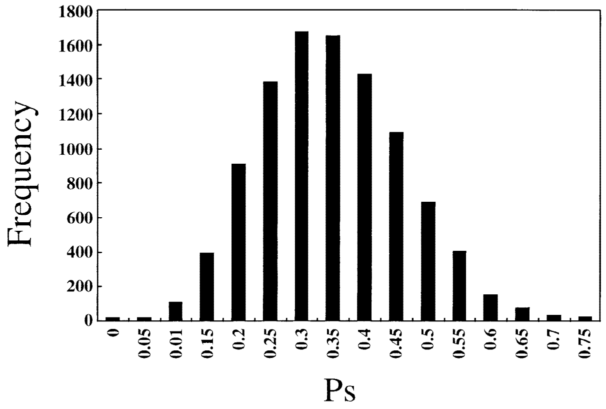

Fig. 1. An example of the distribution of 10,000 Ps values made randomly for a gene pair. The gene is protein kinase C inhibitor.

The statistical test for each gene pair was conducted as follows. (1) We computed the frequency of each codon pair of the two gene sequences aligned. (2) Using these frequencies, we generated random nucleotide sequences such that they have the same length as the window size and they reflect the codon usage of the actual pairwise alignment. (3) We generated 10,000 pairs of random sequences and estimated the Ps value for each pair of random sequences. Thus, a random distribution of 10,000 Ps values is obtained (see Fig. 1 as an example of the distribution). Finally, we computed the probability of the Ps value observed for each window on the actual pairwise alignment by using the distribution of 10,000 Ps values.

Methods for Examining the Cause of Intragenic Variation of Synonymous Substitutions. We investigated the causes of intragenic variation of synonymous substitutions by examining possible correlations between Ps and other measures; the proportion (Pn) of nonsynonymous differences, the codon usage bias, the mRNA structures, the base composition, and a frequency of CpG dinucleotides. These measures, except the mRNA structures, were also estimated for each window. Whenever the correlation analysis was conducted, we used the windows which were not overlapped to each other, in order to ensure independence of calculated measures. As for the codon usage bias, base contents, and frequencies of CpG dinucleotides, we computed the averages between a pair of genes for each window. We then conducted the correlation analysis. First, we calculated Pearson's correlation coefficient between Ps values and one of these measures for each gene pair. Since we selected only gene pairs having at least nine non-overlapping windows in the correlation analysis, the number of gene pairs that we could use was 316. The remaining 102 gene pairs (= 418 − 316) were not used. Second, we conducted the *t*-test for each correlation coefficient by setting the significance level at 5%. We then calculated the probability that the observed number of gene pairs having correlation coefficients with statistical significance was expected by chance with the binomial distribution. Third, we then used the reduced 0.0158% (= 5%/316) level of significance for each correlation coefficient in accordance with the Bonferroni method. In the Bonferroni method (Sokal and Rohlf 1969), the overall significance level divided by the number of comparisons is used as a significance level for each comparison. This is because the overall significance level may become larger than 5% if we use the 5% of significance for each correlation coefficient. If we find at least one significant correlation coefficient by this method, we can exclude a possibility that significance of the overall correlation between Ps and the measure examined took place by chance.

We used ENC (an effective number of codons) as a measure of codon usage bias (Wright 1990); it quantifies how far the codon usage of a gene departs from equal usage of synonymous codons. Note that when the short length of windows is used, the biased value of ENC is

