

The molecular clock revisited: the rate of synonymous vs. replacement change in *Drosophila*

Ling-Wen Zeng, Josep M. Comeron, Bin Chen & Martin Kreitman*

Department of Ecology and Evolution, University of Chicago, 1101 E. 57th St. Chicago, IL 60637, USA; *Author for correspondence (Phone: (773) 702-1222; E-mail: mkre@midway.uchicago.edu)

Key words: protein evolution, *Drosophila*, synonymous, nonsynonymous, Index of Dispersion

Abstract

Rates of synonymous and nonsynonymous substitution were investigated for 24 genes in three *Drosophila* species, *D. pseudoobscura*, *D. subobscura*, and *D. melanogaster*. *D. pseudoobscura* and *D. subobscura*, two distantly related members of the *obscura* clade, differ on average by 0.29 synonymous nucleotide substitutions per site. *D. melanogaster* differs from the two *obscura* species by an average of 0.81 synonymous substitutions per site. Using a method developed by Gillespie, we investigated the variance to mean ratio, or Index of Dispersion, R , of substitutions along the three species' branches to test the fundamental prediction of the neutral theory of molecular evolution, $E(R) = 1$. For nonsynonymous substitutions, the average R , R_a is 1.6, which is not significantly different from the neutral theory prediction. Only 5 of the 24 genes had significantly large R_a values, and 12 of the genes had R_a estimates of less than one. In contrast, the Index of Dispersion for synonymous substitutions was significantly large for 12 of the 24 genes, with an average of $R_s = 4.4$, also statistically significant. These findings contrast with results for mammals, which showed overdispersion of nonsynonymous substitutions, but not of synonymous substitutions. Weak selection acting to maintain codon bias in *Drosophila*, but not in mammals, may be important in explaining the high variance in the rate of synonymous substitutions in this group of organisms.

Introduction

The behavior of the molecular clock has been a major issue in evolutionary genetics ever since its proposal by Zuckerkandl and Pauling (1962, 1965) and Margoliash (1963). The constancy in the rate of amino acid substitution found by Zuckerkandl and Pauling for mammalian hemoglobins (α and β) challenged conventional Darwinian thinking and provided motivation for Kimura's (1969) neutral theory of molecular evolution. A key prediction of this theory is that the rate of substitution per year, K_y , will be equal to the mutation rate to neutral alleles, $\mu_y f_0$, where μ_y is the total mutation rate per year, and f_0 is the fraction of mutations that are selectively neutral. f_0 is assumed to vary among proteins (King & Jukes, 1969); the remaining mutations ($\mu_y(1 - f_0)$) are assumed to be strongly deleterious and thus do not contribute to molecular evolution.

More generally, a molecular clock will be expected for any gene in which the proportion of mutations fixing along a lineage—either by selection or by drift—remains constant. However, conditions for achieving a constant rate of evolution are more stringent under selection than under neutrality. In particular, the rate of substitution of selectively advantageous mutations is determined by the product of four parameters, $K_y = N_e \mu_y f_a s$, where N_e is the effective population size; μ_y the total mutation rate per year; f_a the fraction of advantageous mutants; and s the selection coefficient (Kimura & Ohta, 1971). Because population size is likely to fluctuate over time, Kimura doubted that the rate of adaptive substitution would be constant.

Both models require a constant mutation rate per year to be compatible with the mammalian data. Ohta and Kimura (1971) indicated that the mutation rate is expected to be constant per generation (μ_g) rather than per year, so that for neutral mutations $K_y =$

$(\mu_g/g)f_0$, where g is the generation time. Thus, neutral mutations should exhibit a generation-time effect (GTE); that is, the longer the generation time of the species in a given lineage, the smaller the number of neutral substitutions per year. Such an effect has been observed in mammals for synonymous substitutions (Li, Tanimura & Sharp, 1987), but the effect appears to be much weaker for amino acid replacement changes (Ohta, 1993). For slightly deleterious mutations, $N_e s \cong 1$, $K_y = (\mu_g/4N_e s g)f_n$, where f_n is the fraction of nearly neutral mutations. In this case, a constant number of substitutions per year will be expected only if there is an inverse relationship between N_e and g , $N_e \propto 1/g$. This inverse relationship between population size and generation time, a central premise of the nearly neutral theory, was initially supported by weak empirical evidence (Kimura & Ohta, 1971) with some conspicuous examples—elephants and *Drosophila*—, but is now statistically well established (Chao & Carr, 1993).

Perhaps the strongest evidence against a molecular clock for proteins has come from the study of variation in rates among lineages (Ohta & Kimura, 1971; Langley & Fitch, 1973, 1974; Kimura, 1983; Gillespie, 1984, 1986a,b, 1989; Ohta, 1995). Under neutral theory, substitutions conform to a Poisson process, so that the Index of Dispersion, $R(t)$ (R for convenience), the ratio of the variance in the rate to the mean rate, is expected to equal one (Ohta & Kimura, 1971). A significantly large variance in the rate of evolution among lineages can be evidence for selectively driven substitutions, but it can also be indicative of nonselective factors. These factors, termed lineage effects, act on all genes within a lineage (Gillespie, 1989), and include differences in branch lengths, generation times, mutation rates and/or efficiency of DNA repair (Britten, 1986), and metabolic-rate effects (Martin & Palumbi, 1993).

Kimura (1983) represented the radiation of mammalian orders as a star phylogeny, which eliminated lineage length (and phylogeny) as a variable, and found an average value of R for five genes to be 2.6. A star phylogeny for mammalian orders is not widely accepted, however (see Easteal, 1988). Gillespie (1989, 1991) studied 20 coding sequences from humans, rodents (mouse or rat), and artiodactyls (bovine) (Li, Tanimura & Sharp, 1987). As there is only a single topology for the phylogeny of three species, there can be no error in the tree itself. He used the average rate of evolution of all genes along each branch to estimate lineage effects and used the residual differences among lin-

ages to estimate the variance in the substitution rate. This was achieved by weighting the observed number of synonymous (or nonsynonymous) substitutions in each gene by the average number of synonymous (or nonsynonymous) substitutions for all the genes in each lineage (see Gillespie, 1991, for details). This analysis revealed a clear difference between the average R for replacement (R_a) and synonymous (R_s) substitutions, suggesting that the two kind of substitutions are subject to different evolutionary forces. When the number of replacements for a given gene was weighted with the replacement lineage effects, 12 out of the 20 genes showed a significantly higher R_a than expected under neutrality, with an average R_a of 6.95. On the other hand, when the number of synonymous substitutions was weighted with the synonymous lineage effects, R_s was significant in only five genes, with an average R_s of 4.64.

The estimation of R_s can be seriously biased because of multiple substitutions in a site (Kimura, 1983; Gillespie, 1989; Bulmer, 1989). Bulmer (1989) has pointed out that high rates of synonymous substitutions increase the variance in the longest lineage, and therefore the value of R . After computer simulation, where the Jukes and Cantor (1969) correction for multiple hits was applied, the average values of R_a and R_s were 7.8 and 3.3, respectively. Ohta (1995) analyzed 49 single-copy mammalian genes in a similar manner, finding $R_a = 5.6$ and $R_s = 5.89$. These results indicate some overdispersion of the molecular clock. It should be pointed out, however, that in Gillespie's influential 1989 analysis, the average R was strongly influenced by three highly overdispersed genes, all of which encode hormones (see Wallis, 1996). Without them, the overdispersion index was much smaller and not significantly different from one. Nonetheless, there is general agreement that replacement changes in mammalian evolution are significantly overdispersed.

Lineage effects, including the generation time, have also been investigated by comparing weighted and unweighted estimates of R . Gillespie (1989) and Ohta (1995) found strong lineage effects for synonymous substitutions, but only a weak effect for nonsynonymous substitutions. As rodents have substantially shorter generation times than primates or artiodactyls as well as a higher synonymous substitution rate, the strong GTE for synonymous substitutions supports the view that they are mostly neutral. This interpretation assumes, however, that the rodent lineage did not branch prior to the primate-artiodactyl split (Easteal,

1988, 1990; Li et al., 1990; Li & Graur, 1991; Bulmer, Wolfe & Sharp, 1991; Easteal & Collet, 1994).

