J Mol Evol (1998) 46:37–44

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 1998

Synonymous and Nonsynonymous Substitutions in Mammalian Genes: Intragenic Correlations

Fernando Alvarez-Valin,* Kamel Jabbari, Giorgio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France

Received: 10 July 1997 / Accepted: 13 August 1997

Abstract. Previous investigations indicated that synonymous and nonsynonymous substitution rates are correlated in mammalian genes. In the present work, this correlation has been studied at the intragenic level using a dataset of 48 orthologous genes from species belonging to at least four different mammalian orders. The results obtained show that the intragenic variability in synonymous rates is correlated with that of nonsynonymous rates. Moreover, the variation in GC level (and especially of C level) of silent positions along each gene is correlated with the variation in synonymous rate. These results reinforce the previous conclusions that synonymous and nonsynonymous rates as well as GC levels of silent positions are to some extent under common selective constraints.

Key words: Intragenic correlations — Mammalian genes — Substitutions

Introduction

It is well known that the rate of nonsynonymous substitutions is extremely variable among genes, as revealed a long time ago by investigation of the rates of amino acid substitution (Dickerson 1971; Dayhoff 1972). Indeed, rates range, in a human/murid comparison (Li 1997), from zero (in the genes for histones 3 and 4) to 3.06 × 10^{-9} substitutions per site per year (in the gene for interferon gamma).

The average for the small set of genes taken into consideration by Li (1997) is 0.74 (±0.67) × 10^{-9} substitutions per site per year. The more than 300-fold range in nonsynonymous substitution rates found in different genes (neglecting those showing no substitution) may be ascribed to differences in the rate of mutation and/or to the intensity of selection. As for the first possibility, the difference in mutation rate is not nearly large enough to account by itself for the phenomenon under consideration. This conclusion can be drawn from the fact that synonymous mutation rates for mammalian genes cover a range which is much smaller than that of nonsynonymous substitutions (see below). The most important factor in determining the rate of nonsynonymous substitution appears, therefore, to be the selection intensity (Li and Graur 1991). The higher the deleterious effects of amino acid replacement, the higher the chances of the mutation being eliminated by negative selection (advantageous mutations being very rare compared to deleterious mutations, they are neglected here). Indeed, negative selection is indicated by the actual amino acid changes observed, functionally crucial amino acids never being replaced and conservative substitutions being predominant over nonconservative ones.

As far as the rate of synonymous substitutions is concerned, the average value of 3.51 (±1.01) × 10^{-9} per site per year reported for 47 genes from human and rodents (Li 1997) is misleading in that it conveys the idea of an average rate not very different from that reported for

Permanent address: Sección Biomatemática, Facultad de Ciencias, Universidad de la República, Tristan Narvaja 1674, Montevideo, Uruguay
Correspondence to: G. Bernardi

pseudogenes, thus reinforcing the idea (see Wolfe et al. 1989) that the synonymous substitution rate is close to the mutation rate. In fact, a larger set of over 300 human/murid genes has shown a 20-fold range of synonymous rates (Bernardi et al. 1993; Wolfe and Sharp 1993). This immediately raises a question: If the fastest rate observed is equal to the mutation rate, what is the cause for the 20-fold slower rate exhibited by other genes?

An answer to this question has come from investigations on the frequency and the compositional patterns of synonymous substitutions in orthologous genes from four mammalian orders (Cacciò et al. 1995; Zoubak et al. 1995). These studies showed (1) that the frequencies of conserved, intermediate, and variable positions (defined as the positions showing no change, one change, or more than one change, respectively) of quartet and duet (fourfold and twofold degenerate) codons are different in different genes; (2) that the frequencies of the three classes are significantly different (especially for GC-rich genes) from expectations based on a random substitution process in the majority of genes for quartet codons, and in a minority of genes for duet codons; and (3) that the frequencies of the three classes of positions of quartet codons are correlated with those of duet codons, the conserved positions of quartet and duet codons being, in addition, correlated with the degree of amino acid conservation. Moreover, in the majority of GC-rich genes, the three classes of positions (but especially the conserved positions) exhibited significantly different base compositions compared to expectations based on a ‘‘random’’ substitution process from the ‘‘ancestral’’ (consensus) sequence to the present-day (actual) sequences, whereas significant differences were rare in GC-poor genes.

