J Mol Evol (2000) 50:1–10
DOI: 10.1007/s002399910001

JOURNAL OF MOLECULAR EVOLUTION

Fluctuating Mutation Bias and the Evolution of Base Composition in *Drosophila*

Francisco Rodríguez-Trelles, Rosa Tarrío, Francisco J. Ayala

Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, CA 92697-2525, USA

Received: 4 May 1999 / Accepted: 26 July 1999

Abstract. The idea that the pattern of point mutation in *Drosophila* has remained constant during the evolution of the genus has recently been challenged. A study of the nucleotide composition focused on the *Drosophila saltans* group has evidenced unsuspected nucleotide composition differences among lineages. Compositional differences are associated with an accelerated rate of amino acid replacement in functionally less constrained regions. Here we reassess this issue from a different perspective. Adopting a maximum-likelihood estimation approach, we focus on the different predictions that mutation and selection make about the nonsynonymous-to-synonymous rate ratio. We investigate two gene regions, alcohol dehydrogenase (*Adh*) and xanthine dehydrogenase (*Xdh*), using a balanced data set that comprises representatives from the *melangaster*, *obscura*, *saltans*, and *willistoni* groups. We also consider representatives of the Hawaiian picture-winged group. These Hawaiian species are known to have experienced repeated bottlenecks and are included as a reference for comparison. Our results confirm patterns previously detected. The branch ancestral to the fast-evolving *willistoni*/*saltans* lineage, where most of the change in GC content has occurred, exhibits an excess of synonymous substitutions. The shift in mutation bias has affected the extent of the rate variation among sites in *Xdh*.

Key words: Mutation bias — Nucleotide composition — Nonsynonymous/synonymous rate ratio — Among-site rate heterogeneity — *willistoni* group — *saltans* group

Correspondence to: Francisco Rodríguez-Trelles c/o Francisco J. Ayala; *e-mail:* ibge2@blues.uab.es

Introduction

The pattern of point mutation is generally thought to have been a negligible source of heritable variation during the diversification of the *Drosophila* genus, initiated around 60 My ago (Fitch and Ayala 1994; Powell and DeSalle 1995). This assumption is based on the observations that (1) nucleotide composition varies little in introns (Shields et al. 1988; Moriyama and Hartl 1993; Kliman and Hey 1994) and other allegedly unconstrained regions (Petrov and Hartl 1999), and (2) the pattern of codon usage is fairly homogeneous across species, except when differences can be accounted for by changes in natural selection (Akashi 1995, 1996; Akashi and Schaffer 1997). These studies, however, have been restricted largely to two species, *D. melanogaster* and *D. pseudoobscura*, of the *Sophophora* subgenus, and *D. virilis* of the subgenus *Drosophila*, all three of which have quite similar base compositions (reviewed by Powell 1997).

The idea of the constancy of the pattern of point mutation in the evolution of *Drosophila* has recently been challenged (Rodríguez-Trelles et al. 1999b). The not previously investigated *Drosophila saltans* group exhibits patterns of GC content and codon usage bias markedly different from those previously known in *Sophophora.* The GC content in the third codon position, and to a lesser extent in the first position and the introns, is higher in the *D. melanogaster* and *D. obscura* groups than in the

D. saltans group. Differences are greater for the xanthine dehydrogenase (*Xdh*) region than for the alcohol dehydrogenase (*Adh*), superoxide dismutase (*Sod*), period (*Per*), and *28SrRNA* regions, which are functionally more constrained. In addition, the *saltans* group shows an increased rate of amino acid substitution in *Xdh*, with the new replacements occurring preferentially by amino acids encoded by low-GC content codons. These observations are best explained by a shift in the pattern of point mutation that occurred in the ancestor of the *saltans* lineage, after it split from the lineage that gave rise to the *melanogaster* and *obscura* groups (Rodríguez-Trelles et al. 1999b). Alternatively, they can be explained without invoking changes in the mutational spectrum, either by a change in protein function (i.e., positive selection) or by a reduction in the effectiveness of selection (due to diminished population numbers or reduced recombination) that generally counteracts the underlying mutation bias significant [i.e., in the absence of selection, the nucleotide composition of functionally relevant regions would move toward that exhibited by the less constrained regions of the genome, i.e., ≈65% AT in *Drosophila* (Shields et al. 1988; Moriyama and Hartl 1993; Kliman and Hey 1994; Petrov and Hartl 1999)]. In light of the available evidence, these competing hypotheses have received less credit, yet no statistical arguments have been so far provided to resolve the issue.