Easteal and Collet (1994) found a constant rate per year among the primate and rodent lineages for synonymous substitutions, indicating that the substitution rate is independent of generation time and metabolic rate. They also found a higher rate of nonsynonymous substitutions in the rodent lineage and suggested that replacement changes are slightly deleterious (Ohta, 1973, 1992). Li (1997) has pointed out that Easteal and Collet's results should be taken with caution because of the saturation in the number of transition substitutions in mammalian genes. A reanalysis of the same data produced a significantly higher synonymous substitution rate in the rodent lineage than in the human lineage. In addition, Ohta (1995) has indicated that if members of multigene families are omitted from the analysis, then the synonymous substitution rate in rodents becomes higher.

Thus, from the analysis of mammalian data, it is generally accepted that synonymous and replacement substitutions show different patterns of evolutionary change, consistent with different evolutionary processes: protein evolution is overdispersed and exhibits only weak generation time effects, whereas synonymous substitutions show relatively little overdispersion but show pronounced generation time effects (Li, Tanimura & Sharp, 1987; Gillespie, 1989; Ohta, 1995).

The high average R for nonsynonymous substitutions, i.e., the overdispersed clock, has been attributed to the action of natural selection (Gillespie, 1986a, 1989, 1991) and used as evidence against the neutral theory. Gillespie has suggested that protein evolution is episodic, with occasional bursts of substitution caused by environmental change. In contrast, Ohta (1995) argued for a nearly neutral model of protein evolution, where both random genetic drift and selection influence the rate of substitution. She proposed that the data for nonsynonymous substitutions can be explained by nearly neutral mutations, fluctuations in population size, and compensatory substitutions. Ohta also indicated that positive selection on nonsynonymous substitutions can be detected for particular genes due to duplication and/or functional differentiation (Ohta, 1991). Because of the potentially large effect of multiple hits on the variance of R (Bulmer 1989), both Gillespie (1989) and Ohta (1995) have been cautious in concluding that synonymous substitutions are overdispersed.

The analysis presented here focuses on the variation in substitution rates among lineages in *Drosophi-*

la. We have compared 24 coding sequences from *D. melanogaster* and two *obscura* group species (*D. subobscura* and *D. pseudoobscura*) and calculated the Index of Dispersion for synonymous and replacement substitutions. This analysis differs from previous studies on mammalian lineages in two important ways: (1) there is a clear difference in lineage lengths, because *D. melanogaster* is a member of the sister group to the *obscura* group, and (2) effective population sizes are likely to be several orders of magnitude larger for *Drosophila* species than for mammals. Mammalian lineages can be thought of as having a long-term effective population size of the order 10^4 (Nei & Graur, 1984), whereas *Drosophila* lineages appear to have effective population sizes of at least $10^6 - 10^7$ (Kreitman, 1983; Schaeffer, Aquadro & Anderson, 1987; Riley, Hallas & Lewontin, 1989). If replacement changes are overdispersed in *Drosophila*, as they are in mammals, we can conclude that the process(es) underlying the overdispersed clock is insensitive to effective population size. In contrast, if replacement substitutions are not overdispersed in *Drosophila*, then the fixation of mutations that are definitely deleterious in *Drosophila* but that are weakly deleterious in mammals may be important in the overdispersion of mammalian protein evolution.

Material and methods

Gene sequences

Table 1 lists the locus/gene names of the 24 coding sequences included in this study, approximately half of which were sequenced by us. The sequences were obtained in one of several ways. Eleven genes were obtained from GenBank or published reports with complete sequences for each of the three species. Of the remaining 13 genes included in this study, 5 had been previously sequenced in *D. melanogaster* and either *D. pseudoobscura* or *D. subobscura*, and 8 were previously sequenced in *D. melanogaster* only. The additional species sequences were determined by one or a combination of the following procedures. cDNA libraries of adult flies were constructed in Lambda-Zap (Stratagene) for both of the *obscura* group species, and random clones were sequenced from their 5' ends. BLAST searches (NCBI Entrez) identified clones with a homologous sequence in *D. melanogaster*. In this manner, we identified six and eight additional genes from *D. pseudoobscura* and *D. subobscura*, respectively. The

Table 1. Sequences used in this study

| Gene | <i>D. melanogaster</i> | | <i>D. pseudoobscura</i> | | <i>D. subobscura</i> | |
|-------------------|-------------------------|-----|-------------------------|-----|-----------------------|-----|
| | Acc. N.* | | Acc.N. | | Acc.N. | |
| <i>Adh</i> | X78384 | (c) | X62181 | (c) | M15545 | (c) |
| <i>Adhr</i> | X78384 | (c) | Y00602 | (c) | M55545 | (c) |
| <i>A/A-T/sesB</i> | S43651 | (c) | AF025798 ^a | (n) | AF025799 ^a | (n) |
| <i>Aprt</i> | M18432 | (c) | L06281 | (c) | AF025800 ^a | (c) |
| <i>ATPsyn-β</i> | X71013 | (c) | AF025801 ^a | (p) | AF025802 ^a | (p) |
| <i>Bcd</i> | X07870 | (c) | X55735 | (c) | X78058 | (p) |
| <i>Cp15</i> | X02497 | (c) | Benson (1995) | (c) | X53423 | (c) |
| <i>Cp19</i> | X02497 | (c) | Benson (1995) | (c) | X53423 | (c) |
| <i>Cyp1</i> | M62398 | (c) | AF025803 ^a | (n) | AF025804 ^a | (n) |
| <i>Ddc</i> | X04426 | (c) | Wang et al. (1996) | (p) | Wang et al. (1996) | (p) |
| <i>Eno</i> | X17034 | (c) | AF025805 ^a | (n) | AF025806 ^a | (n) |
| <i>Gad1</i> | X76198 | (c) | AF025807 ^a | (p) | AF025808 ^a | (p) |
| <i>Gapdh2</i> | M11255/256/259 | (c) | AF025809 ^a | (n) | AF025810 ^a | (n) |
| <i>Gld</i> | M29298/X07358/ X13581-2 | (c) | M29299 | (c) | AF025811 ^a | (c) |
| <i>Gpdh</i> | X67650 | (c) | U59682 | (c) | Wells (1996) | (n) |
| <i>Mlc1</i> | M10125 | (c) | L08052 | (c) | AF025812 ^a | (p) |
| <i>Rh1/minaE</i> | K02135 | (c) | X65877 | (c) | AF025813 ^a | (c) |
| <i>Rp49/RpL32</i> | X00848 | (c) | S59382 | (c) | M21333 | (c) |
| <i>Sod</i> | M24421 | (c) | U47871 | (p) | U47888 | (p) |
| <i>Sry-α</i> | X03121 | (c) | L19536 | (c) | L19535 | (c) |
| <i>Tpi</i> | X57576/S70377 | (c) | AF025814 ^a | (n) | AF025815 ^a | (n) |
| <i>Uro</i> | X51940 | (c) | X57113/S94076 | (c) | AF025816 ^a | (n) |
| <i>Vha14</i> | Z26918 | (c) | AF025796 ^a | (p) | AF025797 ^a | (p) |
| <i>Xdh/ry</i> | Y00308/Y00307 | (c) | M33977 | (c) | Y08237 | (c) |

* Complete (c), nearly complete (n), or partial (p) coding region. ^aPresent work.

cDNA sequences were then completely determined using either primer walking or nested deletion strategies (Zeng & Kreitman, 1996a, b). All DNA sequences were determined by cycle sequencing using fluorescent dye-terminator chemistry and an ABI 373A or ABI 377 sequencer.

For the remaining genes, we constructed oligonucleotide primers to amplify fragments of the homologous gene from one or both of the *obscura* group species. The templates were either genomic DNA or DNA prepared from an aliquot of a cDNA library. After sequencing these fragments, additional primers were designed, if necessary, to carry out inverse PCR to obtain flanking regions. We did not attempt to determine the complete coding sequence of every gene.

Care was taken to avoid including in the study genes that were known to be very highly conserved, such as histone or heat shock genes, or genes that were known to be members of closely related multigene families. We also confirmed by *in situ* hybridization that each of

the 24 genes in this study resides on the homologous chromosome arm in all three species.

The sequences were aligned after translation with ClustalW (Thomson, Higgins & Gibson, 1994), and in some cases were adjusted by hand to eliminate unnecessary gaps. Comparisons were carried out only for the regions in which all three sequences were present. All alignments are available upon request.