In the present work we have reinvestigated, at the intragene level, two problems, concerning the correlations of synonymous substitutions with nonsynonymous substitutions and with the base composition of synonymous positions, respectively. The solution of these two problems is of crucial importance for our understanding of the compositional conservation of synonymous codon positions during mammalian evolution and for the broader issue of the role played by selection on synonymous positions.

Sequence Dataset and Analysis

The analysis was performed on 48 orthologous mammalian coding sequences, a subset of the 69 sequences previously studied (Caccio et al. 1995) from which short genes (less than 180 codons) and extremely conserved sequences (more than 98% conservation at the amino acid level) were excluded. In most cases each alignment comprised four different mammalian orders (primates, artiodactyls, lagomorphs, and rodents), but in several cases alignments comprised sequences from five orders. In no case was more than one sequence belonging to a species of the same order included in the analysis. The alignments used are available via the Internet:

http://genetica.edu.uy/mol_evol/mammals_aln
ftp://genetica.edu.uy/mammals_aln

The variations in substitution rates along the coding sequences were determined by using a sliding window. Pairwise synonymous and nonsynonymous nucleotide distances were estimated by Nei and Gojobori’s method (1986). Two profiles were thus obtained, representing synonymous and nonsynonymous divergence, respectively, where each point corresponds to the average of all pairwise distances for each window. Both nonoverlapping windows and overlapping windows (shifting by one codon at at time) were used. The latter were not used for any of the statistical calculation because they are not mutually independent, but were used for graphical purposes, because they produce smoother profiles.

Since relatively short (20 to 30 codons) segments of coding sequences are used in the window analyses, distance estimations are subject to large stochastic errors. For this reason, the window size cannot be excessively small (less than 20 codons). Yet it cannot be too large either, because, since only nonoverlapping windows are considered, the numbers of points to be compared is then too small, especially in short genes. Moreover, when the distance is too large the estimation is inaccurate or even incalculable (Nei 1987). Indeed, in genes having an average synonymous distance of one substitution per synonymous site (or more), some individual windows could reach values that are beyond the limits of the estimation method. The degree of similarity between the synonymous and nonsynonymous profiles was determined using Pearson’s correlation coefficient.

Results and Discussion

Correlations Between Synonymous and Nonsynonymous Distances

Table 1 lists the coding sequences used in this work and their sizes, GC_3 levels (in the human genes), and the correlation between synonymous and nonsynonymous distance profiles of these sequences. The great majority of coding sequences, 73%, showed a positive correlation coefficient. Only a minority of coding sequences, 27%, exhibited a negative correlation, but in no case was the negative correlation statistically significant. Figure 1 shows that the distribution is highly biased toward positive values. Among the genes displaying positive correlations, some were so high that they were obvious even by visual inspection. Figure 2A to D shows synonymous and nonsynonymous distance profiles for four of the highly correlated coding sequences. They indicate that the regions that have little divergence at the amino acid level are also more conserved at the synonymous level, whereas the regions that are less conserved at the amino acid level are more divergent at the synonymous level.

Concerning the statistical significance of the correlations observed, it should be stressed that 16 out of 48 genes (33%) displayed significant *r* values. As already mentioned, in all cases these correspond to positive values. In four genes the correlations were significant at the 0.1% level, in four genes at the 1% level, and in eight genes at the 5% level.

It should be noted that when several cases are con-

Table 1. Coding sequences investigated, their sizes, GC_3 levels, and correlations between synonymous and nonsynonymous rates [a]

