Synonymous Substitutions in the *Xdh* Gene of Drosophila: Heterogeneous Distribution Along the Coding Region

Josep M. Comeron and Montserrat Aguadé

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain

Manuscript received January 25, 1996
Accepted for publication July 25, 1996

ABSTRACT

The *Xdh* (*rosy*) region of *Drosophila subobscura* has been sequenced and compared to the homologous region of *D. pseudoobscura* and *D. melanogaster*. Estimates of the numbers of synonymous substitutions per site (*Ks*) confirm that *Xdh* has a high synonymous substitution rate. The distributions of both nonsynonymous and synonymous substitutions along the coding region were found to be heterogeneous. Also, no relationship has been detected between *Ks* estimates and codon usage bias along the gene, in contrast with the generally observed relationship among genes. This heterogeneous distribution of synonymous substitutions along the *Xdh* gene, which is expression-level independent, could be explained by a differential selection pressure on synonymous sites along the coding region acting on mRNA secondary structure. The synonymous rate in the *Xdh* coding region is lower in the *D. subobscura* than in the *D. pseudoobscura* lineage, whereas the reverse is true for the *Adh* gene.

SYNONYMOUS substitutions were initially thought to be nearly neutral (KIMURA 1968; KING and JUKES 1969) as they do not affect the amino acid sequence of proteins. KIMURA (1983) later suggested that "even synonymous substitutions are subject to negative selection, although the intensity of selection involved must be exceedingly weak . . . and similar for different genes within a genome" (pp. 314–315). Nevertheless, high variability in estimated numbers of synonymous substitutions per site (*Ks*) has been reported frequently in Drosophila (SHIELDS *et al.* 1988; RILEY 1989; SHARP and LI 1989; MORIYAMA and GOJOBORI 1992; SEGARRA and AGUADÉ 1993). Moreover, there is a well-known negative relationship between *Ks* and both the nonrandom usage of synonymous codons (codon bias) and the G+C content at synonymous sites (SHIELDS *et al.* 1988; SHARP and LI 1989). Furthermore, in Drosophila, spatial biases in the mutational pattern have been ruled out as a cause of the observed variability of codon bias and synonymous substitution rates among genes (CARULLI *et al.* 1993; MORIYAMA and HARTL 1993). Different selective hypotheses have been proposed to account for these observations. Selective constraints at the translational level on synonymous sites, varying as a function of the level of expression of each gene (SHIELDS *et al.* 1988; MORIYAMA and HARTL 1993), could enhance the accuracy of protein synthesis (KURLAND 1987a; SHARP and LI 1989; AKASHI 1994) within the mutation-selection-drift model (reviewed in BULMER 1991). On the other hand, selection could also act at the mRNA secondary structure level as already suggested for some enterobacterial genes (LAWRENCE *et al.* 1991; EYRE-WALKER and BULMER 1993). The study of longer coding regions could help to ascertain the mode of selection that is acting by allowing analyses of the putative relationship between variability of divergence estimates and codon bias within genes. Such studies, in contrast to multigene comparisons would have the advantage of being independent of variation in both expression level and mutation rate among chromosomal regions.

The *Xdh* gene (*rosy* in *Drosophila melanogaster*), which codes for xanthine dehydrogenase, is one of the longest genes sequenced both in *D. melanogaster* (KEITH *et al.* 1987; LEE *et al.* 1987) and *D. pseudoobscura* (RILEY 1989). In the present study the sequence of the *Xdh* region of *D. subobscura* has been determined and compared to those of *D. pseudoobscura* and *D. melanogaster*. We have estimated both the numbers of synonymous (*Ks*) and nonsynonymous (*Ka*) substitutions per site between the two species of the obscura group as well as between *D. subobscura* and *D. melanogaster*. The length (slightly more than 4 kb) and nucleotide variability of the *Xdh* coding region has also allowed the analysis of the distribution of synonymous substitutions along the gene as well as its relationship to the putative nonrandom use of synonymous codons (codon bias). Moreover, *Ks* estimates in the *Xdh* region have been analyzed in these species in order to study putative lineage effects.

MATERIALS AND METHODS

Sequencing strategy: Several positive recombinant phages were isolated from a random genomic library of *D. subobscura* (AGUADÉ 1988), using as probe a 4.1-kb *Eco*RI-*Hin*dIII fragment with most of the coding region of the *rosy* gene of *D. melanogaster*. After analysis by Southern blot using two different fragments corresponding to the 5′ and 3′ flanking regions

Corresponding author: Montserrat Aguadé, Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Diagonal 645, 08071 Barcelona, Spain. E-mail: aguade@porthos.bio.ub.es

of *D. pseudoobscura* and *D. melanogaster* as probes, one phage that included the complete *Xdh* (*rosy*) region was chosen. A 10-kb *Cla*I fragment containing the whole coding region was subcloned in pBluescriptII and, subsequently, several smaller and overlapping fragments were subcloned. A set of nested deletions was obtained for each strand of each subclone (HENIKOFF 1984). Both strands were completely sequenced by the dideoxy chain-termination method using double stranded DNA. The DNA sequence was assembled using STADEN's programs (1982).

DNA sequence analysis: The Wisconsin package of the Genetics Computer Group (v7.3) (DEVERAUX *et al.* 1984) has been used for interspecific alignments as well as for detection of enzyme-binding domains. Insertions and deletions in interspecific alignments of coding regions have been manually placed to minimize the number of amino acid differences and of insertion/deletion events. The codon bias index (CBI) (MORTON 1993) has been used as a measure of deviation from a uniform use of synonymous codons. This measure exhibits a much lower dependence on the number of codons analyzed and a lower dispersion due to sampling than "scaled χ^2" (Xi/L) (SHIELDS *et al.* 1988), dispersion that in the case of CBI is length independent (J. M. COMERON and M. AGUADÉ, unpublished results). Neither CAI (SHARP and LI 1987) nor FOP (SHIELDS *et al.* 1988) indexes have been used because data available at present for the obscura species are still scanty.

The numbers of synonymous (Ks) and nonsynonymous substitutions (Ka) per site and their confidence intervals were estimated as described by COMERON (1995). This method modifies LI's (1993) and PAMILO and BIANCHI's (1993) method in order to better quantify the number of synonymous substitutions that are actually transitions or transversions as well as to minimize putative stochastic errors.

All correlation probabilities have been calculated by applying the z-transformation (z^*) suggested by HOTELLING (SOKAL and ROHLF 1995; Chapt. 15) for use in small samples.