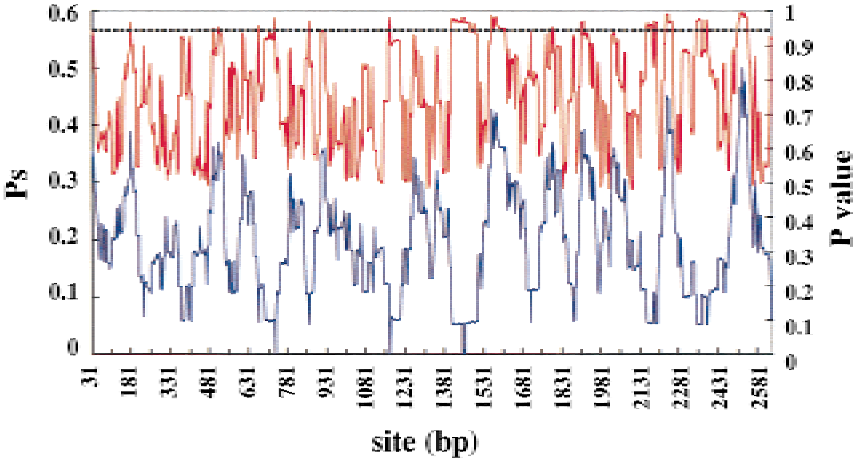

Fig. 2. Window analysis for the gene pair of iron-responsive element binding factor. The estimated Ps value and the corresponding probability (from the statistical test) are assigned at a codon site located at the center of the window. Thus, the start site is base number 31. The Ps values are shown in blue and the corresponding p values are shown with a red line. The 95% of confidence level for each Ps value is shown with a straight dotted line.

likely to be obtained (Comeron and Aguadé 1998). Thus, for the comparison between the Ps and ENC values, we used 300 bp of the window length instead of a regular window size of 60 bp.

We calculated the GC1%, GC2%, and GC3% (the average GC content at the 1st, 2nd, and 3rd positions of the codon, respectively) as base compositions. When we calculated these measures, we excluded codons having no synonymous codons because they do not contribute to synonymous substitutions. Such codons were Met and Trp in the comparison between mouse and rat.

Results

Significant Intragenic Variation of Synonymous Substitution Rates

Interestingly enough, in the results of the statistical test that was conducted to check significance of intragenic variation of the Ps values, 92% of 418 gene pairs showed statistically significant variation of the Ps values within a gene.

These results were not overly affected by a window size. When we used a window size of 540 bases (180 codons) instead of 60 bases, 54% of the gene pairs showed statistically significant intragenic variation of the Ps values. Therefore, in spite of the relatively large window size of 540 bp, which is approximately half of the average gene length over all 418 gene pairs, more than half of the compared gene pairs still showed statistically significant variation of synonymous substitution rates within a gene.

Moreover, although a longer length gene was expected to show significant variation, we did not observe any notable correlation between the gene length and the number of intragenic regions where the Ps value is significantly high or low (data not shown). Thus, almost all examined gene pairs have at least one intragenic region where synonymous variation is statistically significant. In this region, the Ps value is high or low at the 5% level of significance.

As an example, Fig. 2 shows the intragenic variation of Ps values for the gene pair of iron responsive element binding factor. The *P* value for each of Ps values is shown in the figure, demonstrating the intragenic regions where the Ps value is significantly high or low.

Examination of Possible Causes of Intragenic Variation

Moreover, we examined all conceivable possibilities that may cause the intragenic variation of synonymous substitutions. In particular, we examined whether the rate of synonymous substitution was correlated with that of nonsynonymous substitution, the degree of codon usage bias, the stem or loop regions of the mRNA secondary structures, base content, and the frequency of CpG dinucleotides.

(1) Functional Constraints at the Protein Level. We examined the possibility that functional constraints against nonsynonymous substitutions work even on synonymous substitutions. This possibility was deduced from the observation that the gene having a low rate of synonymous substitution also manifests a low rate of nonsynonymous substitution. In fact, this has been observed in bacteria, Drosophila, and mammals (Graur 1985; Li et al. 1985; Britten 1986; Wolfe et al. 1989; Bernardi et al. 1993; Wolfe and Sharp 1993; Mouchiroud et al. 1995; Ohta and Ina 1995; Comeron and Kreitman 1998). In order to examine this possibility, we conducted a window analysis for computing the proportions (Pn) of nonsynonymous differences. Table 1 shows the number of gene pairs in which correlations between the Ps and Pn were found to be significant. As shown in the table, 49 of 316 gene pairs showed the significant correlation at the 5% significance level. When we used a lower level of significance, that is, 0.0158% according to Bonferroni method, only one gene pair showed significance in a correlation between Ps and Pn. Thus, the intragenic variation of synonymous substitution rates may, to some extent, be caused by functional constraints of proteins.

Table 1. Examination of several possibilities for intragenic variation of synonymous substitutions

	No. of sig. pair[a] (5% level)	Prob.[b]	+/–[c]	No. of sig. pair[a] (0.0158% level)	+/–[c]
Pn	49	$2.74*10^{-12}$	40/9	1	1/0
ENC	2	0.071	0/2	0	–/–
GC1%	61	$1.39*10^{-19}$	32/29	0	–/–
GC2%	64	$1.31*10^{-21}$	39/25	0	–/–
GC3%	71	$1.26*10^{-26}$	31/40	2	0/2
C1G2	65	$2.66*10^{-22}$	40/25	1	1/0
C2G3	67	$1.04*10^{-23}$	59/8	5	5/0
C3G1	90	$4.17*10^{-42}$	81/9	4	4/0

The correlation coefficients between the Ps values and these measures were examined. A window size of 60 bp was used for Pn, GC3%, and frequency of CpG. For ENC, a 300 bp window was used. Thus, the number of available genes was reduced to nine for this measure. We used 316 gene pairs for the other measures.