Gene name	L	GC_3	SY-NSY	
Creatin kinase B	381	0.90	0.29	
Apolipoprotein E	305	0.88	0.27	
A1 adenosine receptor	326	0.87	0.78	**
Apolipoprotein A1	257	0.84	-0.27	
Na-H exchange protein	810	0.83	0.64	***
Serine pyruvate aa transferase	392	0.83	0.38	
CD8 alpha chain	227	0.83	0.04	
GMP-phosphodiesterase alpha	852	0.82	0.02	
Dipeptidase	411	0.82	0.39	
Prostaglandin E receptor	335	0.81	0.15	
Glutathione peroxidase	197	0.80	-0.33	
Retinol binding protein	199	0.80	0.00	
Creatin kinase M	382	0.79	0.65	**
H,K ATPase beta subunit	290	0.78	–0.17	
TNFalpha	231	0.78	0.69	
Glucose Glut3	492	0.78	–0.14	
Growth hormone	214	0.75	0.71	*
Prolyl-4-hydroxylase beta	502	0.74	–0.16	
TNF beta	193	0.71	–0.33	
Polymeric Ig receptor	740	0.70	0.66	***
Erythropoietin	185	0.70	–0.30	
Na Glucose transporter	602	0.68	0.68	***
CD4 antigen	444	0.68	0.59	*
D-amino acid oxidase	346	0.64	0.17	
tRNA ligase	470	0.63	0.35	
Endothelin	201	0.62	0.79	*
Interleukin 1B	259	0.60	0.68	*
Interleukin 6 receptor	203	0.58	0.46	
Interleukin 2 receptor	265	0.56	0.78	*
Phagocytic glycoprotein I	353	0.54	0.26	
Urate oxidase	299	0.53	0.36	
Prolactin receptor	554	0.52	0.64	**
Na-K ATPase beta-1 subunit	303	0.51	0.83	**
CD3 epsilon antigen	187	0.51	0.78	*
Tissue factor	282	0.51	0.47	
Selectin	412	0.49	0.12	
Interleukin 1A	263	0.49	0.40	
Flavin–containing monooxygenase	532	0.49	–0.11	
Link protein	354	0.48	–0.48	
Osteopontin	253	0.41	0.22	
Pancreatic triglyceride lipase	460	0.41	–0.01	
Serum albumin	607	0.40	0.68	***
Stem cell factor/Kit ligand	273	0.39	0.47	
HSP 108	802	0.37	0.47	*
Macrophage scavenger	449	0.37	–0.43	
Apolipoprotein H	345	0.36	–0.17	
Calpastatin	593	0.34	–0.10	
Ca-ATPase	1176	0.29	0.33	*

[a] L: gene length (in codons); SY-NSY: correlation between intragenic rates of synonymous and nonsynonymous substitutions
* significant at the 5% level
** at the 1% level
*** at the 0.1% level

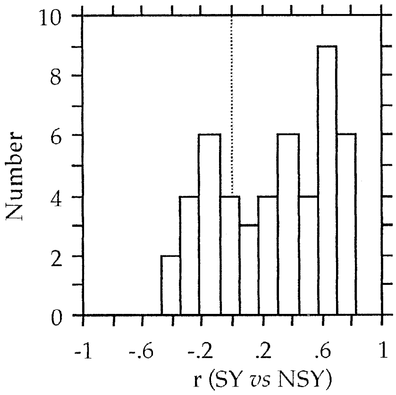

Fig. 1. Distribution of correlation coefficients between synonymous and nonsynonymous profiles.

sidered at the same time, some correlation coefficients are expected to be significant just by chance. In fact, for 48 genes we expect 2.4 coefficients (positive and negative) to be significant at or below the 5% level, 1.92 of these coefficients to be significant only at the 5% level, 0.43 coefficients at the 1% level, and 0.05 at the 0.1% level. Based on the probabilities for each level of significance and using the multinomial distribution [0.04, 0.009, 0.001, 0.95)^48], the probability of obtaining Table 1 and all less likely tables by chance can be calculated, giving the value 6.97 × 10^{-13}, the probability dropping to 2.06 × 10^{-17} if it is taken into account that all significant coefficients are positive. Therefore, even though only one-third of the genes display significant correlations, the extremely low probability that such a proportion could have been the result of random effects indicates that synonymous and nonsynonymous divergences are indeed related. It is noteworthy that there are several examples of coding sequences showing correlations at the limit of significance (e.g., TNF-alpha). The fact that these coding sequences are in general relatively short suggests that, in several cases, the lack of significance is due to the small number of codons. This suggestion is further supported by the fact that if one considers subsets of the coding sequences of progressively increasing size, the proportion of coding sequences with significant correlations increases, reaching a value of 64% (7/11) in the subset of coding sequences that have at least 500 codons in length (Table 2). The association between the size of the coding sequences and the significance of the correlations is statistically significant (2 × 2 chi-square contingency table, $P < 0.05$).