RESULTS

Interspecific comparison: An 8.8-kb fragment encompassing the *Xdh* gene of *D. subobscura* including more than 1 kb of both its 5′ and 3′ flanking regions has been sequenced (Figure 1). In *D. subobscura,* the coding region (five exons of 57, 2613, 1146, 168 and 48 bp) shows the same structure as in *D. pseudoobscura.* Exons three and four of the obscura group species correspond to the third exon of *D. melanogaster.* When the coding region of *D. subobscura* is compared to those of *D. pseudoobscura* and *D. melanogaster,* it shows two and nine insertion events, respectively, that do not disrupt the reading frame. A total of 1333 codons can therefore be compared among the three species. The two large introns show considerable length differences among the three Drosophila species analyzed. Noncoding regions cannot be generally aligned between *D. subobscura* and *D. melanogaster* and only partly aligned between the more closely related species of the obscura group, *D. subobscura* and *D. pseudoobscura.* When the first intron is analyzed, only the region spanning between nucleotides 1279 and 1432 (Figure 1) can be aligned in both interspecific comparisons. This region, which shows the lowest divergence estimate of all noncoding regions compared (Table 1) and maintains the same spatial position relative to exon one in the three species, is a good candidate for the transcriptional control element, *i409,* described in *D. melanogaster* by LEE *et al.* (1987) and roughly localized in the upstream region of the large first intron by these authors.

The terminal region of the lethal complementation group *l(3)s12* (HILLIKER *et al.* 1980; LEE *et al.* 1987; RILEY 1989) has been detected from position 1 to 213 of the sequenced fragment of *D. subobscura.* The last intron of this transcription unit (COTÉ *et al.* 1986) maintains its position and length in *D. subobscura* relative to *D. melanogaster* and *D. pseudoobscura.* Alignment of the intergenic region between *D. subobscura* and *D. melanogaster* is only possible in the region close to the *l(3)s12* locus spanning to position 382 (Figure 1). In the *D. subobscura* and *D. pseudoobscura* comparison, sequences can be aligned in that region to position 1002. In both cases the region immediately upstream the *Xdh* locus does not exhibit enough similarity to be aligned. On the other hand, the 3′ flanking region of *D. subobscura* can be aligned from position 7841 to the end of the *D. melanogaster* sequence (the stretch that COTÉ *et al.* (1986) describe as not transcribed); for this region no sequence of *D. pseudoobscura* was available for comparison. Table 1 summarizes divergence estimates (Ks and Ka) for coding regions while Table 2 shows divergence estimates (K) for noncoding regions.

Distribution of the number of nucleotide substitutions along the *Xdh* coding region: When studying the distribution of nucleotide substitutions along a coding region, the following aspects have to be considered: (1) comparison of numbers of nucleotide substitutions among exons can produce spurious results when very small exons are included in the analysis, (2) the observed number of nucleotide substitutions is an underestimate of the real number of substitutions unless sequences from very closely related species are compared and, (3) the use of the chi-square test can be inappropriate as this test does not take into account all the different parameters that have an influence on the divergence estimate variances. In order to overcome these problems, we study adjacent segments with the same number of synonymous sites along the coding region; given that the ratio of synonymous/nonsynonymous sites is almost constant along the *Xdh* coding region, the same segments are used for both Ks and Ka heterogeneity analyses. Secondly, Ks and Ka values are tested directly for heterogeneity in the distribution of nucleotide substitutions by applying computer simulation (COMERON 1995). Finally, departure of the observed distribution of divergence estimates from homogeneity has been analyzed by estimating the probability, using two different Monte-Carlo simulations, of obtaining sets of adjacent regions containing values as extreme or more than those observed. Six different segment sizes have been used: 75, 100, 125, 150, 200 and 250 synonymous sites; the number of segments has been adjusted to the total length of the *Xdh* sequence. A given set is considered "heterogeneous" when the generated set

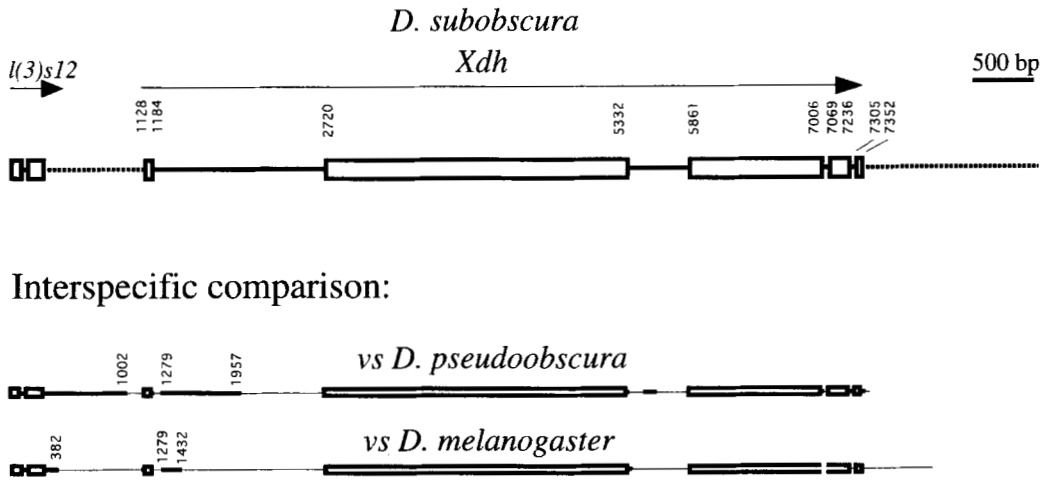

FIGURE 1.—Scheme of the 8.8-kb *Bam*HI-*Hin*dII *Xdh* sequenced region in *D. subobscura* and deposited in EMBL sequence database library under accession no. Y08237. All nucleotide positions in this figure and in the text are referred to this sequence. Arrows above *Xdh* and *l(3)s12* genes indicate the direction of transcription. Boxes indicate exons, solid lines indicate introns and dashed lines indicate intergenic region. In the interspecific comparisons thick solid lines indicate regions where comparison has been possible, while thin lines indicate the rest of regions sequenced in these species. Nucleotide positions are indicated at the beginning and end of exons in *D. subobscura* while in the interspecific comparisons they indicate the extent of aligned noncoding regions.

contains both a higher and a lower estimate than the maximum and minimum observed divergence estimates for the different segment sizes. Note that if divergence estimates were biased, Monte-Carlo results would be similarly biased as divergence values are estimated by the same method. In the first computer simulation, the *Ks* estimate for the global coding region has been used to randomly generate nucleotide substitutions in a particular set of segments. This approach not only takes into account the *Ks* and *Ka* estimates to generate the target sequence but also the numbers of sites and nucleotide composition of the analyzed region (G+C content). Also, the substitution pattern is based on both the estimated transition:transversion ratio and the G+C content in order to maintain these features. Both a Poisson distribution of the expected number of substitutions and a random distribution of these substitutions along the sequence have been assumed. Heterogeneity of *Ka* estimates has been analyzed in a similar way. Ten thousand independent replicates have been performed for each segment size. The results of these Monte-Carlo simulations indicate that while *Ka* estimates show significant heterogeneity ($P < 0.005$) along the coding region of the *Xdh* gene in both interspecific comparisons for all segment sizes, *Ks* estimates only are significantly heterogeneous for the *D. subobscura vs. D. melanogaster* comparison ($P < 0.01$ for segments of 75 and 100 synonymous sites and $P < 0.05$ for segments of 125 and 150 synonymous sites).