[a] Total number of gene pairs showing the significant correlation with Ps.

[b] Probability of the number of observed significant correlation coefficients.

[c] Number of gene pairs showing significant positive and negative correlation with Ps.

One such possible constraint may originate from translational efficiency that depends upon the frequency of occurrence of rare codons, because amino acids encoded by the rare codons eventually affect protein function. Otherwise, a nonrandom mutation may affect the rates of both synonymous and nonsynonymous substitutions. Thus, the possibility that functional constraints against nonsynonymous substitutions work even on synonymous substitutions is unclear at this stage, although we observed significant correlation between the rate of synonymous substitution and that of nonsynonymous substitution.

(2) Bias of Codon Usages. We examined codon usage bias, though it is also related to the above-mentioned possibility. Numerous studies of Drosophila genes have shown that the degree of codon usage bias is negatively correlated with the rate of synonymous substitution (Shields et al. 1988; Sharp and Li 1989; Moriyama and Gojobori 1992; Comeron and Kreitman, 1998). In bacteria and yeast, the degree of codon usage bias is correlated with the level of gene expression, and the codon used most frequently in each synonymous codon family shows a clear relationship with tRNA-abundance (Ikemura 1981; Ikemura 1982; Sharp and Li 1986). Moreover, it has been shown that genes having a strong bias of codon usage have evolved with a slower rate of synonymous substitution (Sharp and Li 1986; Powell and Moriyama 1997). It has also been suggested that selection for translation accuracy works on synonymous substitutions (Akashi 1994).

To test a possible relationship between the intragenic variation of synonymous substitution rates and the degree of codon usage bias, we calculated the ENC value by the window analysis. The ENC values quantify how far the codon usage of a gene departs from equal usage of synonymous codons (Wright 1990). Because we used 300 bp of the window length instead of a regular window size of 60 bp, the number of gene pairs compared reduced to 9. As shown in Table 1, only 2 out of 9 gene pairs showed significance in a correlation between Ps and ENC. This number of pairs showing significant correlations was not enough to conclude statistical significance of the overall correlation between Ps and ENC. This is because the probability that the observed number of gene pairs having significant correlation coefficients was expected by chance was not less than the 5% of significance level. Indeed, when the lower level of significance is adopted, we could not find any gene pair showing significant correlation between Ps and ENC.

Two previous studies, which focused on the intragenic variation, also observed no correlation between the rate of synonymous substitutions and the degree of codon usage bias (Lawrence et al. 1991; Eyre-Walker and Bulmer 1991).

(3) The Secondary Structure of mRNAs. The observation of no correlation between the intragenic variation of synonymous substitution rates and the codon usage bias would be understandable if selection was only acting on the mRNA secondary structure (Eyre-Walker and Bulmer 1993). In other words, it suggests that functional constraints working on synonymous substitutions are at the mRNA level, not at the protein level. Indeed, several studies have reported that there is a relationship between mRNA secondary structure and synonymous substitutions in the genes of bacteria and hepatitis C virus (Lawrence et al. 1991; Comeron and Aguadé 1996; Smith and Simmonds 1997). We then investigated possible functional constraints for the maintenance of mRNA secondary structure affecting synonymous substitutions. Unfortunately, however, only three gene pairs (histone subunit 1, glycoprotein hormone alpha subunit, and regenerating protein I) in our data set have descriptions of their mRNA sequences in the entries of the DDBJ/EMBL/GenBank database. This is because, although we could use the cDNA data for prediction of a mRNA secondary structure, the lack of 5′ and 3′ untranslated regions affect the result of prediction. Among three gene pairs, only the histone subunit 1 showed statistical significance in intragenic variation at the 5% level. Although the number of data is very limited, we estimated an mRNA secondary structure for each of histone subunit 1 genes by the mfold software version 2.3 (Zuker 1989). As shown in Fig. 3, the intragenic region, where Ps is significantly high, was contained in both the predicted stem and loop regions. On the other hand, the intragenic region where Ps is significantly low was contained in the predicted loop region of mRNA structures