Intragenic Correlation Between Synonymous Substitution Rates and GC Levels in Synonymous Positions

In order to investigate the intragenic correlation between variation in GC_3 (excluding Met and Trp codons) and variation in substitution rate, GC_3 levels were measured along the coding sequences using a sliding window. For each alignment, we calculated the correlation coefficient between the substitution rate profiles with the mean GC_3 profile (GC_3 averaged over all species for each window). Table 3 summarizes the results of these analyses as well

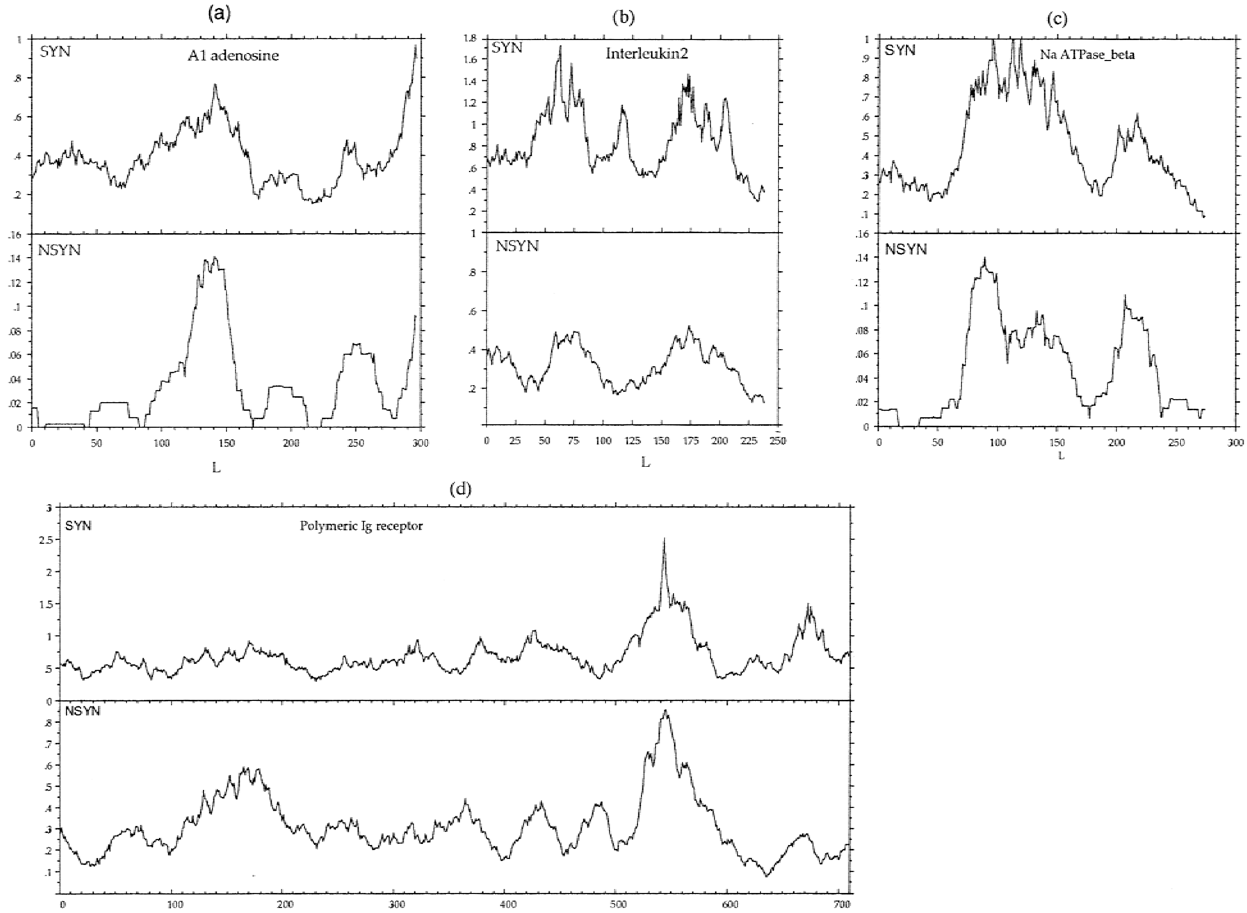

Fig. 2. Four examples of genes displaying correlations between synonymous and nonsynonymous rates. **A** A1-adenosine, **B** Interleukin 2 receptor, **C** Na-glucose transporter, **D** Polymeryc Ig receptor.