R estimation

The number of substitutions along each of the three evolutionary branches connecting each species to the common ancestor was first obtained according to the method of Sarich and Wilson (1973). This method of estimation produced negative branch lengths in five genes, precluding further analysis. To avoid negative branch lengths, we also used parsimony to estimate the numbers of substitutions along each branch. For each variable site, an ancestral sequence was obtained by the parsimony criterion (i.e., when two of the three

sequences shared a common base). For sites in which all three sequences were different, one sequence was chosen at random to represent the ancestor. On average, 4.3 and 0.9 percent of the sites with synonymous and replacement changes, respectively, were different in all three species. For codons with two or more variable sites, the order of mutation of the sites was chosen to maximize the number of synonymous substitutions, because they are much more frequent. The observed number of substitutions along each branch was determined by comparing each sequence with the ancestral sequence. A two-parameter method to correct for multiple hits was applied to the data to yield final estimates of the number of synonymous (K_s) and nonsynonymous (K_a) substitutions per site, as described in Comeron (1995). This method is particularly suitable for pairs of species in which the number of substitutions between them is near saturation. The Index of Dispersion (R) and lineage weights (w) were estimated according to Gillespie (1989).

Weighting methods

Three different sets of weights were applied to synonymous substitutions to obtain the Index of Dispersion R : (1) synonymous weights (ws, Rs_s), (2) synonymous weights that assume equal lengths for the two *obscura* lineages (ws_{s-p}, Rs_{s-p}), and (3) nonsynonymous weights (wa, Ra_a). In a similar way, nonsynonymous substitutions for each gene were weighted using nonsynonymous weights (wa, Ra_a), nonsynonymous weights that assume equal lengths for the two *obscura* branches (wa_{s-p}, Ra_{s-p}), and synonymous weights (ws, Ra_s).

Levels of significance

Two approaches were used to test the null hypothesis of a Poisson distribution of substitutions. The first method (the Poisson approach) is equivalent to that described in Gillespie (1989). The estimated number of substitutions along a branch for a locus was corrected for lineage effects by multiplying each estimate by the appropriate lineage weight (ws or wa) to obtain the mean number per lineage. Null distributions of the six R values were obtained for each locus by producing 10,000 random Poisson-distributed numbers with a mean equal to the estimated number of substitutions for each lineage. Confidence intervals for each locus were obtained directly from each null distribution.

In the second approach (the Sequence approach), we generated for each gene a pseudo-random coding sequence with the same number of codons as that contained in the actual gene and the same percentage of G+C at third positions of codons. Synonymous and nonsynonymous substitutions were introduced onto this sequence, taken to be the common ancestor, to obtain three independent final sequences. The mean numbers of substitutions per lineage was the same as those estimated from the actual data, and for each replicate this number was assumed to be Poisson distributed. Substitutions were allowed to be transitions or transversions (biased to maintain the average G+C content in the different codon positions) with the same frequency as that estimated from the average for all the genes. Transition and transversion fractions were considered separately for synonymous and nonsynonymous substitutions. Finally, the parsimony method, described above, was used to estimate the number of synonymous and nonsynonymous substitutions along each lineage, and R was estimated as described above. Confidence intervals were obtained from the null distribution of the different estimates of R after 1,000 independent replicates for each gene.

Results

Number of synonymous and nonsynonymous substitutions

A total of 26 coding sequences were available for comparison in the three species, *D. melanogaster*, *D. subobscura*, and *D. pseudoobscura*. *Antennapedia* and *Zen*, available in GenBank for the three species, were excluded from the analysis because the sequences for these genes consisted only of the highly conserved homeodomain region, representing a very small fraction of the protein. The 24 genes contained a total of 7418 codons, with an average of 26.0% synonymous sites. Table 2 presents the number of synonymous and nonsynonymous substitutions along each lineage for each gene, estimated according to the parsimony method. The average estimated numbers of synonymous substitutions per site for the *D. subobscura*, *D. pseudoobscura*, and *D. melanogaster* lineages are 0.138, 0.153, and 0.587, respectively. Similarly, the numbers of nonsynonymous substitutions per site are 0.013, 0.011, and 0.057 for the *D. subobscura*, *D. pseudoobscura*, and *D. melanogaster* lineages, respectively. These values are very similar to the correspond-

Table 2. Estimated numbers of synonymous and nonsynonymous substitutions¹

| | Synonymous substitutions | | | Nonsynonymous substitutions | | | Codons ² |
|-----------------|--------------------------|-------------------------|------------------------|-----------------------------|-------------------------|------------------------|---------------------|
| | <i>D. subobscura</i> | <i>D. pseudoobscura</i> | <i>D. melanogaster</i> | <i>D. subobscura</i> | <i>D. pseudoobscura</i> | <i>D. melanogaster</i> | |
| <i>Adh</i> | 38.37 | 23.66 | 80.40 | 5.04 | 8.08 | 25.22 | 254 |
| <i>Adhr</i> | 21.07 | 27.14 | 180.52 | 10.62 | 3.01 | 22.37 | 272 |
| <i>A/A-T</i> | 6.64 | 13.51 | 49.48 | 4.54 | 2.01 | 17.65 | 288 |
| <i>Aprt</i> | 23.07 | 21.80 | 98.03 | 7.10 | 3.03 | 32.05 | 181 |
| <i>ATPsyn-β</i> | 26.69 | 20.02 | 225.19 | 0.00 | 1.00 | 5.02 | 323 |
| <i>Bcd</i> | 10.99 | 2.12 | 47.22 | 9.37 | 4.06 | 33.30 | 93 |
| <i>Cp15</i> | 13.80 | 18.67 | 54.90 | 7.18 | 12.43 | 60.00 | 108 |
| <i>Cp19</i> | 17.38 | 15.68 | 72.40 | 11.24 | 6.48 | 92.54 | 167 |
| <i>Cyp1</i> | 4.10 | 7.32 | 48.47 | 1.00 | 1.00 | 8.30 | 157 |
| <i>Ddc</i> | 44.95 | 52.66 | 155.57 | 10.09 | 8.06 | 33.02 | 321 |
| <i>Eno</i> | 18.66 | 17.84 | 108.04 | 2.00 | 2.00 | 42.14 | 409 |
| <i>Gad1</i> | 24.41 | 33.44 | 128.16 | 4.01 | 3.01 | 14.85 | 369 |
| <i>Gapdh2</i> | 18.89 | 59.47 | 219.32 | 1.00 | 0.00 | 10.62 | 304 |
| <i>Gld</i> | 52.47 | 61.74 | 337.96 | 8.03 | 5.02 | 68.25 | 612 |
| <i>Gpdh</i> | 23.26 | 27.93 | 171.29 | 1.00 | 1.00 | 5.18 | 350 |
| <i>Mlc1</i> | 3.11 | 0.00 | 7.58 | 0.00 | 1.00 | 3.03 | 86 |
| <i>Rh1</i> | 40.20 | 40.79 | 90.02 | 1.00 | 6.03 | 10.05 | 370 |
| <i>Rp49</i> | 6.29 | 8.56 | 43.07 | 1.00 | 0.00 | 7.63 | 134 |
| <i>Sod</i> | 10.09 | 11.07 | 48.34 | 0.00 | 1.00 | 20.37 | 114 |
| <i>Sry-α</i> | 85.18 | 122.04 | 302.32 | 74.09 | 69.55 | 232.22 | 514 |
| <i>Tpi</i> | 25.00 | 41.26 | 68.87 | 1.00 | 2.01 | 25.82 | 235 |
| <i>Uro</i> | 62.18 | 38.47 | 128.05 | 13.66 | 13.15 | 51.99 | 334 |
| <i>Vha14</i> | 3.09 | 8.72 | 76.39 | 0.00 | 0.00 | 1.00 | 90 |
| <i>Xdh</i> | 206.22 | 220.40 | 685.24 | 41.55 | 42.40 | 181.56 | 1333 |
| Average | 32.48 | 36.63 | 136.17 | 7.98 | 7.65 | 40.09 | 323 |

¹ The estimated numbers of synonymous and nonsynonymous substitutions for each lineage have been obtained as described in Material and methods.

² Codons indicate the effective number of codons where the comparison among the three species is possible.

ing estimates of the average number of substitutions per site between pairs of species, shown in Table 3.