In this paper, we tackle the above hypotheses on quantitative terms. We concentrate on the different predictions that mutation and natural selection make about the ratio of nonsynonymous-to-synonymous substitutions (d_N/d_S) for lineages. Specifically, if the shift in GC pressure had actually occurred in the ancestor of *saltans*, it would be reflected more in neutral parts of the genome than in functionally significant parts, so that we would expect a reduced d_N/d_S ratio in the critical branch; on the contrary, most frequent scenarios of natural selection predict an increase in the d_N/d_S ratio. We focus on two loci, *Adh* and *Xdh*, from which a large number of nucleotide sequences is available. It is possible in this way to choose a balanced data set that includes three representatives of each of the four major species groups of *Sophophora* (namely, the *melanogaster*, *obscura*, *willistoni*, and *saltans* groups), with known phylogenetic relationships. The *willistoni* group is more closely related to the *saltans* group than to the other species groups, but it has not been investigated before in this respect. Also, we include three Hawaiian species of the picture-winged group (available only for *Adh*). The species of this Hawaiian group are known to have experienced repeated episodes of relaxed effectiveness of selection due to reduced population numbers (DeSalle and Templeton 1988; Ohta 1993); consequently, these species constitute a valuable reference to evaluate the d_N/d_S ratios in the branches of interest. We chose the maximum-likelihood approach developed by Yang (1998) (see also Yang and Nielsen 1998) for estimating the nonsynonymous and synonymous substitution rates, which allows the d_N/d_S ratio to vary among lineages. While time-consuming, maximum-likelihood methods are preferable over conventional methods because they allow us to accommodate more realistic models of evolution (Yang 1998; Yang and Nielsen 1998). Finally, we investigate the effects that fluctuating mutation bias has had on the extent of the among-site rate variation of the *Xdh* gene.

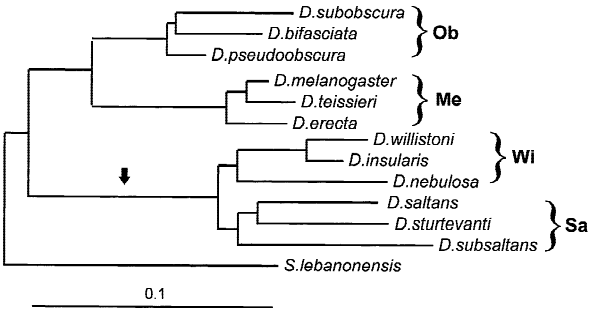

Fig. 1. The maximum-likelihood tree of the *Xdh* amino acid sequences (Hawaiian representatives not included), obtained using the empirical model of Jones et al. (1992; Yang et al. 1998), allowing discrete gamma-distributed rates (eight rate categories) among sites. The tree is obtained with the PAML 1.4 program (Yang 1997). A shift in nucleotide composition is inferred along the branch pointed out by the *arrow*. Ob, Me, Wi, and Sa represent the *obscura*, *melanogaster*, *willistoni*, and *saltans* groups, respectively.