In the second simulation, the codon positions of the two observed sequences have been randomized and the sequences subsequently divided into equivalent and contiguous segments. Similar results have been obtained, showing significant heterogeneity in the distribution of *Ka* estimates along the coding region for both interspecific comparisons ($P < 0.005$) for all segment

TABLE 1

Estimates of synonymous and nonsynonymous nucleotide divergence at the *Xdh* region

		Xdh					
	l(3)s12	Exon 1	Exon 2	Exon 3	Exon 4	Exon 5	Total
D. subobscura vs. D. pseudoobscura							
Ks	0.7630	0.3533	0.4867	0.5036	0.6455	1.3735	0.4992
Ka	0.0334	0.0318	0.0335	0.0166	0.0224	0.0517	0.0284
D. subobscura vs. D. melanogaster							
Ks	2.2593[a]	0.8790	1.0099	1.0339	1.3674	—	1.0246
Ka	0.1072	0.0313	0.0859	0.0475	0.0597	0.1118	0.0735
No. of bp compared[b]	267	39	2601	1143	168	48	3999

Ks and *Ka*, synonymous and nonsynonymous nucleotide divergence, respectively.

[a] Divergence estimate obtained by using the JUKES and CANTOR's method (1969) because of the inapplicability of KIMURA's two-parameter method (1980).

[b] The number of base pairs compared differs from the corresponding numbers in the sequence of *D. subobscura* reflecting the 6, 4 and 1 codon insertion events in this species when compared to the other two species.

TABLE 2

Estimates of nucleotide divergence in alignable noncoding regions at the *Xdh* region

	l(3)s12 intron	*Xdh*					
		5′ flanking region		Intron 1		Intron 2[b]	Intron 4
Region compared (position)[a]	45–101	217–382	383–1002	1279–1432	1433–1957	5806–5860	5237–5304
D. subobscura vs. D. pseudoobscura (K)	0.2032	0.3771	0.4867	0.1185	0.3944	0.2764	0.3606
D. subobscura vs. D. melanogaster (K)	0.6831	0.5412	—	0.3613	—	—	—
No. of bp compared	54	162	584	146	499	52	64
No. of bp in *D. subobscura*	57	913		1535		528	68

K, nucleotide divergence.

[a] See text and Figure 1 for details of the different regions analyzed.

[b] While intron 2 exhibits a very short stretch of similarity between the two obscura species, intron 3 (61 bp in *D. subobscura*) does not show enough similarity to be aligned.

sizes. Again, the distribution of *Ks* estimates only showed significant heterogeneity for the *D. subobscura vs. D. melanogaster* comparison with a probability lower or much lower than 0.05 for segments of 75, 100, 125 and 150 synonymous sites.

One way to graphically display these results is presented in Figure 2 which shows the distribution of divergence estimates (*Ks* and *Ka*) along the *Xdh* coding region as well as the confidence intervals (COMERON 1995) of the overall divergence estimates at 0.05 and 0.01 levels using a sliding window of 150 synonymous sites.

Putative regional variation in mutational bias along the coding region (KLIMAN and HEY 1994) has also been tested by comparing the base composition of synonymous sites, as G+C numbers, among adjacent segments of 150 synonymous sites. No significant heterogeneity has been detected in any of the three analyzed species ($P > 0.99$ for *D. subobscura*; $P > 0.80$, for *D. pseudoobscura*, and $P > 0.50$ for *D. melanogaster*; d.f. = 8).

Intragenic codon bias analysis: The CBI (MORTON 1993) for the *Xdh* coding region is very similar in the two obscura group species (0.535 and 0.513 for *D. subobscura* and *D. pseudoobscura*, respectively) and higher than that obtained for *D. melanogaster* (0.349). Moreover, while the pattern of codon bias along the coding region is quite similar for *D. subobscura* and *D. pseudoobscura*, the pattern for *D. melanogaster* exhibits differences in some regions (see Figure 3).

The relationship between codon bias and heterogeneity of *Ks* and *Ka* estimates along the *Xdh* coding region has been studied using the largest segment size showing significant heterogeneity of both synonymous and nonsynonymous substitutions (150 synonymous sites) in order to diminish the high associated variance of any codon bias index. Comparison of equally sized regions cancels any putative length effect on the codon bias measures (see MATERIALS AND METHODS). No significant correlation has been detected between *Ks* estimates and mean codon bias either in the *D. subobscura vs. D. melanogaster* ($P > 0.40$) or in the *D. subobscura vs. D. pseudoobscura* ($P > 0.20$) comparisons. Furthermore, there is no significant correlation between codon bias and *Ka* estimates for these data ($P > 0.90$ and $P > 0.50$, for *D. subobscura vs. D. pseudoobscura* and *D. subobscura vs. D. melanogaster* comparisons, respectively). Similar results have been obtained when using other codon bias measures like the G+C content of synonymous sites and *Xi/L* (data not shown).

DISCUSSION

Divergence at noncoding sequences: Although we have sequenced more than 4 kb of noncoding regions (taking into account both introns and flanking regions) in *D. subobscura*, most of them can not be aligned even between the two more closely related species of the obscura group, which indicates that they have evolved at a very high rate. On the other hand, the number of substitutions in such alignable regions is significantly lower than that of synonymous substitutions at the *Xdh* coding region ($P < 0.001$ in both interspecific comparisons). These alignable regions would be highly constrained to vary, putatively due to the presence of regulatory or control elements. This situation differs from that of flanking regions of *Adh* and *rp49*, also sequenced in the same three species. In the case of the 5′ flanking region of *Adh* (975 bp), sequences can be aligned in all three pairwise comparisons, which would be compatible either with generalized constraint in this region or with a lower mutation rate. In the case of the 5′ flanking region of *rp49*, sequences are only alignable between the two species of the obscura group, indicating an intermediate situation.