Index of Dispersion (R)

The synonymous (ws) and nonsynonymous (wa) lineage weighting factors are shown in Table 4. These estimates of relative lineage lengths were determined by summing the 24 substitution rates in each lineage. The nonsynonymous weights are nearly identical for the two *obscura* species, but the synonymous weight is 13% larger for *D. pseudoobscura* as a consequence of the greater number of synonymous substitutions assigned to this lineage. This suggests that *D. pseudoobscura* has accumulated synonymous substitutions at a slightly faster rate than *D. subobscura*.

Estimates of R for synonymous and nonsynonymous substitutions using three different weighting

schemes are shown in Table 5. The level of significance for each value of R , determined by simulation using the Sequence approach (see Material and methods), is shown for each gene, as well as for the average and for all (concatenated) sequences. Synonymous substitutions show a significant departure ($P < 0.05$) from the neutral expectation for 12 of the 24 genes when the synonymous weighting factor (R_{s_s}) is applied. Six of the R values are significant at $P < 0.01$. The average value of R_{s_s} is 4.365 ($P < 0.03$) for synonymous substitutions.

In contrast, the average value of R for nonsynonymous substitutions is 1.638 ($P > 0.15$) when the nonsynonymous weighting factor (R_{a_s}) is applied. Only five genes show a significant departure at $P < 0.05$; none are significant at the $P = 0.01$ level. Another indication that the dispersion indices are different for synonymous and nonsynonymous changes is that this

Table 3. Estimated numbers of synonymous (K_s) and nonsynonymous (K_a) substitutions per site between pairs of species

| | <i>D. subobscura</i> - <i>D. pseudoobscura</i> | | <i>D. melanogaster</i> - <i>D. subobscura</i> | | <i>D. melanogaster</i> - <i>D. pseudoobscura</i> | |
|-----------------|---|--------|--|--------|---|--------|
| | K_s | K_a | K_s | K_a | K_s | K_a |
| <i>Adh</i> | 0.3469 | 0.0221 | 0.7177 | 0.0497 | 0.5828 | 0.0549 |
| <i>Adhr</i> | 0.2542 | 0.0222 | 1.1368 | 0.0567 | 1.1266 | 0.0442 |
| <i>A/A-T</i> | 0.0921 | 0.0094 | 0.2506 | 0.0332 | 0.2773 | 0.0293 |
| <i>Aprt</i> | 0.3344 | 0.0276 | 0.9034 | 0.1035 | 0.9808 | 0.0908 |
| <i>ATPsyn-β</i> | 0.1808 | 0.0013 | 1.0399 | 0.0067 | 1.0499 | 0.0081 |
| <i>Bcd</i> | 0.2058 | 0.0614 | 0.9698 | 0.2034 | 0.7507 | 0.1732 |
| <i>Cp15</i> | 0.3995 | 0.0856 | 0.8901 | 0.3005 | 0.9385 | 0.3136 |
| <i>Cp19</i> | 0.2548 | 0.0498 | 0.6848 | 0.2914 | 0.6956 | 0.2720 |
| <i>Cyp1</i> | 0.0972 | 0.0051 | 0.4623 | 0.0262 | 0.4642 | 0.0262 |
| <i>Ddc</i> | 0.4903 | 0.0247 | 1.0115 | 0.0583 | 1.1746 | 0.0551 |
| <i>Eno</i> | 0.1151 | 0.0042 | 0.3414 | 0.0461 | 0.3528 | 0.0455 |
| <i>Gad1</i> | 0.2535 | 0.0083 | 0.6164 | 0.0240 | 0.6827 | 0.0204 |
| <i>Gapdh2</i> | 0.3299 | 0.0031 | 1.0484 | 0.0168 | 1.0011* | 0.0145 |
| <i>Gld</i> | 0.2567 | 0.0098 | 0.9183 | 0.0546 | 1.003 | 0.0536 |
| <i>Gpdh</i> | 0.2032 | 0.0026 | 0.7904 | 0.0081 | 0.7894 | 0.0078 |
| <i>Mlc1</i> | 0.0556 | 0.0045 | 0.1875 | 0.0150 | 0.1116 | 0.0196 |
| <i>Rhl</i> | 0.3097 | 0.0085 | 0.5241 | 0.0135 | 0.5187 | 0.0205 |
| <i>Rp49</i> | 0.1198 | 0.0037 | 0.4875 | 0.0284 | 0.5354 | 0.0246 |
| <i>Sod</i> | 0.2773 | 0.0036 | 0.7030 | 0.0828 | 0.7731 | 0.0886 |
| <i>Sry-α</i> | 0.7285 | 0.1276 | 1.4431 | 0.2577 | 1.4361 | 0.2544 |
| <i>Tpi</i> | 0.4067 | 0.0053 | 0.6382 | 0.0502 | 0.6958 | 0.0519 |
| <i>Uro</i> | 0.4637 | 0.0350 | 0.9672 | 0.0833 | 0.7514 | 0.0848 |
| <i>Vha14</i> | 0.1677 | 0.0000 | 1.6392 | 0.0044 | 1.7044 | 0.0088 |
| <i>Xdh</i> | 0.5006 | 0.0281 | 1.0283 | 0.0760 | 1.1054 | 0.0717 |
| Average | 0.2852 | 0.0231 | 0.8083 | 0.0788 | 0.8126 | 0.0764 |

* K_s estimate obtained by using the Jukes and Cantor's method (1969) because of the inapplicability of Kimura's two-parameter method (1980).

Table 4. Weight factors for synonymous and nonsynonymous substitutions

| | <i>D. subobscura</i> | <i>D. pseudoobscura</i> | <i>D. melanogaster</i> |
|-------------|----------------------|-------------------------|------------------------|
| w_s | 0.475 | 0.535 | 1.990 |
| w_{s-p} | 0.505 | | 1.990 |
| w_a | 0.429 | 0.412 | 2.159 |
| w_{a-s-p} | 0.421 | | 2.159 |

difference is equally large when the alternative weighting factors are used to estimate R (Rs_a and Ra_s).

The comparison between Rs_{s-p} and Ra_{a-s-p} with Rs_s and Ra_a , respectively, identify residual effects that can be attributed to lineage differences between *D. subobscura* and *D. pseudoobscura*. The average R values, calculated with equal and unequal weighting

for the *obscura* branches, are nearly identical, indicating minimal lineage effects. For the concatenated sequences, however, the R for synonymous substitutions is nearly significant ($Rs_{s-p} = 3.965$, $P = 0.06$) and can be attributed to a faster synonymous substitution rate in *D. pseudoobscura* compared to *D. subobscura*. We will return to this point in the Discussion. In contrast, R for nonsynonymous substitutions for the concatenated sequences calculated with equal weights is nearly zero ($Ra_{a-s-p} = 0.110$), as expected under the constant rate model.

To determine whether genes with a significant departure of R from the neutral expectation had a perceptible effect on the weighting factors and, therefore, on our previous results, the weighting factors were also estimated after removing these genes. When the new weights are applied, the average Rs_s and Ra_a values,

Table 5. Estimates of R , Index of Dispersion, for synonymous and nonsynonymous substitutions among the *D. melanogaster*, *D. subobscura*, and *D. pseudoobscura* lineages