Materials and Methods

Species and Sequences. We have examined 17 drosophilid species, with known phylogenetic relationships, as shown in Figs. 1, 5, and 6 [these hypotheses are supported by data of several sorts (see Powell 1997, Tatarenkov et al. 1999); for the species of the *saltans* group, see Rodríguez-Trelles et al. (1999a)]. The *Adh* and *Xdh* gene sequences include three representatives of each of the four major species groups of the *Sophopora* subgenus, as follows. *D. melanogaster* (GenBank accession numbers X78384 and Y00307 for *Adh* and *Xdh*, respectively), *D. teissieri* (X54118 and AF169401), and *D. erecta* (X54116 and AF169400) from the *melanogaster* group; *D. pseudoobscura* (U64560 and M33977), *D. subobscura* (X55391 and AF058976-7, Y08237), and *D. bifasciata* (AF169402-03 for *Xdh*) from the *obscura* group; *D. willistoni* (U95251 and AF093206), *D. insularis* (U95273 and AF093210), and *D. nebulosa* (U95275 and AF93213) from the *willistoni* group; and *D. saltans* (AF045113 and AF058978), *D. sturtevanti* (AF045114 and AF058983), and *D. subsaltans* (AF045117 and AF058980) from the *saltans* group. Three *Adh* sequences from representatives of the Hawaiian picture-winged group of the subgenus *Drosophila*, namely, *D. silvestris* (M6321), *D. affinidisjuncta* (M37262), and *D. picticornis* (M63392), and one *Xdh* sequence from *Scaptodrosophila lebanonensis* (AF058984), are used as outgroups. Except those from *D. melanogaster* and *D. pseudoobscura*, the *Xdh* sequences have been obtained in our laboratory. The *Xdh* region investigated includes about half of exon 2 (371 codons), intron 2 (around 60 bp except for *D. subobscura*, where it is 528 bp long, and *D. melanogaster* and *D. erecta*, where it is about 250 bp long), and most of exon 3 (324 codons), or about 52% of the *Xdh* coding region. Details about the amplification and sequencing primers and strategy are given by Tarrío et al. (1998).

The sequences of *Adh* are obtained from the literature and consist of 135 codons of exon 2. The available *Adh* sequence from *D. bifas-*

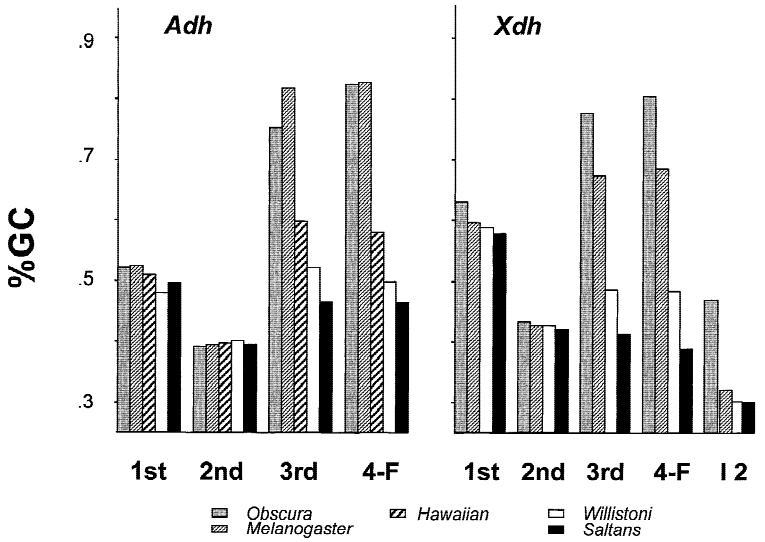

Fig. 2. GC content in codon positions and fourfold degenerate sites (4-F) of genes, and in intron 2 (I2) of *Xdh*, for the species groups in this study.

ciata is incomplete; therefore we have used instead the sequence of *D. ambigua* (X54813), which is not expected to alter significantly the conclusions of this study.

Coding sequences were aligned using the CLUSTAL W (vs. 1.5) program (Thompson et al. 1994) and no gaps were observed.