Table 2. Subsets of genes of progressively increasing size, and the proportion exhibiting significant intragenic correlations between synonymous and nonsynonymous rates

Genes size	Number of genes	Number of genes showing significant correlations	Percentage
>180	48	16	33%
>300	30	11	37%
>400	18	8	44%
>500	11	7	64%

as the behavior of C_3 and G_3 taken separately and shows that almost all GC_3-rich coding sequences exhibit negative correlations between the profiles of GC_3 and those of synonymous rates, these correlations being significant in several cases.

Conversely, for the genes with lower GC_3 (<60%), the proportion of negative correlations drastically decreases. In addition, near the GC_3-poor end of the distribution, there are two examples of significant positive correlations. The number of genes displaying significant correlations is much higher than random expectation (multinomial distribution, $P = 1.9 \times 10^{-9}$). This inverse relationship between the correlation coefficients of synonymous rate and GC_3 and the profiles of GC_3 is evident in Fig. 3A. (The inclusion of mouse genes in the sample studied detracts from the quality of the correlation because mouse GC-rich and GC-poor genes are remarkably less GC-rich and GC-poor, respectively, than those of the other mammals investigated; Mouchiroud et al. 1988.)

There are two possible, not mutually exclusive, interpretations for this inverse relation. The first one is that in GC_3-rich coding sequences most synonymous substitutions would be expected to be from G or C to A or T. Therefore those regions of the coding sequence that underwent higher synonymous divergence are at the same time the regions that should have suffered a higher reduction in their GC_3 level. By contrast, in GC_3-poor coding sequences, most synonymous substitutions are expected to be from A or T to G or C, so in these coding sequences a higher synonymous divergence would be accompanied by a higher increase in their GC_3 level. Therefore, according to this interpretation, this inverse relation is just the somewhat predictable result of the divergence process. The second interpretation is that in GC_3-rich genes the segments with high GC_3 levels, and in the GC_3-poor genes the GC_3-poor segments, tend to be

Table 3. Intragenic correlations between synonymous base compositions and the profiles of synonymous and nonsynonymous rates

	Correlation between synonymous profile and			Correlation between nonsynonymous profile and
Gene name	GC_3	C_3	G_3	GC_3
Creatin kinase B	-0.57 *	-0.25	-0.07	0.10
Apolipoprotein E	-0.38	-0.33	0.20	-0.31
A1 adenosine receptor	-0.81 **	-0.35	-0.10	-0.35
Apolipoprotein A1	-0.24	-0.33	0.04	0.48
Na-H exchange protein	-0.67 ***	-0.58 **	0.02	-0.66 ***
Serine pyruvate aa transferase	-0.24	-0.40	0.29	-0.35
CD8 alpha chain	-0.34	-0.13	-0.11	0.81 *
GMP-phosphodiesterase alpha	-0.54 **	-0.30	-0.16	0.02
Dipeptidase	-0.30	-0.61 *	0.57 *	-0.26
Prostaglandin E receptor	0.30	0.44	-0.10	0.06
Glutathione peroxidase	-0.54	0.10	-0.26	0.85 *
Retinol binding protein	-0.18	0.17	-0.25	0.20
Creatin kinase M	-0.60 **	0.05	-0.46 *	-0.15
H,K ATPase beta subunit	-0.72 *	-0.33	-0.10	-0.10
TNFalpha	-0.61	-0.82 *	0.37	-0.80 *
Glucose Glut3	-0.73 **	-0.08	-0.61 *	0.38
Growth hormone	-0.47	-0.59	0.12	-0.32
Prolyl-4-hydroxylase beta	-0.85 ***	-0.61 **	-0.32	0.35
TNF beta	0.27	0.68	-0.55	-0.87 *
Polymeric Ig receptor	-0.42 *	-0.43 *	0.13	-0.45 *
Erythropoietin	0.21	-0.27	0.25	-0.24
Na Glucose transporter	-0.40	-0.22	-0.10	-0.27
CD4 antigen	-0.11	0.04	-0.12	-0.27
D-amino acid oxidase	0.21	-0.04	0.31	-0.42
tRNA ligase	-0.55 *	-0.35	-0.25	-0.54 *
Endothelin	-0.11	-0.03	-0.13	-0.04
Interleukin 1B	-0.12	-0.17	0.07	-0.08
Interleukin 6 receptor	0.00	-0.45	0.43	0.12
Interleukin 2 receptor	-0.74 *	-0.44	-0.54	-0.71 *
Phagocytic glycoprotein I	0.00	0.37	-0.32	-0.35
Urate oxidase	-0.59	-0.46	-0.64 *	-0.14
Prolactin receptor	0.36	-0.08	0.56 *	-0.04
Na-K ATPase beta-1 subunit	-0.32	-0.15	-0.36	-0.32
CD3 epsilon antigen	-0.45	-0.47	-0.32	-0.79 *
Tissue factor	0.34	0.13	0.52	0.93 ***
Selectin	-0.01	-0.18	0.26	-0.52
Interleukin 1A	-0.18	-0.06	-0.20	-0.21
Flavin-containing monooxygenase	-0.28	-0.23	-0.14	-0.54 *
Link protein	0.48	0.33	0.37	-0.51
Osteopontin	-0.08	0.73 *	-0.49	0.36
Pancreatic trigyceride lipase	-0.01	0.24	-0.26	0.33
Serum albumin	0.36	0.44	0.02	0.00
Stem cell factor/Kit ligand	-0.13	-0.51	0.61	0.26
HSP 108	-0.01	-0.02	0.01	-0.02
Macrophage scavenger	0.66 **	0.70 **	0.38	-0.33
Apolipoprotein H	0.21	0.15	0.21	-0.25
Calpastatin	0.41	0.40	0.06	0.04
Ca-ATPase	0.41 **	0.31 *	0.23	0.05