| | Synonymous substitutions | | | Nonsynonymous substitutions | | |
|-----------------|--------------------------|------------|-----------|-----------------------------|------------|----------|
| | Rs_s | Rs_{s-p} | Rs_a | Ra_a | Ra_{a-p} | Ra_s |
| | ws | ws_{s-p} | wa | wa | wa_{s-p} | ws |
| <i>Adh</i> | 6.058** | 4.440* | 6.471** | 0.835 | 0.731 | 0.263 |
| <i>Adhr</i> | 6.830* | 7.062** | 2.593 | 3.519* | 3.756* | 3.720** |
| <i>A/A-T</i> | 1.280 | 1.691 | 1.830 | 0.599 | 0.663 | 0.915 |
| <i>Aprt</i> | 0.326 | 0.137 | 0.238 | 1.064 | 1.154 | 1.798 |
| <i>ATPsyn-β</i> | 15.128*** | 15.064*** | 6.774** | 0.685 | 0.677 | 0.785 |
| <i>Bcd</i> | 5.002* | 4.691* | 3.887* | 1.315 | 1.457 | 1.831 |
| <i>Cp15</i> | 0.326 | 0.663 | 1.715 | 1.190 | 1.059 | 1.659 |
| <i>Cp19</i> | 0.340 | 0.144 | 0.191 | 3.812* | 3.868* | 7.483*** |
| <i>Cyp1</i> | 2.779 | 2.878 | 1.482 | 0.143 | 0.143 | 0.394 |
| <i>Ddc</i> | 0.857 | 1.277 | 4.441* | 0.498 | 0.560 | 0.396 |
| <i>Eno</i> | 1.846 | 1.759 | 0.187 | 4.309* | 4.312* | 6.801*** |
| <i>Gad1</i> | 0.553 | 1.092 | 1.563 | 0.128 | 0.157 | 0.193 |
| <i>Gapdh2</i> | 12.885*** | 14.962*** | 15.082*** | 1.440 | 1.430 | 1.950 |
| <i>Gld</i> | 5.510* | 5.895** | 1.334 | 2.700 | 2.735 | 5.419** |
| <i>Gpdh</i> | 4.534* | 4.722* | 1.361 | 0.001 | 0.000 | 0.043 |
| <i>Mlc1</i> | 2.099 | 1.953 | 2.102 | 0.672 | 0.655 | 0.588 |
| <i>Rhl</i> | 4.204* | 3.997* | 7.368*** | 3.403 | 3.248* | 2.380 |
| <i>Rp49</i> | 0.724 | 0.834 | 0.343 | 0.950 | 0.948 | 1.247 |
| <i>Sod</i> | 0.114 | 0.143 | 0.129 | 3.486* | 3.513 | 4.930** |
| <i>Sry-α</i> | 5.330* | 8.166** | 16.917*** | 5.111* | 5.217* | 2.004 |
| <i>Tpi</i> | 5.556** | 7.046** | 10.734** | 2.240 | 2.244 | 3.661* |
| <i>Uro</i> | 10.021*** | 7.393*** | 10.735*** | 0.396 | 0.403 | 0.115 |
| <i>Vha14</i> | 8.758** | 8.765** | 5.372* | 0.267 | 0.268 | 0.338 |
| <i>Xdh</i> | 3.707 | 3.778 | 16.563*** | 0.559 | 0.504 | 0.295 |
| Concatenated | 0.000 | 3.965 | 30.299*** | 0.000 | 0.110 | 8.114*** |
| Average | 4.365* | 4.523* | 4.975* | 1.638 | 1.654 | 2.050 |

Level of significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ based on computer simulations (see text).

4.436 and 1.793, respectively, remain nearly the same as those obtained when all genes were included in the estimation procedure. Overdispersed genes, therefore, do not contribute in any special way to the estimation of the lineage effects.

In accord with studies of R in mammals, our results indicate that synonymous and nonsynonymous substitutions in *Drosophila* have different variances in their rates of evolution. In contrast to mammalian data, however, it is the synonymous changes rather than the nonsynonymous changes that show significant rate variability across lineages. In *Drosophila*, the rates of nonsynonymous substitutions are, on average, compatible with the neutral hypothesis, whereas synonymous substitution rates are significantly more variable than expected under a constant rate model.

Significant departure of R

Gillespie (1989) suggested that multiple substitutions will bias the estimation of R , especially for synonymous substitutions. To investigate this bias, he simulated the evolution of an average sequence and applied the Jukes and Cantor (1969) one-parameter correction to obtain a 'true' R . Bulmer (1989) indicated that the expected value of R will be higher than one when there are multiple hits and that this effect will be more pronounced for highly diverged sequences and non-star phylogenies. In the present study, the null distribution of R , and hence the statistical significance of each estimated value, was determined from simulated data, applying the same method for estimating the number of substitutions along each branch (the Sequence approach) as for the actual data. Our null distribution,

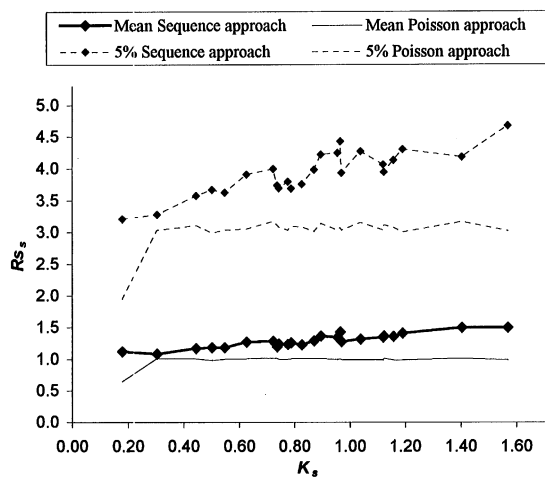


Figure 1. Average values of R_{s_s} and maximum values at 5% significance level for the 24 genes in relation to the overall numbers of synonymous substitutions per site (K_s).

therefore, takes into account the fact that the expected R under the Poisson model can be greater than one.

Figure 1 shows the average values and the 5% confidence limits of R obtained from the simulated null distribution for each gene in relation to the estimated number of synonymous substitutions per site (K_s) for each gene. This figure also compares the results of the Sequence and the Poisson approaches. As expected, the average values of R increase with high levels of K_s in the Sequence approach and they are not related to the absolute number of substitutions. The Poisson approach, in contrast, yields mean R values near one for all K_s values, underestimating the expected R . Similarly, the maximum accepted values at 5% level of significance increase drastically with K_s in the Sequence approach, while they are relatively constant for the Poisson approach.

The average R under the constant rate model reaches a value of 1.5 for genes with high rates of synonymous substitutions (such as *Sry- α* and *Vha14*), and the 5% significance level for R_{s_s} can be as high as 4.7. The linear regression of the expected R , $E(R)$, with K_s , $E(R) = (0.33K_s) + 1$ ($r^2 = 0.86$, $P < 0.001$), was estimated from the data assuming that the linear regression obeys $E(R) = 1$ for $K_s = 0$. Not surprisingly, the coefficient of variation (CV) is close to 1.0 for all genes ($CV = 0.03K_s + 1$).

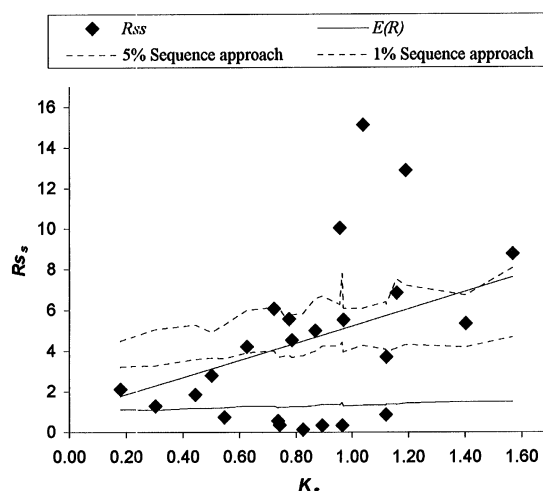


Figure 2. Relationship between R_{s_s} and the overall numbers of synonymous substitutions per site (K_s). The maximum values at 5% and 1% significance level and the expected R , $E(R)$, obtained from computer simulation are also indicated ($R_{s_s} = 4.21 * K_s + 1$; $r^2 = 0.21$, $P < 0.05$).

Relationship between R_{s_s} and the number of synonymous substitutions per site

Ohta (1995) proposed that R should increase linearly with the number of substitutions if lineage effects are the cause of differences in the substitution rates, but that it will be independent if variability in the number of substitutions is caused by episodic selection (Gillespie, 1987, 1989).

Figure 2 shows the relationship between the observed R for synonymous substitutions (R_{s_s}) and the total number of synonymous substitutions per site. Setting $R = 1$ for $K_s = 0$, there is a significant regression of R_{s_s} with K_s ($P < 0.05$). Nevertheless, as described above and as noted by Ohta (1995), this positive relationship can also arise by an upward bias of R caused by multiple hits. In our case, the positive relationship of R_{s_s} with K_s exceeds the 5% confidence interval obtained by computer simulation, suggesting an additional contribution from lineage effects. The analysis of $R_{s_{s-p}}$ values with K_s gives equivalent results ($P < 0.01$). The correlation between R_{a_a} and the number of nonsynonymous substitutions per site (Figure 3) is marginally significant, largely due to *Cp19* and *Sry- α* having high values of both R_{a_a} and K_a ($r^2 = 0.17$, $P < 0.05$). As expected, $E(R)$ for nonsynonymous substitutions is little affected by the multiple hits effect.