GC Content, Codon-Use Bias, and Amino Acid Composition. The nucleotide composition of the sequences is determined by the G + C content. As a measure of departure from optimal codon use in *D. melanogaster*, we use the *Fop* index (Ikemura 1985) with the set of major codons defined by Akashi (1995). To study the relationship between the nucleotide and the amino acid composition of the sequences, we classify amino acids into three groups, according to codon GC content (see Li 1997). Group I consists of codons with high GC: alanine (A), glycine (G), proline (P), and tryptophan (W) (e.g., alanine is encoded by GCU, GCC, GCA, or GCG). Group II consists of codons with an intermediate GC content: cystein (C), aspartic acid (D), glutamic acid (E), histidine (H), glutamine (Q), serine (S), threonine (T), and valine (V) (e.g., aspartic acid is encoded by either GAU or GAC). Group III consists of codons with a low GC content: phenylalanine (F), isoleucine (I), lysine (K), methionine (M), asparagine (N), and tyrosine (Y) (e.g., phenylalanine is encoded by either UUU or UUC). Arginine (R) and leucine (L) are not included in these groups, because R is encoded by an intermediate (AGA, AGG) as well as a high-GC codon family (CGU, CGC, CGA, CGG), and L is encoded by a low-GC (UUA, UUG) and an intermediate-GC (CUU, CUC, CUA, CUG) codon family. If amino acid frequencies are impacted by nucleotide composition, the frequency of group I, *f*(I), will increase and *f*(III) will decrease as the GC content increases, while *f*(II) will change little. The association among GC, *f*(I), *f*(II), and *f*(III), is investigated with the Pearson lineal coefficient of correlation. Since the species are part of a hierarchically structured phylogeny, treating them as statistically independent observations may lead to overestimation of the nominal significance level in hypothesis testing (Felsenstein 1985). Therefore we have also studied the previous correlation by means of Felsenstein's (1985) pairwise independent contrast test. Contrast tests are performed with the Contrast program in the computer package Phylip 3.5 (Felsenstein 1993).

Parameter Estimation and the Likelihood-Ratio Test. Estimation of the numbers of synonymous and nonsynonymous substitution rates per site (d_S and d_N) in evolutionary lineages is conducted by maximum likelihood. Unlike approximate methods (e.g., Nei and Gojobori 1986; Li 1993; Ina 1995), maximum-likelihood methods properly accommodate transition/transversion rate biases and codon-usage biases, factors that are very important in the estimation of d_S and d_N (Yang 1998; Yang and Nielsen 1998). Furthermore, the likelihood approach is applicable to joint comparison of multiple sequences (Goldman and Yang 1994; Yang 1998; Yang and Nielsen 1998). In contrast to the use of parsimony-reconstructed ancestral states as observed data, which involve random errors and systematic biases, maximum-likelihood methods average over all possible ancestral sequences at each interior node in a tree, which is weighted appropriately according to the relative likelihood of occurrence (Yang 1998).

Maximum-likelihood methods assume a tree topology and a model of sequence change. Given a tree topology, the probability of observing the data, $f(x_i, \theta)$, is used as the likelihood function for estimating the parameter θ, which includes branch lengths, and parameters in the substitution model. As a tree topology for the species in this study we use the hypotheses shown in Figs. 1 and 5. We use the codon-substitution model of Goldman and Yang (1994), with a single distance between any pair of amino acids (Yang 1998; Yang and Nielsen 1998); codon frequencies are calculated using the nucleotide frequencies at the three codon positions. Besides codon-based models, we also use nucleotide- and amino acid-based models to compare the rates of amino acid substitution between *Adh* and *Xdh* and to investigate the rate variation among sites along the *Xdh* region. The transition probability matrixes of models and details on parameter estimation are given by Yang (1998).