Relationship between divergence estimates and codon bias along and among genes: Heterogeneity in *Ka* estimates among genes has been easily explained since different coding regions can have different overall functional constraints at the protein level. Like in comparisons among genes, the significant heterogeneity detected in the distribution of nonsynonymous substitutions along the *Xdh* coding region can be explained since different parts of the encoded protein can have

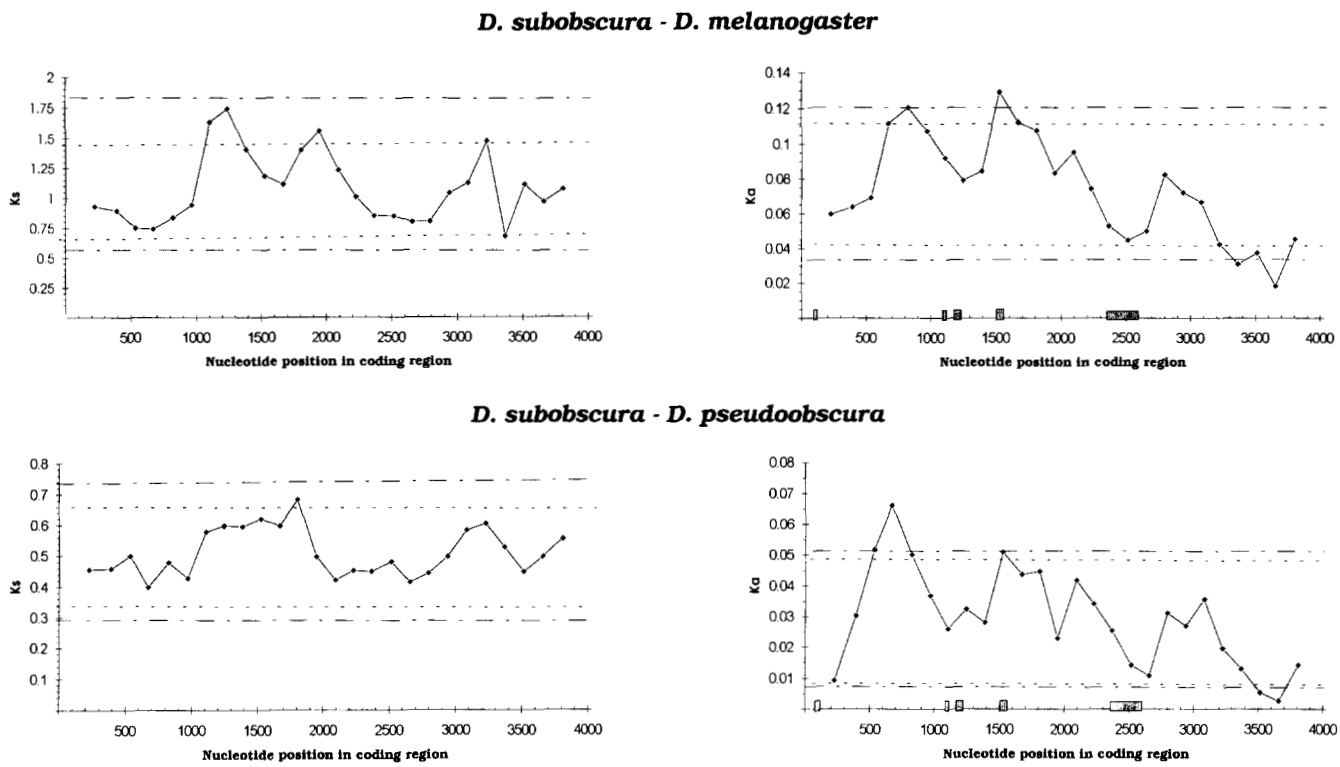

FIGURE 2.—*Ks* and *Ka* divergence estimates along the *Xdh* coding region for both the *D. subobscura vs. D. pseudoobscura,* and the *D. subobscura vs. D. melanogaster* comparisons. Sliding windows with 150 synonymous sites and on average 433 total nucleotides have been used for analysis with a movement between windows of 50 synonymous sites. Horizontal dashed lines indicate the confidence intervals (at 0.05 (-----) and 0.01 (·–·–·) levels) of the nucleotide divergence estimates, obtained by computer simulation (10,000 replicates) using the overall divergence estimate for the *Xdh* region and considering the length, G+C content and transition:transversion ratio of each window, as described in COMERON (1995). In the distribution of *Ka* estimates, boxes along the coding region indicate the location of some binding domains of the *Xdh* enzyme: 2Fe/2S, ATP/GTP, FAD/NAD+/NADH, and Mo-pterin binding domains (26, 24, 63 + 63, and 288 bp long, respectively) (AMAYA *et al.* 1990; HUGHES *et al.* 1992).

different selective constraints. In this sense, the number of amino acid replacement substitutions shows a relative reduction in most regions where binding domains of the *Xdh* enzyme have been located (Figure 2).

Variability of *Ks* estimates among genes in Drosophila has been related to natural selection, the effect of which would be the reported negative relationship between *Ks* estimates and codon usage bias (SHIELDS *et al.* 1988; SHARP and LI 1989). They have also proposed that the degree of selective constraint is sensitive to expression level. To contrast, the detected intragenic heterogeneity of *Ks* estimates along a given coding region can not be explained in this way. Like in the case of some enterobacterial genes (where small segments of only a few codons were analyzed; LAWRENCE *et al.* 1991; EYRE-WALKER 1994), we have found significant heterogeneity of *Ks* estimates along the *Xdh* coding region between *D. melanogaster* and *D. subobscura.* Thus, we will analyze in detail the relationship between *Ks* and codon bias among different genes and along a coding region.

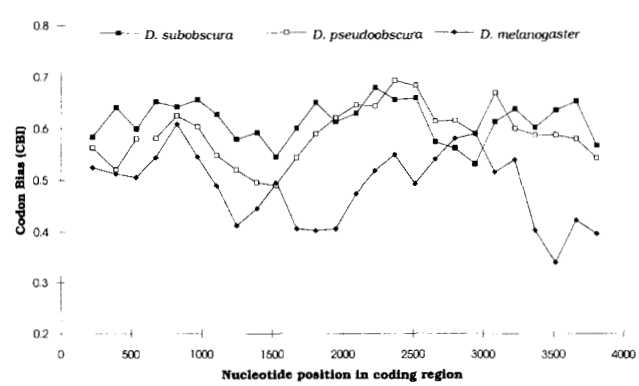

FIGURE 3.—Codon bias values (CBI) along the *Xdh* coding region of *D. subobscura, D. pseudoobscura* and *D. melanogaster* for the same windows analyzed in Figure 2.

There are several hypotheses, within a general mutation-selection-drift theory (reviewed in BULMER 1991), of how natural selection affects the usage of synonymous codons. They all involve transfer RNA (tRNA) abundance and expression level of the encoded protein; natural selection would act to enhance accuracy (reducing either amino acid misincorporation or the cost of proofreading) or to enhance elongation rates at the translational level (KURLAND 1987a,b; PRECUP and PARKER 1987; AKASHI 1994). The observed negative relationship between codon bias and *Ks* among genes does not preclude *a priori* any of these hypotheses. In prokaryotes this observed relationship has been associated with differences in expression level of each gene and

with different abundance of tRNA (IKEMURA 1985; SHARP and LI 1986). Although the same situation has been assumed in *D. melanogaster* (SHIELDS *et al.* 1988), in multicellular organisms expression levels and tRNA abundance could be different both among tissues and among developmental stages. In fact, codon bias at the *Adh* gene does not seem to be correlated with gene expression among different Drosophila species (JUAN *et al.* 1994). On the other hand, selection on messenger RNA (mRNA) secondary structure has been proposed to explain the use in *Escherichia coli* of nonoptimal codons in regions with low *Ks* estimates (LAWRENCE *et al.* 1991; EYRE-WALKER and BULMER 1993), and where no negative relationship between *Ks* and codon bias has been detected.