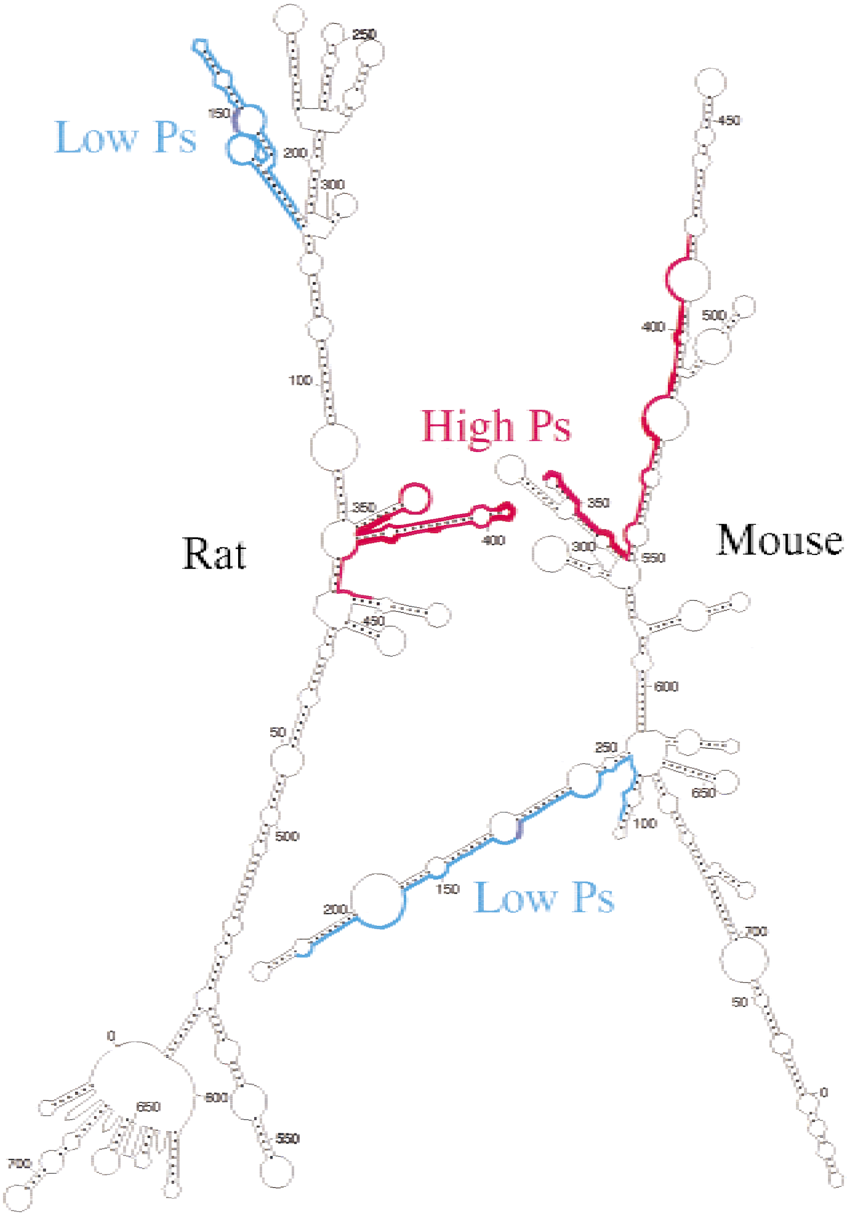

Fig. 3. Estimated mRNA secondary structures of histone subunit 1 genes of both mouse and rat. Red and blue lines correspond to the region where significantly high and low Ps values were shown, respectively. Red and blue regions contain sites in which synonymous substitutions were observed.

of both mouse and rat. However, the structural features corresponding specifically to these regions are quite different between mouse and rat. Thus, the possibility of functional constraints at the mRNA level is doubtful, at present, as a cause of the intragenic variation of synonymous substitution rates.

(4) Base Composition. We considered the possibility of functional constraints acting at the DNA level (Ticher and Graur 1989; Wolfe and Sharp 1993). It has been recently suggested that there are functional constraints which maintain a particular base composition (Alvarez-Valin et al. 1998). To test this possibility, we calculated the GC1%, GC2%, and GC3%, the GC content at the 1st, 2nd, and 3rd positions of the codon, respectively, and examined their correlations with Ps. As a result, for GC1%, GC2%, and GC3%, 61, 64, and 71 gene pairs out of 316 showed significant correlations with Ps at the 5% level of significance, respectively (Table 1). Among these three measures, only GC3% showed overall significance of its correlation with Ps because two gene pairs showed significant correlations when we used a lower significance level.