*,**,*** significances as in Table 1

more conserved at the synonymous level, reflecting some functional constraints that act toward maintaining GC_3 richness in the GC_3-rich segments of GC_3-rich genes and GC_3 poorness in the GC_3-poor segments of GC_3-poor genes.

In order to investigate further this relation for all coding sequences in which the correlation between synonymous profile and GC_3 profile was significant, we compared the profile of synonymous divergence in a pair of species vs the GC_3 profile of a third species. Since all species belong to different orders and since a star phylogeny among the mammalian orders investigated can be assumed for the present purpose, no correlation would be expected between these profiles if the first interpretation

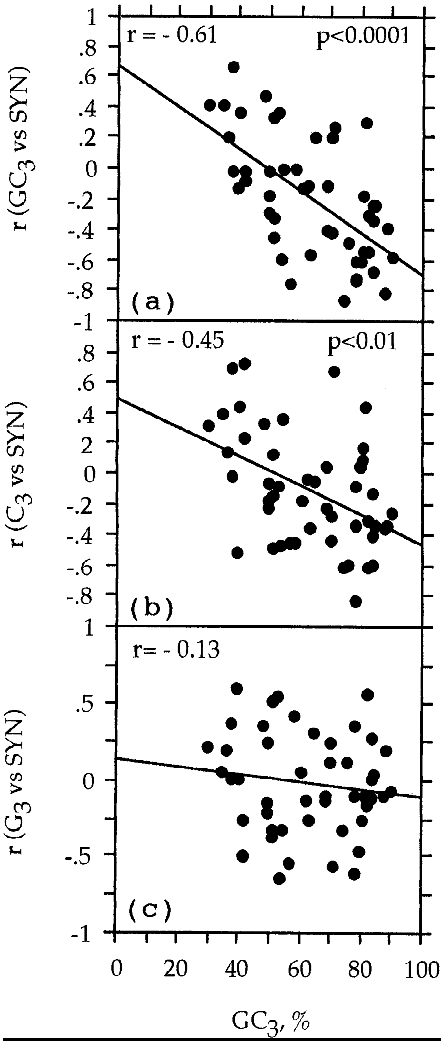

Fig. 3. Scatterplot of the correlation coefficients between the profiles of GC_3 **(A)**, C_3 **(B)**, G_3 **(C)** and synonymous distances vs the total GC_3 level of the gene.

were correct. Indeed, if the process of synonymous divergence were random and spatially homogeneous, then there would be no reason to expect spatial correlation between GC_3 of the common ancestor, or of the third species, and the synonymous substitution pattern between the two other species. The analyses show (Table 4), however, that this is not the case, since in all comparisons the correlations were significant and with the same sign as in Table 3. The only exception was the gene encoding for glucose Glut3, which exhibited a correlation at the limit of significance ($P = 0.06$). Therefore, even though the first process may contribute to the correlations presented in Table 3, it seems clear that constraints on synonymous sites are needed to explain the results presented in Table 4.