The different hypotheses to explain codon usage make different predictions about the relationship between divergence estimates and codon bias both among genes and within a given coding region. If natural selection is acting to reduce amino acid misincorporation, one would expect a negative correlation between codon bias and both *Ks* and *Ka*. In fact, functionally constrained regions (with low *Ka* values) would exhibit high codon bias and therefore low *Ks* estimates. On the other hand, if natural selection is acting to enhance elongation rates (KURLAND 1987b), a negative relationship would only be expected between *Ks* and codon bias along the coding region and among genes. Finally, no relationship between divergence estimates and codon bias would be expected if selection was acting only on mRNA secondary structure differences (BULMER 1991; EYRE-WALKER and BULMER 1993).

Among genes: The negative correlation expected under certain models between *Ks* and codon bias (CBI) has been detected among 20 coding regions compared between *D. melanogaster* and species of the obscura group (Figure 4) ($P < 0.001$, $R = -0.60$). On the other hand, a significant correlation between *Ka* and codon bias ($P < 0.01$) as well as between *Ks* and *Ka* ($P < 0.01$) is only found when genes with a high *Ka*/*Ks* ratio are excluded (*cp15, cp16, cp18, cp19* and *Arr*), which could be subject—as suggested by AKASHI (1994)—to selection for elongation rates (see Figure 4).

Within genes: No negative correlation between either *Ks* or *Ka* estimates and codon bias has been detected along the *Xdh* coding region in the *D. subobscura vs. D. melanogaster* comparison. The putative effect of analyzing only nine regions along the *Xdh* coding region in not finding a significant correlation between *Ks* and codon bias estimates has been ruled out by Monte-Carlo simulations where subsets of nine genes were randomly drawn from the set of 15 genes described above ($P < 0.002$).

When the *Gart* (*ade-3*) gene, which is the only other long enough coding region that has been sequenced in *D. melanogaster* and *D. pseudoobscura*, is similarly analyzed (using adjacent segments of 150 synonymous sites): (1) *Ks* estimates show a heterogeneous distribution along the coding region, (2) the distribution of codon bias is different in the two species, and (3) neither *Ks* nor *Ka* estimates are significantly correlated with codon bias ($P > 0.4$ and $P > 0.05$, respectively). However, it could be argued that failure to find any significant relationship between codon bias and divergence estimates (*Ks* or *Ka*) along those genes could be simply due to a mild effect of selection, given that their overall *Ks* estimates are moderately high. A third long coding region [*sevenless (sev)*], sequenced both in *D. melanogaster* and *D. virilis* and for which there seems to be some indication of selection acting to enhance accuracy (AKASHI 1994), has also shown the same kind of results when it has been similarly analyzed: despite significant heterogeneity of *Ks* estimates along the coding region ($P < 0.05$), there is no correlation between codon bias and either *Ks* or *Ka* ($P > 0.2$ and $P > 0.5$, respectively).

Therefore, the heterogeneous distribution of *Ks* estimates along these long coding regions (*Xdh, Gart* and *sev*) does not seem to be mainly due to those forces responsible for heterogeneity among genes, as no significant relationship between *Ks* estimates and codon bias is detected within genes. Alternative explanations assuming variability of mutation rate across the coding region would predict a *Ks vs. Ka* interdependence along the coding region (LAWRENCE *et al.* 1991) that is not detected either in the *Xdh* ($P > 0.9$ and $P > 0.5$ for *D. subobscura vs. D. melanogaster* and *D. subobscura vs. D. pseudoobscura* comparisons, respectively) or *sevenless* coding regions ($P > 0.1$), although significant in the *Gart* coding region ($P < 0.05$). Present data are consistent with those found in *E. coli* indicating that, in some regions, selection on synonymous sites would not enhance either accuracy or elongation rates but would be related to mRNA secondary structure (LAWRENCE *et al.* 1991; EYRE-WALKER and BULMER 1995). Unlike in *E. coli* (EYRE-WALKER and BULMER 1993), neither the relative number of synonymous substitutions nor codon bias show any significant decrease near the start of the gene (see Figures 2 and 3), which might simply indicate different constraints on mRNA secondary structure between *E. coli* and Drosophila.

Epistatic selection acting at the intragenic level to maintain secondary structures of pre-mRNA has been proposed (STEPHAN and KIRBY 1993; KIRBY *et al.* 1995) for the Drosophila *Adh* locus. Selection would constrain the substitution at particular nucleotide positions and/or clusters of nucleotides that would produce hairpins and stabilize such RNA structures. A positive correlation between stem length and physical length of the loop has been detected (STEPHAN and KIRBY 1993) and, at the same time, longer stem lengths would have weaker selection pressure on single sites within the stem (W. STEPHAN, personal communication). Also, it is expected that the longer the mRNA, the longer the physical mean distance between stems. Then, if selection at this level actually plays a significant role, a positive correlation is expected among genes between *Ks* and both the coding length and the gene length (considering only