(5) Spontaneous Mutation Rate. Finally, we investigated the remaining possibility that the intragenic variation of synonymous substitutions reflect heterogeneity of the mutation rate within a gene. Such nonrandomness of mutation has been known to be ‘hotspots’ of mutation. In order to examine this possibility for nonrandomness of mutation, we calculated the average frequency of CpG dinucleotides, because almost all regions of vertebrate

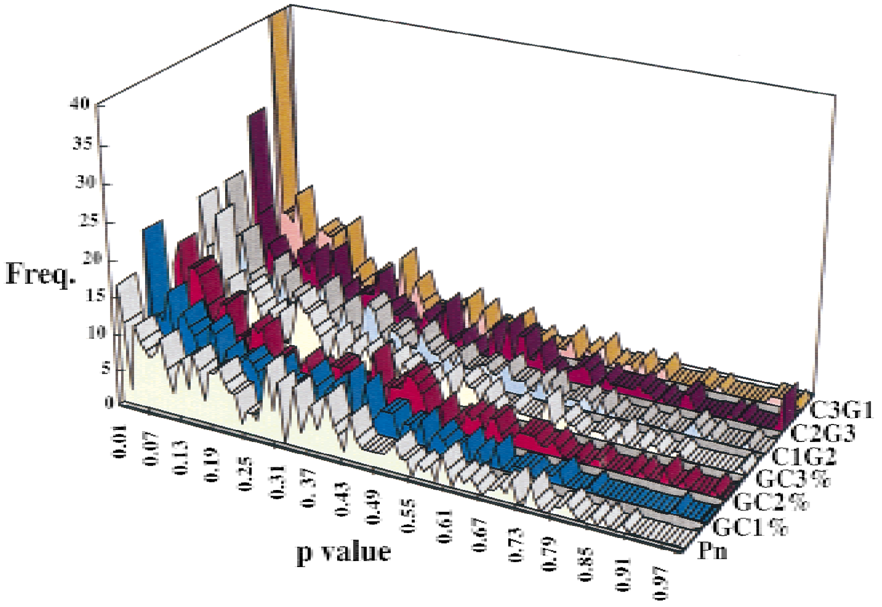

Fig. 4. Frequency distribution for the p values of 316 correlation coefficients.

genomes are subject to methylation and it is generally accepted that the methylcytosine, which is known as a mutable site, exists primarily in the CpG dinucleotide (Bird 1993; Holiday and Grigg 1993). As shown in Table 1, 65 gene pairs out of 316 showed significant correlations between Ps and C1G2 (CG dinucleotides of the first and second codon positions), and 67 gene pairs showed significant correlation between Ps and C2G3 (CG dinucleotides of the second and third codon positions). Moreover, C3G1 (CG dinucleotides of the third and first codon positions spanning two codons) showed statistically significant correlations with the Ps values for 90 gene pairs. When we used a lower level of significance, only one gene pair showed a significant correlation between Ps and C1G2. Interestingly enough, a larger number of 5 gene pairs showed significant correlations between Ps and C2G3 at a lower level of significance. Moreover, for C3G1, a larger number of 4 gene pairs showed significant correlations with Ps at a lower level of significance. Thus, gene pairs having significant correlations with Ps were more frequently observed in the correlation analysis for C2G3 and C3G1 than in the other correlation analysis. These results lead us to the possibility that intragenic variation of synonymous substitutions reflects heterogeneity of the mutation rate within a gene.

Intragenic Variation of Synonymous Substitutions Is Caused Mainly by a Nonrandom Mutation

Among the above-mentioned 5 possibilities, when we calculated the probability that the observed number of significant correlation coefficients is expected by chance, all measures except ENC showed that correlations with the Ps are statistically significant at the 5% level. Thus, at this stage, a possibility for the degree of codon usage bias was rejected. Moreover, when we used a lower level of significance, GC1% and GC2% did not show significant correlation with Ps at all. Therefore, we eliminated the possibilities for base compositions at the 1st and 2nd positions of the codon. On the other hand, the possibility for base compositions (GC3%) at the 3rd position of the codon remained because two gene pairs showed significant correlations between Ps and GC3%. Thus, including this possibility, the remaining possibilities were (1) functional constraints at the protein level, (2) base composition at the 3rd position of the codon, and (3) spontaneous mutation rate. These possibilities were examined by five measures of Pn, GC3%, C1G2, C2G3, and C3G1.