An analysis similar to that of GC_3 and synonymous rate was carried out independently for C_3 and G_3 in order to understand the behavior of these bases (Table 3). While the general trend for C_3 is similar to that described above for GC_3 (Fig. 3B), the behavior of G_3 appears to be less strongly related (Fig. 3C). In addition, the number of genes displaying significant correlations between G_3 profiles and synonymous rate profiles is not significantly higher than random expectation (binomial distribution, $P = 0.22$). In contrast with this, the number of significant correlations observed between C_3 and synonymous profiles is higher than random expectation (multinomial distribution, $P = 5.3 \times 10^{-3}$). This results leads to the conclusion that the relation described above between GC_3 and synonymous rate depends mostly upon C-ending codons.

Finally, the correlation between GC_3 profiles and nonsynonymous profiles was also analyzed (Table 3). Since no correlation would be expected a priori between these profiles, it is surprising that the number of coding sequences showing significant correlations is much higher than random expectation (multinomial distribution, $P = 1.13 \times 10^{-6}$). As in the case concerning synonymous profiles and GC_3, the majority of the genes display negative correlations. In contrast to the correlations between profiles of synonymous distances and GC_3, in this case negative correlations are not preferentially associated with GC_3-rich genes. This lack of a parallel behavior is rather puzzling, considering that the only link between nonsynonymous profiles and GC_3 could be the dependence between synonymous and nonsynonymous profiles already described. However, a closer examination of this discrepancy (Fig. 4) leads to the observation that the correlation between nonsynonymous profile and GC_3 profile depends both on the correlation of synonymous vs nonsynonymous profiles and on the correlation between synonymous vs GC_3 profile. In fact, in those genes for which synonymous rate is positively correlated with nonsynonymous rate, the correlation between nonsynonymous and GC_3 profiles is of the same type as that between synonymous and GC_3 profiles, while the opposite is found in genes that show a negative covariation between synonymous and nonsynonymous profile.

Conclusions

Several authors have reported that the synonymous and the nonsynonymous rates for entire genes are correlated (Li et al. 1985; Ticher and Graur 1989; Wolfe and Sharp 1993; Mouchiroud et al. 1995; Ohta and Ina 1995). In this work we have investigated this correlation at the intragenic level. It is worth noting that although the main problems tackled in this paper were already investigated in previous work from our laboratory (Bernardi et al. 1993; Mouchiroud et al. 1995; Cacciò et al. 1995; Zoubak et al. 1995), the approach used differs in a very important respect, namely in that here we measured the

Table 4. Correlation between the profile of synonymous divergence in a pair of species and the GC_3 profile from a third species

Gene name	Divergence between	CG_3 profile from	Correlation[a]	
Creatin kinase B	Human-rabbit	Dog	-0.68	*
A1-adenosine receptor	Human-rat	Calf	-0.62	*
Na-H exchange protein	Pig-rabbit	Human	-0.47	*
GMP-phosphodiesterase alpha	Human-dog	Calf	-0.57	**
Creatin kinase M	Human-rabbit	Dog	-0.73	***
H,K ATPase beta subunit	Human-pig	Dog	-0.68	*
Glucose Glut3	Rat-calf	Human	-0.47	
Prolyl-4-hydroxylase beta	Human-rat	Calf	-0.56	*
Polymeric Ig receptor	Calf-rabbit	Rat	-0.56	**
tRNA ligase	Calf-rabbit	Human	-0.78	***
Interleukin 2 receptor	Human-cat	Mouse	-0.76	*
Macrophage scavenger	Human-calf	Rabbit	0.62	*
Ca-ATPase	Human-sheep	Rabbit	0.31	*