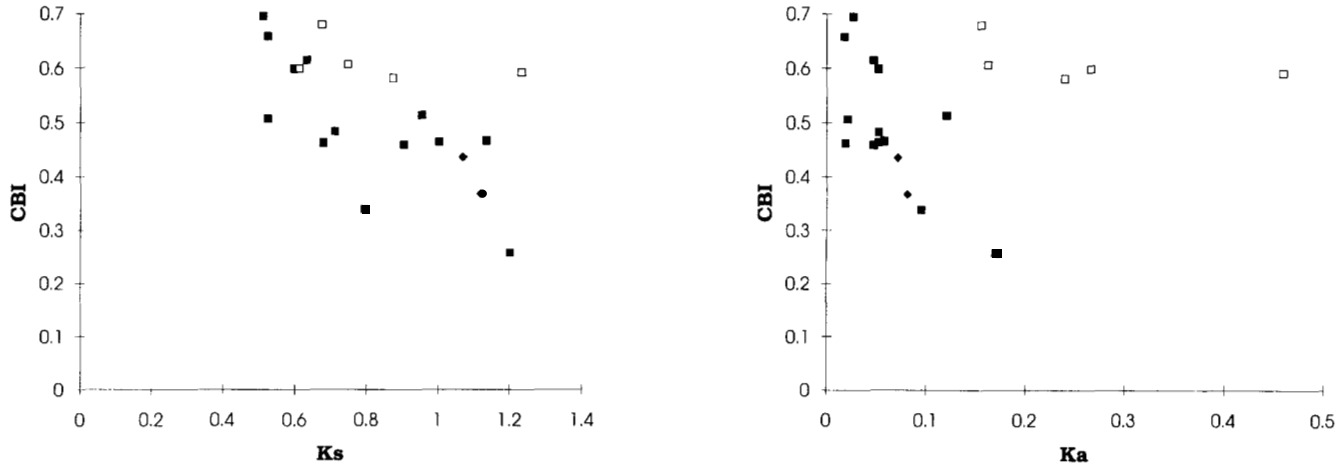

FIGURE 4.—Relationship between bias in synonymous codon usage (measured as CBI), and *Ks* and *Ka* divergence estimates between *D. melanogaster* and species of the obscura group. *D. pseudoobscura* is the species used in the comparison unless noted below. All sequences were drawn from the GeneBank/EMBL sequence library. Gene and accesion numbers of used sequences: *Adh,* alcohol dehydrogenase, mean of *D. subobscura* and *D. pseudoobscura* (M14802, Y00602, M55545); *Antp,* antennapedia, *D. subobscura* (M14495, M14496, X60995, X62246); *Arr, D. miranda* (M30140, X54084); *bcd,* bicoid (X07870, X55735); *Cp15,* chorion protein s15, *D. subobscura* (X02497, X53423); *Cp16,* chorion protein s16, *D. subobscura* (X16715, X53423); *Cp18,* chorion protein s18, *D. subobscura* (X02497, X53423); *Cp19* chorion protein s19 *D. subobscura* (X02497, X53423); *Esterase6/5b* (J04167, M55907); *Gart/ade3,* glycinamide ribotide transformylase (J02527, X06285); *Gld,* glucose dehydrogenase (M29298, M29299); *hsp82,* heat shock protein 82 (X03810, X03812); *Pcp,* cuticle protein ((J02527, M255907); *Rh1/ninaE,* opsin Rh1 (K02315, X65877); *Rh2,* opsin Rh2 (M12896, X65878); *Rh3,* opsin Rh3 (M17718, X65879); *Rp49/M(3)99D,* ribosomal protein 49, mean of *D. subobscura* and *D. pseudoobscura,* (X00848, S59382, M21333); *Ubx,* ultrabithorax (X05723, X05179); *Uro,* urate oxidase (X51940, X57113); *Xdh/ry,* Xanthine dehydrogenase, mean of *D. subobscura* and *D. pseudoobscura* (Y00308, M33977, present work Y08237). Diamonds indicate *Xdh* and *Gart* genes; open squares indicate *Arr, Cp15, Cp16, Cp18* and *Cp19* genes.

the length of the exons or that of exons and introns, respectively). The total coding length of eighteen genes, from the set of 20 genes described in Figure 4, is available. When these genes are considered, there is a significant correlation between *Ks* estimates and coding length ($P < 0.05$, $R = 0.51$). However, when the length of introns has also been taken into account, such correlation becomes nonsignificant ($P < 0.10$; $R = 0.44$). Equivalent results are obtained when the five genes with a high *Ka/Ks* ratio described above are not considered. *Ks* heterogeneity along particular coding regions could be, therefore, explained by differential selection on synonymous substitutions along the coding region acting on mRNA, or pre-mRNA, secondary structure. Also,

TABLE 3

Relative-rate test for synonymous substitutions between *D. subobscura* and *D. pseudoobscura vs. D. melanogaster*

Gene	*Ks (D. subobscura)*	*Ks (D. pseudoobscura)*	*Ds*	*MDs*[a,b]	
				$P = 0.05$	$P = 0.01$
Xdh	1.0246	1.1115	−0.0869*	0.0860	0.1171
Adh	0.7079	0.5581	0.1498**	0.1114	0.1483
rp49	0.4876	0.5354	−0.0478	0.0919	0.1348

See APPENDIX for details. WU and LI's test gives the same result for *Adh,* while for *Xdh* the difference is only marginally significant ($P < 0.10$). MUSE and GAUT's test, nevertheless, does not detect any significant difference in synonymous substitution rates between the *D. subobscura* and *D. pseudoobscura* lineages for any of the three genes analyzed. The sequences have been drawn from the GeneBank/EMBL sequence library as detailed in Figure 4, except for the *rp49* sequence of *D. melanogaster* (A. CADIC-JAQUIER and M. ROSBASH, personal communication) and for the *Xdh* sequence of *D. subobscura* (present work).

[a] *MDs* indicates the maximum absolute difference between *Ks* estimates for the two more distant interspecific comparisons accepted under the null hypothesis of equal substitution rates for the two obscura group lineages.

[b] Simulation parameters: *Xdh,* 1333 codons; 77% G+C at third positions of codons; transition ($\alpha = 1.2$):transversion ($\beta = 1.0$); *Ks*(s-p) = 0.4992, *Ka* (s-p) = 0.0284, *Ka*(s-m) = 0.0735, *Ka* (p-m) = 0.0692. *Adh,* 251 codons; 76% G+C at third positions of codons; transition ($\alpha = 2$):transversion ($\beta = 1$); *Ks*(s-p) = 0.3259, *Ka*(s-p) = 0.0223, *Ka*(s-m) = 0.0442, *Ka*(p-m) = 0.0496. *rp49,* 133 codons; 76% G+C at third positions of codons; transition ($\alpha = 1$):transversion ($\beta = 1$); *Ks*(s-p) = 0.1198, *Ka*(s-p) = 0.0038, *Ka*(s-m) = 0.0284, *Ka*(p-m) = 0.0246. "s", "p" and "m" denote *D. subobscura, D. pseudoobscura* and *D. melanogaster,* respectively.

these preliminary results would indicate that selection at this level might show stronger effects on mRNA than on pre-mRNA structures.

Among genes, however, selection would additionally act in an expression dependent and much stronger way to enhance accuracy and/or elongation rates. In fact, when the relationship between codon bias and both *Ks* and *Ka* estimates is considered for different genes (Figure 4), both *Xdh* and *Gart* fall within the general pattern of a negative relationship. Although selection acting at the RNA level would also affect nonsynonymous sites, selection coefficients preventing amino acid substitutions are expected to be of higher magnitude and, therefore, selection at the RNA level would mainly be detected at synonymous sites.

Furthermore, the nine adjacent *Ka* estimates along the *Xdh* coding region in the *D. subobscura-D. pseudoobscura* comparison are significantly correlated with those found in the same position in the *D. subobscura-D. melanogaster* comparison ($P < 0.01$), while no correlation is detected when *Ks* estimates are similarly analyzed ($P > 0.05$). Moreover, codon bias along the *Xdh* coding region in *D. melanogaster* shows a different pattern from those observed for both obscura species (Figure 3). These observations could be an indication of selection acting differentially in the two obscura lineages as compared to the *D. melanogaster* lineage (AKASHI 1995), selection that might be related to observed differences in gene structure and/or intron length between the two obscura species and *D. melanogaster*.

Nonconstancy of synonymous substitution rates among different lineages: Differences in substitution rates among different taxa have been ascribed to lineage effects due to differences in DNA repair efficiency (BRITTEN 1986), in generation time and rate of germline DNA replication (WU and LI 1985; LI *et al.* 1987), and in metabolic rate (MARTIN and PALUMBI 1993). Thus, assuming similar DNA repair and metabolic rates when closely related species are analyzed, behavior of neutral mutations would show a generation-time effect (modulated by the germline DNA replication rate). However, if the selection coefficients are not small enough, differences in the effective population number (*Ne*) (KIMURA 1968) and/or enviromental diversity (OHTA 1987; OHTA and TACHIDA 1990) would play a role in determining the fate of mutants. Nonsynonymous mutations, on the other hand, are predicted to show weaker lineage effects than synonymous mutations (OHTA 1973, 1987; LI 1979; WU and LI 1985; LI *et al.* 1987; GILLESPIE 1989).