Then we investigated the frequency distribution for the probabilities of correlation coefficients, as shown in Fig. 4. For each correlation analysis, we computed the probability for each of 316 correlation coefficients. As shown in the figure, a probability lower than 1% was most frequently observed in the correlation analysis for all of these five measures. However, C3G1 showed the largest number of correlation coefficients having the probability lower than 1% when compared with the other measures. The second largest number of correlation coefficients having a probability lower than 1% was observed in the correlation analysis of C2G3. Therefore, we considered that the frequencies of C2G3 and C3G1 dinucleotides may be related to the cause of intragenic variation of synonymous substitutions.

Fig. 5 shows one example of a gene pair of vesicle transporter protein. Out of 66 codon pairs where synonymous substitutions can be observed, 30 codons have dinucleotide C2G3 or C3G1 in either one of a codon pair. At these codon sites having synonymous changes, the most frequently observed substitution for C3G1 is a substitution from C to T at T3G1, whereas the one for C2G3

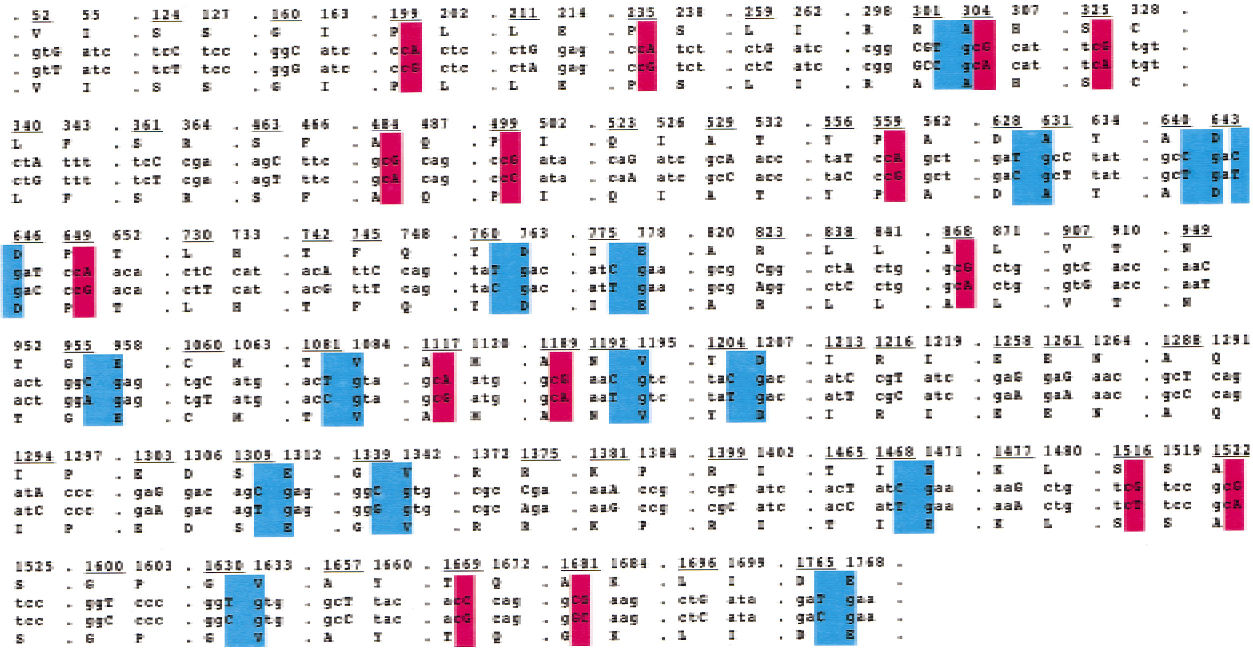

Fig. 5. An example of a pairwise alignment of a vesicle transporter protein. The lines indicate in order: codon site number (bp), amino acid sequence for mouse, DNA sequence for mouse, DNA sequence for rat, and amino acid sequence for rat, respectively. Codon sites with synonymous changes and the neighboring codons are shown. Dots indicate identical codons or codon sites with nonsynonymous changes. The number for codon sites underlined indicates codon sites with synonymous changes. Blue and red boxes show codon sites with synonymous changes at C3G1 and C2G3 in one of the genes, respectively. In this alignment, out of 66 codon pairs where synonymous substitutions can be observed, 15 codons have C3G1 and 15 have C2G3 in one of the species.

is a substitution from G to A, resulting in C2A3. This observation is consistent with mutation at a methylated CpG dinucleotide producing a TpG and its complementary CpA dinucleotide.