If the observed correlation between R_{s_s} and K_s is related to the rate of synonymous substitution, it

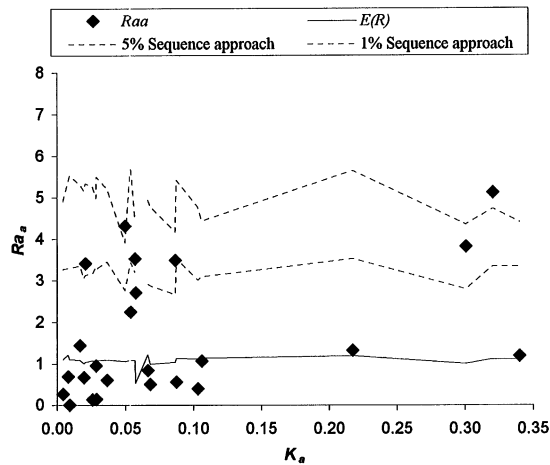


Figure 3. Relationship between Ra_a and the overall numbers of nonsynonymous substitutions per site (K_a). The maximum values at 5% and 1% of significance level and the expected R , $E(R)$, obtained from computer simulation, are also indicated.

should also be observed for the K_s values in each of the three lineages. Rs_s shows a significant correlation with K_s , however, only in the *D. melanogaster* lineage: $r^2 = 0.28$ ($P < 0.01$) for the *D. melanogaster* lineage alone; $r^2 = 0.27$ ($P < 0.01$) for the *D. subobscura*-*D. melanogaster* lineage; and $r^2 = 0.25$ ($P < 0.05$) for the *D. pseudoobscura*-*D. melanogaster* lineage. The relationship between Rs_s and K_s is nonsignificant for the two *obscura* lineages. These results suggest that the correlation between Rs_s and K_s is mainly due to genes with high K_s , which in the *D. melanogaster* lineage have a higher than expected number of synonymous substitutions.

Lack of correlation between Rs_s and Ra_a

As pointed out by Gillespie (1986a), if mutation rate changes were the cause of differences in rates of substitutions, then a correlation would be expected between R values for synonymous and nonsynonymous substitutions. As shown in Figure 4, there is no relationship between Rs_s and Ra_a ($r^2 = 0.01$, $P > 0.50$). As expected, significant correlations are observed between Rs_s and Rs_a , the two measures of the synonymous dispersion index ($r^2 = 0.31$, $P < 0.01$), and between Ra_s and Ra_a , the two measures of the nonsynonymous dispersion index ($r^2 = 0.65$, $P < 0.001$). Taken together, these results again indicate that synonymous and nonsynonymous substitutions are evolving under different evolutionary forces.

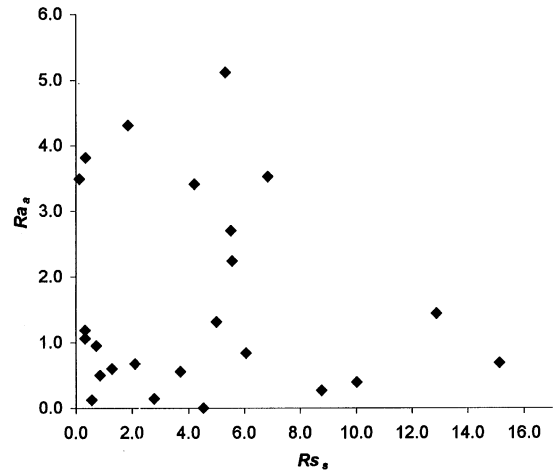


Figure 4. Relationship between the Index of Dispersion for synonymous (Rs_s) and nonsynonymous (Ra_a) substitutions.

Discussion

This study addresses two principal questions about rates of molecular evolution. First, is protein evolution overdispersed in *Drosophila* lineages, as it is in mammalian lineages? Second, are the patterns of synonymous and nonsynonymous substitutions different? The data provide unambiguous answers to both questions. First, rates of protein evolution are relatively constant across the three *Drosophila* lineages. The average nonsynonymous R taken over the 24 genes is only 1.64, and it is not significantly different from the neutral theory's prediction of $R = 1$. Only 5 out of the 24 genes exhibit significantly large values of R , whereas R is less than one for 12 of the 24 genes. We tentatively conclude that the data are consistent with genetic drift as a driving force in protein evolution in *Drosophila*.

Synonymous substitutions, on the other hand, exhibit significantly high values of R for 12 of the 24 genes, as well as for the average ($R = 4.365$, $P < 0.03$). Before discussing possible causes of this rate variation, we will first comment on methodological issues concerning the estimation procedures. In the present study, the levels of significance of R for the different genes, as well as for the average, were obtained by computer simulation, thus removing many of the factors that can bias the estimation procedure. Our analysis considered the multiple hits effect, the putative effect of K_s/K_a ratio, the G+C content, the transition/transversion ratio, lineage weighting methods, and the method for estimating the numbers of synonymous and nonsynonymous substitutions in each lineage. Using parsimony to estimate

the number of substitutions along each lineage rather than the standard method of Sarich and Wilson (1973) allowed us to obtain positive branch lengths for every gene. In addition, it allowed us to consider transitions and transversions separately, so that we could estimate the number of multiple hits with a two-parameter method. Both methods yielded similar estimations of R . Thus, neither of the two major results—the small Index of Dispersion for nonsynonymous substitutions and the significantly large Index of Dispersion for synonymous substitutions—can be the result of any of these factors.

Detection of lineage effects

Previous studies (Gillespie, 1989; Ohta, 1993, 1995) compared unweighted with weighted estimates of R to examine the influence of lineage effects on R . Because *D. melanogaster* is clearly the outgroup of the two *obscura* group species, we included this obvious lineage effect in all of the analyses. The substitution rate per year may also be higher in *D. melanogaster* than in the *obscura* lineages because it has a shorter generation time (Goddard, Caccone & Powell, 1990).

Other lineage effects could only be studied between the two *obscura* species. Because *D. subobscura* and *D. pseudoobscura* are sister taxa, the two lineages have evolved for the same absolute time. Furthermore, they can be assumed to have approximately the same generation time. As noticed above, synonymous substitutions show a high value of R when the same weight is given to the two *obscura* species, reflecting the fact that they have accumulated at a 13% higher rate in the *D. pseudoobscura* lineage than in the *D. subobscura* lineage. This difference does not appear to be related to mutation rates, as the nonsynonymous substitution rate estimates are nearly equal in the two lineages (Table 2). We, therefore, consider an alternative explanation.

Akashi (1995) estimated the strength of selection acting on synonymous mutations by comparing polymorphism and divergence of synonymous substitutions in five genes of *D. melanogaster* and *D. simulans*. By categorizing mutations appropriately (either towards or away from major codons), he showed that these synonymous mutations in *D. simulans* are nearly neutral. He also detected a substantially greater number of synonymous substitutions in the *D. melanogaster* lineage, with the vast majority of changes being away from a major codon. He suggested that a smaller population size in *D. melanogaster* may have resulted in a relaxation of selection against these slightly dele-

rious mutations and, hence, in a higher fixation rate in this lineage.

Similarly, we suggest that greater fluctuation in the population size of *D. pseudoobscura* over its evolutionary history might have resulted in a slight overall increase in the number of synonymous substitutions. Nucleotide polymorphism data at the *Xdh* locus for these two species suggests that the current effective population size may be larger in *D. subobscura* than in *D. pseudoobscura* (Comeron, 1997). This is unlikely, however, to have been the case during most of the species' histories because the average codon bias, as measured by ENC, the effective number of codons, (Wright, 1990) is very similar for the 24 genes (ENC [*D. subobscura*] = 40.07, ENC [*D. pseudoobscura*] = 40.33). We suggest, therefore, that the higher rate of synonymous substitution in *D. pseudoobscura* compared to *D. subobscura* may be the result of weak selection for codon bias and greater fluctuation in population size in the *D. pseudoobscura* lineage. In spirit, this is the intersection of Gillespie's episodic selection model with Ohta's nearly neutral model, with both population size fluctuation and positive selection playing important roles in the process. With population size fluctuation of sufficiently long duration and appropriate magnitude, the synonymous substitution rate will be accelerated by the fixation of slightly deleterious mutations by genetic drift when population size is smaller as well as by the fixation of slightly advantageous mutations by selection when the population size is larger. In principle, episodic weak selection might also have contributed to a higher than expected Index of Dispersion for synonymous substitutions.