Lineage effects for synonymous substitutions have been commonly detected when analyzing highly diverged species, including analyses of some Drosophila species (WU and LI 1985; BRITTEN 1986; MORIYAMA 1987; GODDARD *et al.* 1990). The null hypothesis of equal rates of synonymous substitutions in the two obscura lineages (*D. subobscura* and *D. pseudoobscura*) has been tested for the *Xdh, Adh* and *rp49* coding regions, the only coding regions sequenced in these species.

A relative-rate test based on confidence intervals obtained by computer simulation has been used as it is more sensitive in detecting differences in synonymous rates than previous methods (see APPENDIX). As shown in Table 3, the rate is significantly higher in the *D. pseudoobscura* than in the *D. subobscura* lineage for *Adh,* while the null hypothesis of equal rates cannot be rejected for *rp49* and only barely rejected for *Xdh.* Although *Xdh* seems to deviate in the opposite sense than *Adh* and this could point to gene-specific effects, only a comparison of a large number of genes between these two species will allow assessing whether the *D. subobscura* and *D. pseudoobscura* lineages actually show lineage effects.

We thank J. ROZAS, W. STEPHAN and J. BRAVERMAN for critical comments on the manuscript, and M. KREITMAN and R. R. HUDSON for useful suggestions and discussions. We also thank M. RILEY for the *Xdh* clones of *D. melanogaster* and *D. pseudoobscura,* and S. V. MUSE for sending us a copy of the program described in MUSE and GAUT (1994). We are also grateful to two anonymous reviewers for comments and criticism. This work was supported by a predoctoral fellowship from Ministerio de Educación y Ciencia, Spain, to J.M.C., and Grants PB88-0196 and PB91-0245 from Dirección General de Investigación Científica y Técnica and GRQ93-1100 from Comissió Interdepartamental de Recerca i Innovació Tecnològica to M.A.

LITERATURE CITED

AGUADÉ, M., 1988 Nucleotide sequence comparison of the *rp49* gene region between *Drosophila subobscura* and *D. melanogaster.* Mol. Biol. Evol. **5:** 433–441.

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

AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076.

AMAYA, Y., K.-I. YAMAZAKI, M. SATO, K. NODA, T. NISHINO *et al.* 1990 Proteolitic conversion of xanthine dehydrogenase from the NAD-dependent type of the O2-dependent type. J. Biol. Chem. **265:** 14170–14175.

BRITTEN, R. J., 1986 Rates of DNA sequence evolution differ between taxonomic groups. Science **231:** 1393–1398.

BULMER, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897–907.

CARULLI, J., D. E. KRANE, D. L. HARTL and H. OCHMAN, 1993 Compositional heterogeneity and patterns of molecular evolution in the Drosophila genome. Genetics **134:** 837–845.

COMERON, J. M., 1995 A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J. Mol. Evol. **41:** 1152–1159.

COTÉ, B., W. BENDER, D. CURTIS and A. CHOVNICK, 1986 Molecular mapping of the *rosy* locus in *Drosophila melanogaster.* Genetics **112:** 769–783.

DEVEREUX, J., P. HAEBERLI and O. SMITHIES, 1984 A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12:** 387–395.

EYRE-WALKER, A., 1994 Synonymous substitutions are clustered in enterobacterial genes. J. Mol. Evol. **39:** 448–451.

EYRE-WALKER, A., and M. BULMER, 1993 Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. **21:** 4599–4603.

EYRE-WALKER, A., and M. BULMER, 1995 Synonymous substitution rates in enterobacteria. Genetics **140:** 1407–1412.

GILLESPIE, J. H., 1989 Lineage effects and the index of dispersion of molecular evolution. Mol. Biol. Evol. **6:** 636–647.

GODDARD, K., A. CACCONE and J. R. POWELL, 1990 Evolutionary implications of DNA divergence in the *Drosophila obscura* group. Evolution **44:** 1656–1670.

HENIKOFF, S., 1984 Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene **28:** 351–359.

HILLIKER, A. J., S. H. CLARK and A. CHOVNICK, 1980 Cytogenetic analysis of the chromosomal region immediately adjacent to the *rosy* locus in *Drosophila melanogaster*. Genetics **95:** 95–110.

HUGHES, R. K., W. A. DOYLE, A. CHOVNICK, J. R. WHITTLE, F. F. BURKE *et al.*, 1992 Use of *rosy* mutant strains of *Drosophila melanogaster* to probe the structure and function of xanthine dehydrogenase. Biochem. J. **285:** 507–513.

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

JUAN, E., M. PAPACEIT and A. QUINTANA, 1994 Nucleotide sequence of the genomic region encompassing *Adh* and *Adh-dup* of *D. lebanonensis* (Scaptodrosophila): gene expression and evolutionary relationships. J. Mol. Evol. **38:** 455–467.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism* III, edited by H. N. MUNRO. Academic Press, New York.

KEITH, T. P., M. A. RILEY, M. KREITMAN, R. C. LEWONTIN, D. CURTIS *et al.*, 1987 Sequence of the structural gene for Xanthine dehydrogenase (*rosy* locus) in *Drosophila melanogaster*. Genetics **116:** 67–73.

KIMURA, M., 1968 Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. **11:** 247–269.

KIMURA, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111–120.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

KING, J. L., and T. H. JUKES, 1969 Non-Darwinian evolution. Science **164:** 788–798.

KIRBY, D. A., S. V. MUSE and W. STEPHAN, 1995 Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92:** 9047–9051.

KLIMAN, R. M., and J. HEY, 1994 The effects of mutation and natural selection on codon bias in the genes of Drosophila. Genetics **137:** 1049–1056.

KURLAND, C. G., 1987a Strategies for efficiency and accuracy in gene expression. 2. Growth optimized ribosomes. Trends Biochem. Sci. **12:** 169–171.

KURLAND, C. G., 1987b Strategies for efficiency and accuracy in gene expression. 1. The major codon preference: a growth optimization strategy. Trends Biochem. Sci. **12:** 126–128.

LAWRENCE, J. G., D. L. HARTL and H. OCHMAN, 1991 Molecular considerations in the evolution of bacterial genes. J. Mol. Evol. **33:** 241–250.

LEE, C. S., D. CURTIS, M. MCCARRON, C. LOVE, M. GRAY *et al.*, 1987 Mutations affecting expression of the *rosy* locus in *Drosophila melanogaster*. Genetics **116:** 55–66.

LI, W.-H., 1979 Maintenance of genetic variability under the pressure of neutral and deleterious mutations in a finite population. Genetics **92:** 647–667.