Our results, including the correlation analysis described above, always showed the strong correlation between synonymous substitution rates and frequencies of CpG dinucleotides. Since methylated CpG dinucleotide has been known as a mutable site in mammals, we finally concluded that at least in mammals, intragenic variation of synonymous substitutions is caused mainly by a nonrandom mutation due to the methylation of CpG dinucleotides rather than by functional constraints.

Note that it is also possible that functional constraints of the base composition cause intragenic variation of synonymous substitution rates. This is because two gene pairs showed significant correlations between Ps and GC3% when we used a lower significance level. However, we suggest that these correlations can be explained by a nonrandom mutation due to the methylated CpG dinucleotides in the following observations. We first observed that two gene pairs having significant correlations between Ps and GC3% always showed ‘negative’ correlations (Table 1). On the other hand, C1G2, C2G3, and C3G1 were always shown to have ‘positive’ correlations for the gene pairs having statistically significant correlations with Ps at a lower level (Table 1). Therefore, the opposite correlations of these measures with Ps lead to the possibility that synonymous substitutions can be frequently observed at the codon sites having CpG dinucleotides and, at the same time, GC3% is reduced at the codon sites. Indeed, we observed this possibility in the gene pair in Fig. 5. Among 13 C3G1 codon pairs having synonymous changes, the most frequently observed substitution is from C3G1 to T3G1. Moreover, the most frequently observed substitution is from C2G3 to C2A3 among 12 C2G3 codons pairs with synonymous changes. Thus, synonymous substitutions at the codon pairs having CpG dinucleotides make the cause of reduction of GC3%. Therefore, intragenic variation of synonymous substitutions is mainly caused by a nonrandom mutation due to the methylation of CpG dinucleotides rather than by functional constraints of the base composition.

Discussion

As mentioned before, it has been observed that synonymous substitution rates were correlated positively with nonsynonymous substitution rates at the intergenic level in various organisms (Graur 1985; Li et al. 1985; Britten 1986; Wolfe et al. 1989; Bernardi et al. 1993; Wolfe and Sharp 1993; Mouchiroud et al. 1995; Ohta and Ina 1995; Comeron and Kreitman 1998). At the intergenic level, it was reported that these rates were also correlated with each other in mammalian genes (Alvarez-Valin et al. 1998). As shown in Table 1, we also observed the positive correlation between Ps and Pn. Thus, the significant correlation between synonymous and nonsynonymous substitution rates can be observed at both intragenic and intergenic levels. These observations can be explained by nonrandom mutations at methylated CpG sites. In fact, all substitutions from methylated CpG to TpG or CpA do not always cause synonymous substitutions. For example, at the codon sites having C1G2 dinucleotides, the substitutions from C1G2 to T1G2 and from C1G2 to C1A2 cause nonsynonymous substitutions. When nonrandom mutations at methylated CpG dinucleotides occur frequently, they cause more numbers of nonsynonymous substitutions as well as synonymous substitutions. Thus, the significant positive correlation between Ps and C1G2, as shown in Table 1, can be explained by nonrandom mutations at the methylated CpG dinucleotides. Similarly, the substitutions at C2G3 and C3G1 can cause both synonymous and nonsynonymous substitutions. Therefore, the genes or the intragenic regions having more numbers of methylated CpG sites are likely to have more numbers of both synonymous and nonsynonymous substitutions.

Although our correlation analysis suggests, at this stage, that intragenic variation of synonymous substitutions in mammalian genes is caused mainly by a nonrandom mutation due to the methylation of CpG dinucleotides rather than by functional constraints, the number of available data was quite limited especially in the analysis of the bias of codon usage and the secondary structure of mRNAs. Moreover, as for the other organisms, further studies will be needed since the role of methylation and the methylation pattern may be quite different from mammalian genomes having CpG islands. Therefore, detailed analysis with more substantial number of data and extensive analysis for the other organisms are needed to identify minutely the cause of intragenic variation of synonymous substitutions in various organisms.