Relationship between the Index of Dispersion and codon bias

Are the differences in the estimates of R among genes also related to codon bias? Synonymous substitutions in *Drosophila* have been shown to be inversely related to the magnitude of codon bias (Sharp & Li, 1989): genes with strong codon bias have relatively low rates of synonymous substitution. We found significantly higher values of R_{s_s} in genes with high K_s . In a study that will be published separately, we have analyzed the relationship between K_s and codon bias for the three independent lineages. The results show that K_s is significantly correlated with the average codon bias (ENC) for the *D. melanogaster* lineage ($r^2 = 0.26$, $P < 0.01$) but not for the *D. subobscura* or *D. pseudoobscura* lineages ($r^2 = 0.08$, $P > 0.10$) and $r^2 = 0.15$

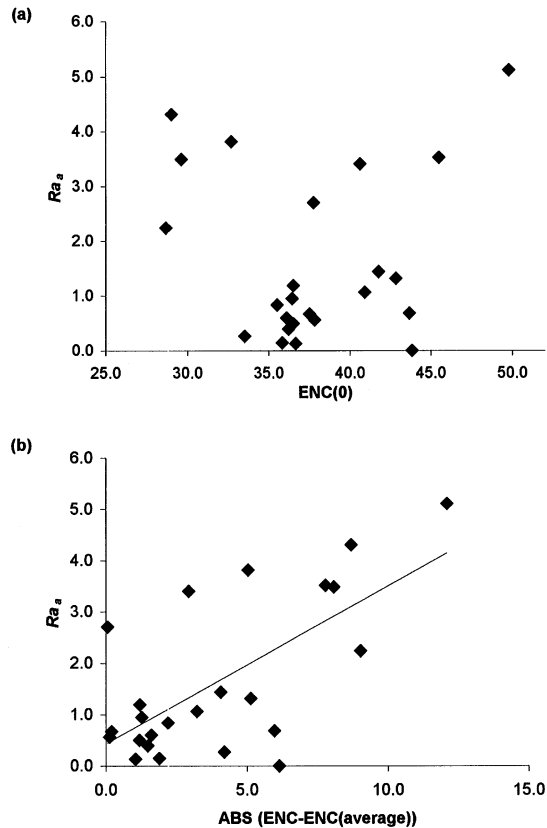


Figure 5. Relationship between Ra_a and, (a) the codon bias (measured as ENC, effective number of codons) and (b) the absolute difference of ENC for each gene from the average ENC.

($P > 0.05$), respectively). Because we also find a positive correlation between K_s and R_s in the *D. melanogaster* lineage, R_s may also be influenced by codon bias selection in this lineage.

The relationship between K_s and K_a is significant for both *obscura* lineages ($r^2 = 0.36$, $P < 0.01$, $r^2 = 0.28$, $P < 0.01$ for *D. subobscura* and *D. pseudoobscura*, respectively), but not for the *D. melanogaster* lineage ($r^2 = 0.0001$, $P > 0.90$). Perhaps more surprisingly, nonsynonymous substitutions also exhibit an interesting relationship with codon bias. In the *D. melanogaster* lineage, high K_a is associated with either low or high codon bias, whereas a low K_a is restricted to genes with intermediate codon bias. The correlation between Ra_a and the absolute departure of ENC from the average ENC is highly significant ($r^2 = 0.45$, $P < 0.001$; Figure 5b), whereas the overall correlation between ENC and Ra_a is nonsignificant ($r^2 = 0.03$) (Figure 5a).

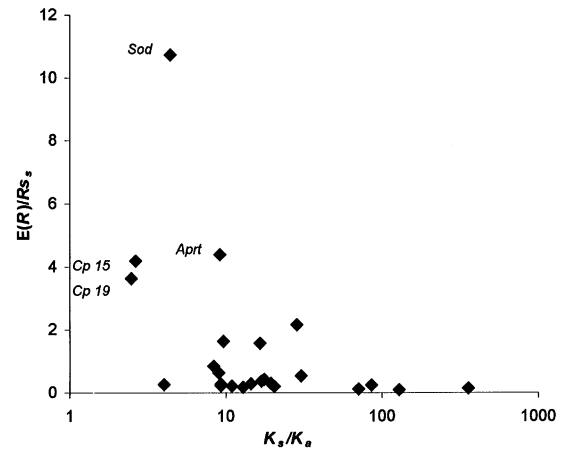


Figure 6. Relationship between the ratio K_s/K_a (logarithmic scale) in the *D. melanogaster* lineage and the ratio of the expected to the estimated R for synonymous substitutions ($E(R_s)/R_s$). $E(R)/R_s = 3.2(K_s/K_a)^{-0.66}$, $r^2 = 0.37$, $P < 0.001$.

Although these results suggest the existence of an evolutionary mechanism coupling codon bias selection with amino acid substitution, we refrain from speculating about it until this putative relationship is confirmed with independent datasets. The analysis does indicate, however, that codon bias selection may be relevant for understanding both the average and the variance in the rate of molecular evolution.

Values of R lower than 1

Seven genes exhibit R_s values lower than the $E(R)$, and four (*Sod*, *Aprt*, *Cp15*, and *Cp19*) show R_s values in the range 0.1–0.5. Although the R_s values for these three genes are not low enough to be significant, they warrant further examination. For nonsynonymous substitutions, 12 genes out of the 24 show a $Ra_a < 1$. *Sod* and *Cp19*, however, show a significant overdispersion of nonsynonymous substitutions, related to an increased substitution rate in the *D. melanogaster* lineage. We therefore examined the relationship between R_s (relative to $E[R]$) and K_a (relative to K_s) in the *D. melanogaster* lineage. Figure 6 shows that high relative nonsynonymous rates are strongly associated with low synonymous dispersion indices. Equivalent results are obtained when the substitutions are from all lineages.

Thus, even though nonsynonymous substitutions have accumulated at a nearly constant rate in *Drosophila*, there is nevertheless an inverse relationship between nonsynonymous and synonymous R . Gillespie, in his series of papers on substitution processes (1993, 1994a,

b) has indicated that $R < 1$ can be a property of overdominance and TIM models (Takahata, Tishi & Matsuda, 1975; Takahata & Kimura, 1979) in a rapidly changing environment. A potential explanation might be a scenario in which constraints on protein structure/function influence both K_s and K_a , as would be expected if there is selection for translational accuracy (Akashi, 1994). This may lead to more nearly constant rates of synonymous substitution for highly conserved proteins. But, although there is a significant correlation between K_s and K_a and between K_a and codon bias in the *obscura* species, no significant correlations were found in the *D. melanogaster* lineage. This suggests that if such a mechanism is operating, it may not be pervasive.

Comparison between Drosophila and mammalian lineages

Unlike mammals, protein evolution is relatively constant in *Drosophila*, conforming to a key prediction of the neutral theory. Protein evolution is either not subject to episodic selection in *Drosophila*, as it may be in mammals, or episodic selection is not the cause of the high Index of Dispersion in mammals. Given that population sizes are larger in *Drosophila* than in mammals, positive selection due to environmental changes would be expected to be more, rather than less, efficacious in *Drosophila*. It is, of course, possible that a larger proportion of the selection coefficients for amino acid replacement changes is close to the reciprocal of mammalian population sizes. The rate of protein evolution would then be sensitive to fluctuations in population size in mammals, but not necessarily so in *Drosophila*. On the other hand, synonymous substitutions are likely to be nearly neutral in *Drosophila*, and the rate of synonymous evolution may be sensitive to population size in these species (Akashi, 1994, 1995). Mammalian genes do not exhibit a large range of codon usage bias, suggesting that population sizes are too small to support its evolution. Therefore, population size may be an important determinant of variation in the rates of both protein and synonymous substitutions, but with different consequences in different groups of organisms.

Acknowledgements

We thank Stavroula Assimacopoulis and Brian Charlesworth for carrying out *in situ* hybridization studies, Barbara Stranger for screening cDNA

libraries, and Marcos Antezana and Molly Przeworski for useful comments about the manuscript. This work was supported by NIH grant 1P01GM50355 to M. Kreitman.