[a] Asterisks refer to significances as in Table 1

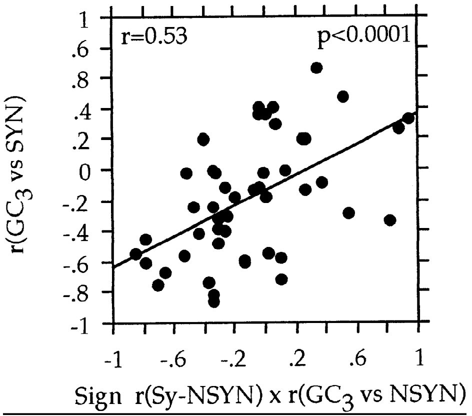

Fig. 4. Scatterplot of the correlation between the profiles of GC_3 and synonymous rates vs the correlation between the profiles of GC_3 and nonsynonymous distances multiplied (in each gene) by the sign of the correlation between nonsynonymous profile and synonymous profile.

substitution rates along the coding sequences instead of the average distance for entire genes or the average distance for all quartet (fourfold degenerate) and duet (twofold degenerate) codons taken separately. Moreover, the spatial distribution of substitutions was taken into account since distances were determined here by using a sliding window. This approach allowed us to investigate the intragenic variability in substitution rates and thereby to show that the spatial pattern of synonymous substitutions is not random but related to that of nonsynonymous substitutions. Considering that the spatial pattern of amino acid conservation is generally accepted to reflect the effect of negative selection for maintaining functionally important amino acids (see Kimura 1991), the intragenic correlations described here strongly reinforce our previous conclusion that the processes of synonymous and nonsynonymous divergence are, to some extent, under common selective constraints.

It has been previously suggested that the correlation between silent and amino acid replacement rates might be the result of either doublet mutations (Wolfe and Sharp 1993) or nonhomogeneous mutational rates along the genome (Wolfe et al. 1989). The first possibility has been already ruled out by Mouchiroud et al. (1995) by excluding from the analyses those codons positions that could have undergone doublet mutations. Concerning the second possibility, Ohta and Ina (1995) have shown that the variability in mutation rates cannot by itself explain the correlation unless mutation rate and selective constraint are negatively correlated (Ina 1995). That is, however, an ad hoc hypothesis for which no justification has been provided. A serious complication associated with this hypothesis is that it implies that not only the intragenic variability in synonymous rates, but also in nonsynonymous rates, would be the result of different mutational rates; otherwise the correlation could not appear. According to this view, it is then evident that, contrary to what is generally accepted, the intragenic spatial pattern of amino acid conservation would reflect the variation in mutational rate rather than the effect of negative selection.

It is therefore more likely that the intragenic correlations described here rely on common selective constraints acting on synonymous and nonsynonymous divergence processes. One possible cause of the existence of such correlated constraints between synonymous and nonsynonymous mutations might be that it is related to selection for translational accuracy. Indeed, it has been shown that in *Drosophila* genes the frequency of optimal codons is higher in conserved (and thus putatively functionally important) amino acids than in nonconserved ones, and this has been attributed to the effect of selection for increasing translational accuracy (Akashi 1994). Thus, the lower synonymous rates described here for conserved amino acids could be the result of negative selection for maintaining codon biases in sites conserved at the amino acid level. This interpretation is further supported by the fact that the intragenic variability in synonymous rates is correlated with the variability in GC_3 content. As stated above, this correlation in all likelihood reflects functional constraints that act toward maintaining GC_3 levels (and in particular C_3 levels) in the gene segments that evolve more slowly.

Acknowledgements. One of us (F.A.V.) was the recipient of a UNESCO-TWAS Fellowship.

References

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

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

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

Dayhoff MO (1972) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington, DC

Dickerson RE (1971) The structure of cytochrome c and the rates of molecular evolution. J Mol Evol 1:26–45

Ina Y (1995) Correlation between synonymous and nonsynonymous substitutions and variation in synonymous substitution numbers. In: Nei M, Takahata N. Current topics on Molecular evolution. Institute of Molecular Evolutionary Genetics, Penn State University

Kimura M (1991) Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. Proc Natl Acad USA 88:5969–5973

Li W-H (1997) Molecular evolution. Sinauer, Sunderland, MA

Li W-H, Graur D (1991) Fundamentals of molecular evolution. Sinauer, Sunderland, MA

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

Mouchiroud D, Gautier C, Bernardi G (1988) The compositional distribution of coding sequences and DNA molecules in man and murids. J Mol Evol 27:311–320

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

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

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

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

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

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