LI, W.-H., 1993 Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36:** 96–99.

LI, W.-H., M. TANIMURA and P. M. SHARP, 1987 An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J. Mol. Evol. **25:** 330–342.

MARTIN, A. P., and S. R. PALUMBI, 1993 Body size, metabolic rate, generation time, and the molecular clock. Proc. Natl. Acad. Sci. USA **90:** 4087–4091.

MORIYAMA, E. N., 1987 Higher rates of nucleotide substitution in Drosophila than in mammals. Jpn. J. Genet. **62:** 139–147.

MORIYAMA, E. N., and T. GOJOBORI, 1992 Rates of synonymous substitution and base composition of nuclear genes in Drosophila. Genetics **130:** 855–864.

MORIYAMA, E. N., and D. L. HARTL, 1993 Codon usage bias and base composition of nuclear genes in Drosophila. Genetics **134:** 847–858.

MORTON, B. R., 1993 Chloroplast DNA codon use: evidence for selection at the psb A locus based on tRNA availability. J. Mol. Evol. **37:** 273–280.

MUSE, S. V., and B. S. GAUT, 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. **11:** 715–724.

OHTA, T., 1973 Slightly deleterious mutant substitutions in evolution. Nature **246:** 96–98.

OHTA, T., 1987 Very slightly deleterious mutations and the molecular clock. J. Mol. Evol. **26:** 1–6.

OHTA, T., and H. TACHIDA, 1990 Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. Genetics **126:** 219–229.

PAMILO, P., and N. O. BIANCHI, 1993 Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. Mol. Biol. Evol. **10:** 271–281.

PRECUP, J., and J. PARKER, 1987 Missence misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. **262:** 11351–11356.

RILEY, M., 1989 Nucleotide sequence of the *Xdh* region in *Drosophila pseudoobscura* and an analysis of the evolution of synonymous codons. Mol. Biol. Evol. **6:** 33–52.

SEGARRA, C., and M. AGUADÉ, 1993 Nucleotide divergence of the *rp49* gene region between *Drosophila melanogaster* and two species of the obscura group of Drosophila. J. Mol. Evol. **36:** 243–248.

SHARP, P. M., and W.-H. LI, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:** 28–38.

SHARP, P. M., and W.-H. LI, 1987 The rate of synonymous substitutions in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:** 222–230.

SHARP, P. M., and W.-H. LI, 1989 On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. **28:** 398–402.

SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS and F. WRIGHT, 1988 "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

SOKAL, R. R., and F. J. ROHLF, 1995 *Biometry*. 3rd ed. W. H. Freeman and Co., New York.

STADEN, R., 1982 Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Res. **10:** 4731–4751.

STEPHAN, W., and D. A. KIRBY, 1993 RNA folding in Drosophila shows a distance effect for compensatory fitness interactions. Genetics **135:** 97–103.

TAJIMA, F., 1993 Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135:** 599–609.

WU, C.-I., and W.-H. LI, 1985 Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82:** 1741–1745.

Communicating editor: W-H. LI

APPENDIX

Relative-rate test of nucleotide substitutions: A new relative-rate test, which compares both synonymous and nonsynonymous substitution rates between two related lineages (A and B) when they are both referred to an outgroup species (C), is used. Thus, following COMERON (1995), confidence intervals for the difference between comparisons A-C and B-C under the null hypothesis of equal substitution rates along the two more closely related lineages A and B since their divergence (O) are obtained by computer simulation. When synonymous substitutions are analyzed, the difference between the two distant comparisons (Ds), $Ds = Ks(AC) - Ks(BC)$, is expected to be zero and lower than the absolute maximum difference obtained by computer simulation (MDs) for a given level of significance. Nonsynonymous substitutions are analyzed in a similar way, $Da = Ka(AC) - Ka(BC)$ and this value compared to the maximum absolute difference accepted under the null hypothesis (MDa). Such a maximum accepted difference is not only affected by the AB and OC evolutionary distances but also by the number of codons, the nucleotide substitution pattern and the G+C content at third positions of codons (data not shown).

TABLE A1

Computer simulation analysis of relative-rate tests

	Percentage of rejections ($P < 0.05$)							
	Tajima's test		Wu and Li's test		Muse and Gaut's test		Current test	
	$L = 250$	$L = 1000$	$L = 250$	$L = 1000$	$L = 250$	$L = 1000$	$L = 250$	$L = 1000$
	(a) Ks (A − B) = 0.375, Ks (AB − C) = 0.75[a]							
Ratio of synonymous rate, lineage OA:lineage OB								
1.00:1.00	3.6	3.8	4.6	6.2	3.0	1.0	—	—
1.00:1.25	5.6	9.6	8.8	18.0	4.0	11.0	10.0	13.0
1.00:1.50	6.8	16.4	14.0	44.2	8.0	47.0	16.2	37.2
1.00:2.00	10.0	57.2	35.2	88.0	24.0	83.0	37.2	82.0
	(b) Ks (A − B) = 0.15, Ks (AB − C) = 0.30[a]							
Ratio of synonymous rate, lineage OA:lineage OB								
1.00:1.00	1.2	0.8	1.2	0.4	1.0	0.0	—	—
1.00:1.25	2.0	3.6	2.0	9.6	1.0	6.0	11.6	24.4
1.00:1.50	3.2	16.4	7.6	46.8	1.0	27.0	18.8	63.2
1.00:2.00	8.4	66.4	29.2	89.6	9.0	87.0	55.6	94.8

[a] Other simulation parameters: $Ka = Ks/5$ in all cases, G+C content at third positions of codons, 70%; transition ($\alpha = 1$):transversion ($\beta = 1$); L denotes the number of codons. The number of replicates (n) is 500 for each condition and test, except for Muse and Gaut's (1994) test where $n = 100$ as detailed by the authors because of its time-consuming procedure. In all cases equal rates of nonsynonymous substitutions among lineages OA and OB have been assumed.

Therefore, pseudo-random coding regions with a given number of codons and G+C content at third positions are generated and both synonymous and nonsynonymous substitutions are randomly distributed. Both kinds of substitutions are independently biased in order to take into account the transition:transversion ratio and the G+C content. According to the null hypothesis, the number of synonymous substitutions applied to the original sequence to obtain the A sequence is $Ks(AB)/2$, and the same number is applied to obtain the B sequence. The C sequence is then obtained after applying $[Ks(AC)+Ks(BC)-Ks(AB)]/2$ synonymous substitutions. Nonsynonymous substitutions are distributed in an equivalent way. After independent replicates, the maximum absolute difference (MDs and MDa) accepted under the null hypothesis is obtained for different levels of significance. Table A1 shows some results of computer simulation analyses with different divergence estimates and/or length of the coding region. The method proposed here is in general more sensitive in detecting differences of synonymous substitution rates between the two more closely related lineages than those previously described by Tajima (1993), Wu and Li (1985) and Muse and Gaut (1994), as shown by comparing the percentages of rejection of the null hypothesis.