References

- 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.
- Benson, A.R., 1995. The molecular evolution of the *obscura* group *Chorion s15*: A prominent role for codon bias. PhD thesis, Harvard University.
- Britten, R.J., 1986. Rates of DNA sequence evolution differ between taxonomic groups. *Science* 231: 1393–1398.
- Bulmer, M., 1989. Estimating the variability of substitution rates. *Genetics* 123: 615–619.
- Bulmer, M., K.H. Wolfe, & P.M. Sharp, 1991. Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. *Proc. Natl. Acad. Sci. USA* 88: 5974–5978.
- Chao, L. & D.E. Carr, 1993. The molecular clock and the relationship between population size and generation time. *Evolution* 47: 688–690.
- Comeron, J.M., 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.* 41: 1152–1159.
- Comeron, J.M., 1997. Estudi de la variabilitat nucleotídica a *Drosophila*: RegiÓ *Xdh* a *D. subobscura*. PhD thesis. Barcelona, Spain. Universitat de Barcelona.
- Eastale, S., 1988. Rate constancy of globin gene evolution in placental mammals. *Proc. Natl. Acad. Sci.* 85: 7622–7626.
- Eastale, S., 1990. The pattern of mammalian evolution and the relative rate test of molecular evolution. *Genetics* 124: 165–173.
- Eastale, S. & C. Collet, 1994. Consistent variation in amino-acid substitution rate, despite uniformity of mutation rate: Protein evolution in mammals is not neutral. *Mol. Biol. Evol.* 11: 643–647.
- Gillespie, J.H., 1984. The molecular clock may be an episodic clock. *Proc. Natl. Acad. Sci. USA* 81: 8009–8013.
- Gillespie, J.H., 1986a. Variability of evolutionary rates of DNA. *Genetics* 113: 1077–1091.
- Gillespie, J.H., 1986b. Rates of molecular evolution. *Annu. Rev. Ecol. Syst.* 17: 637–665.
- Gillespie, J.H., 1987. Molecular evolution and the neutral allele theory. *Oxford Surveys Evol. Biol.* 4: 10–37.
- Gillespie, J.H., 1989. Lineage effects and the index of dispersion of molecular evolution. *Mol. Biol. Evol.* 6: 636–647.
- Gillespie, J.H., 1991. *The Causes of Molecular Evolution*. Oxford series in Ecology and evolution. Oxford University Press. New York.
- Gillespie, J.H., 1993. Substitution processes in molecular evolution. I. Uniform and clustered substitutions in haploid model. *Genetics* 134: 971–981.
- Gillespie, J.H., 1994a. Substitution processes in molecular evolution. II. Exchangeable models from population genetics. *Evolution* 48: 1101–1113.

- Gillespie, J.H., 1994b. Substitution processes in molecular evolution. III. Deleterious alleles. *Genetics* 138: 943–952.
- Goddard, K., A. Caccone & J.R. Powell, 1990. Evolutionary implications of DNA divergence in the *Drosophila obscura* group. *Evolution* 44: 1656–1670.
- Jukes, T. H., & 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.
- Kimura, M., 1969. The rate of molecular evolution considered from the standpoint of population genetics. *Proc. Natl. Acad. Sci. USA* 63: 1181–1188.
- Kimura, M., 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
- Kimura, M. & T. Ohta, 1971. On the rate of molecular evolution. *J. Molec. Evol.* 1: 1–17
- King, J.L. & T.H. Jukes, 1969. Non-Darwinian evolution. *Science* 164: 788–798.
- Kreitman, M., 1983. Nucleotide polymorphism at the *alcohol dehydrogenase* locus of *Drosophila melanogaster*. *Nature* 304: 412–417.
- Langley, C.H. & W.M. Fitch, 1973. The constancy of evolution: a statistical analysis of the α and β haemoglobins, cytochrome c, and fibrinopeptide A, pp. 246–262 in *Genetic Structure of Populations*, edited by N.E. Morton, Univ. of Hawaii Press, Honolulu.
- Langley, C.H. & W.M. Fitch, 1974. An estimation of the constancy of the rate of molecular evolution. *J. Mol. Evol.* 3: 161–177.
- Li, W.-H., 1997. *Molecular evolution*. Sinauer Assoc., Inc.
- Li, W.-H. & D. Graur, 1991. *Fundamentals of Molecular Evolution*. Sinauer Assoc., Inc., Sunderland.
- Li, W.-H., M. Gouy, P. M. Sharp, C.O'Huigin & Y.-W. Yang, 1990. Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. *Proc. Natl. Acad. Sci. USA* 87: 6703–6707.
- Li, W.-H., M. Tanimura & P. M. Sharp, 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J. Mol. Evol.* 25: 330–342.
- Margoliash, E., 1963. Primary structure and evolution of cytochrome c. *Proc. Natl. Acad. Sci. USA* 50: 672–679.
- Martin, A. P., & S. R. Palumbi, 1993. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. USA* 90: 4087–4091.
- Nei, M. & D. Graur, 1984. Extent of protein polymorphism and the neutral mutation theory. *Evol. Biol.* 17: 73–118.
- Ohta, T., 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246: 96–98.
- Ohta, T., 1991. Multigene families and the evolution of complexity. *J. Mol. Evol.* 33: 34–41.
- Ohta, T., 1992. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* 23: 263–286.
- Ohta, T., 1993. An examination of the generation-time effect on molecular evolution. *Proc. Natl. Acad. Sci. USA* 90: 10676–10680.
- Ohta, T., 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* 40: 56–63.
- Ohta, T. & M. Kimura, 1971. On the constancy of the evolutionary rate of cistrons. *J. Mol. Evol.* 1: 18–25.
- Riley, M., M.E. Hallas & R.C. Lewontin, 1989. Distinguishing the forces controlling variation at the *Xdh* locus in *Drosophila pseudoobscura*. *Genetics* 123: 359–369.
- Sarich, V.M. & A.C. Wilson, 1973. Generation time and genomic evolution in Primates. *Science* 179: 1144–1147.
- Schaeffer, S.W., C.F. Aquadro & W.W. Anderson, 1987. Restriction-map variation in the *alcohol dehydrogenase* region of *Drosophila pseudoobscura*. *Mol. Biol. Evol.* 4: 254–265.
- Sharp, P.M., & W.-H. Li, 1989. On the rate of DNA sequence evolution in *Drosophila*. *J. Mol. Evol.* 28: 398–402.
- Takahata, N. & M. Kimura, 1979. Genetic variability maintained in a finite population under mutation and autocorrelated random fluctuation of selection intensity. *Proc. Natl. Acad. Sci. USA* 76: 5813–5817.
- Takahata, N., K. Iishi & H. Matsuda, 1975. Effect of temporal fluctuation of selection coefficient on gene frequency in a population. *Proc. Natl. Acad. Sci. USA* 72: 4541–4545.
- Thomson, J.D., D.G. Higgins & T.J. Gibson, 1994. CLUSTAL W: improving the sensitivity of progressive sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673–4680.
- Wang, D., J.L. Marsh & F.J. Ayala, 1996. Evolutionary changes in the expression pattern of a developmentally essential gene in three *Drosophila* species. *Proc. Natl. Acad. Sci. USA* 93: 7103–7107.
- Wallis, M., 1996. The molecular evolution of vertebrate growth hormones: A pattern of near-stasis interrupted by sustained bursts of rapid change. *J. Mol. Evol.* 43: 93–100.
- Wells, R.S., 1996. Nucleotide variation at the *Gpdh* locus in the genus *Drosophila*. *Genetics* 143: 375–384.
- Wright, F., 1990. 'The effective number of codons' used in a gene. *Gene* 87: 23–39.
- Wu, C.-I. & W.-H. Li, 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. USA* 82: 1741–1745.
- Zeng, L.-W. & M. Kreitman, 1996a. Simple strategy for sequencing cDNA clones. *Biotechniques* 1996 Sep; 21(3): 446–452
- Zeng, L.-W. & M. Kreitman, 1996b. Rapid and cost-effective DNA sequencing strategy for PCR products. *Trends in Genetics, Technical Tips Online #TL10017*.
- Zuckerandl, E. & L. Pauling, 1962. Molecular disease, evolution, and genetic heterogeneity, pp. 189–225 in *Horizons in Biochemistry*, edited by M. Kasha and B. Pullman. Academic Press. New York.
- Zuckerandl, E. & L. Pauling, 1965. Evolutionary divergence and convergence in proteins, pp. 97–166 in *Evolving Genes and Proteins*, edited by V. Bryson and H.J. Vogel. Academic Press. New York.