Acknowledgment. We thank G. Bernardi and J. Felesenstein for helpful discussion, and J. Zhang for sending us a source program of the modified Nei and Gojobori method. This study was supported by a grant from the Ministry of Education, Culture and Science. M. Bellgard is supported by a postdoctoral fellowships of the Japan Society for the Promotion of Science.

References

Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster:* Natural selection and translational accuracy. Genetics 136:927–935

Alvarez-Valin F, Jabbari K, Bernardi G (1998) Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. J Mol Evol 46:37–44

Bernardi G, Mouchiroud D, Gautier C (1993) Silent substitutions in mammalian genomes and their evolutionary implications. J Mol Evol 37:583–589

Bird AP (1993) Functions for DNA methylation in vertebrates. Cold Spring Harbor Symp Quant Biol 58:281–285

Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393–1398

Cacciò S, Zoubak S, D'Onofrio G, Bernardi G (1995) Nonrandom frequency patterns of synonymous substitutions in homologous mammalian genes. J Mol Evol 40:280–292

Comeron JM, Aguadé M (1996) Synonymous substitutions in the *Xdh* gene of Drosophila: heterogeneous distribution along the coding region. Genetics 144:1053–1062

Comeron JM, Aguadé M (1998) An evaluation of measures of synonymous codon usage bias. J Mol Evol 47:268–274

Comeron JM, Kreitman M (1998) The correlation between synonymous and nonsynonymous substitutions in Drosophila: mutation, selection or relaxed constraints? Genetics 150:767–775

Eyre-Walker A, Bulmer M (1993) Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res 21: 4599–4603

Graur D (1985) Amino acid composition and the evolutionary rates of protein-coding genes. J Mol Evol 22:53–62

Holiday R, Grigg GW (1993) DNA methylation and mutation. Mutat Res 285:61–67

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codons choice that is optimal for the *E. coli* translational system. J Mol Biol 151:389–409

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 transfer RNAs. J Mol Biol 158:573–598

Ina Y (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J Mol Evol 40:190–226

Ina Y, Mizokami M, Ohba K, Gojobori T (1994) Reduction of synonymous substitutions in the core protein gene of hepatitis C virus. J Mol Evol 38:50–56

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

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

King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798

Lawrence JG, Hartl DL, Ochman H (1991) Molecular considerations in the evolution of bacterial genes. J Mol Evol 33:241–250

Li W-H, Wu C-I, Luo C-C (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Moriyama EN, Gojobori T (1992) Rates of synonymous substitution and base composition of nuclear genes in Drosophila. Genetics 130:855–864

Mouchiroud D, Gautier C, Bernardi G (1995) Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J Mol Evol 40: 107–113

Nei M, Gojobori T (1986) Simple methods for estimating the number of synonymous and nonsynonymous substitutions. Mol Biol Evol 3:418–426

Ohta T, Ina Y (1995) Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. J Mol Evol 41:717–720

Powell JR, Moriyama EN (1997) Evolution of codon usage bias in Drosophila. Proc Natl Acad Sci USA 94:7784–7790

Sharp PM, Li W-H (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res 14:7737–7749

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

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in

Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Smith DB, Simmonds P (1997) Characteristics of nucleotide substitution in the hepatitis C virus genome: constraints on sequence change in coding regions at both ends of the genome. J Mol Evol 45:238–246

Sokal RR, Rohlf FJ (1969) Biometry, 3rd ed. W. H. Freeman and Co, New York

Tateno Y, Ikeo K, Imanishi T, Watanabe H, Endo T, Yamaguchi Y, Suzuki Y, Takahashi K, Tsunoyama K, Kawai M, Kawanishi Y, Naitou K, Gojobori T (1997) Evolutionary motif and its biological and structural significance. J Mol Evol 44 (Suppl 1):S38–S43

Ticher A, Graur D (1989) Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein-coding genes. J Mol Evol 28:286–298

Wolfe KH, Sharp PM (1993) Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J Mol Evol 37:441–456

Wolfe KH, Sharp PM, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Wright F (1990) The ‘effective number of codons’ used in a gene. Gene 87:23–29

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA 95:3708–3713

Zoubak S, D’Onofrio G, Cacciò S, Bernardi G, Bernardi G (1995) Specific compositional patterns of synonymous positions in homologous mammalian genes. J Mol Evol 40:293–307

Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. Science 244:48–52