Likelihood-ratio tests are applied to test several hypotheses of interest. For a given tree topology (i.e., Fig. 1), a model (H_1) containing p free parameters and with log-likelihood L_1 fits the data significantly better than a nested submodel (H_0) with $q = p - n$ restrictions and likelihood L_0 if the deviance $D = -2\log\Lambda = -2(\log L_1 - \log L_0)$ falls in the rejection region of a χ^2 distribution with n degrees of freedom (Yang 1996). We use several starting values in the iterations to guard against the possible existence of multiple local optima. These analyses are conducted with the BASEML and CODEML programs from the PAML vs 1.4 package (Yang 1997).

Results

GC Content, Codon Usage, and Amino Acid Composition. Figure 2 shows the GC-content variation of the *Adh* and *Xdh* gene regions across species groups for each codon position and fourfold degenerate sites. Also shown

are the GC-content values of intron 2 of *Xdh*. The results confirm the trends already reported (Rodríguez-Trelles et al. 1999b); the base composition varies extensively and consistently among the lineages of *Sophophora* in both the *Adh* and the *Xdh* gene regions. The compositional pattern of the *D. willistoni* group, which had not been investigated before, is very similar to that of *D. saltans*, its sister clade; the GC content in fourfold degenerate sites and third codon positions, and to a lesser extent in the first position and the *Xdh* intron, is lower in these two species groups than in the *D. melanogaster* and *D. obscura* groups. The Hawaiian species (available only for *Adh*) show an intermediate GC content. Differences are greater for *Xdh* than for *Adh*, suggesting that the former region evolves faster than the later. We examine this possibility by means of a likelihood-ratio test of the null model that both gene regions evolve at equal rates. Thus we have combined the *Adh* and *Xdh* data in a single data set for both nucleotides and amino acid sequences. For nucleotides we have considered first and second, and third, codon positions separately, with the substitution model of Hasegawa et al. (1985); for amino acids we used an empirical model based on the matrix of Jones et al. (1992), using amino acid frequencies as free parameters [referred to as the JTT-F model by Yang et al. (1998)]. Setting different rate parameters for *Adh* and *Xdh* improves significantly the likelihood over the equal-rates model at the amino acid level (1-df test), with the *Xdh* protein evolving 1.43 ± 0.22 times faster than *Adh* ($2\log\Lambda = 6.04$; $p = 0.013$) and the third codon positions of *Xdh* evolving 1.61 ± 0.13 faster ($2\log\Lambda = 35.08$; $p < 0.001$); for first and second positions the difference is marginally significant ($2\log\Lambda = 3.64$; $p = 0.056$), with *Xdh* evolving 1.28 ± 0.17 faster than *Adh* in these codon sites.

If a given locus experiences different mutation pressures in different lineages, then positive correlations should be observed between the GC compositions of the codons and the introns (assuming that the intron base composition reflects the mutational equilibrium of the genome), the stronger the less constrained is the coding region. With Felsenstein's (1993) contrast test, the GC content of intron 2 correlates significantly with the first ($r = 0.81$, $p = 0.002$) and the third ($r = 0.84$, $p = 0.001$) codon positions in *Xdh*. Moreover, the GC content at fourfold degenerate sites of *Xdh*, which can reasonably be assumed to be under very weak functional constraints and less affected by sampling error than the intron 2 GC content, also correlates significantly with the first ($r = 0.84$, $p = 0.001$), second ($r = 0.60$, $p = 0.04$), and third ($r = 0.96$, $p < 0.001$) codon positions of *Xdh* and the third codon positions of *Adh* ($r = 0.61$, $p = 0.049$).

Figure 3 represents the frequency of optimal codons (*Fop*) (Ikemura 1985) across species groups in *Adh* and *Xdh*, computed with the set of major codons defined by Akashi (1995) for *D. melanogaster*. The *D. saltans* and *D. willistoni* groups deviate the most from the optimal codon usage among these species, while the Hawaiian representatives exhibit intermediate *Fop* values. The largest *Fop* values correspond to the *Adh* region in the *melanogaster* and *obscura* species, and the lowest to the *Xdh* region in the species of the *D. saltans* and *D. willistoni* groups.

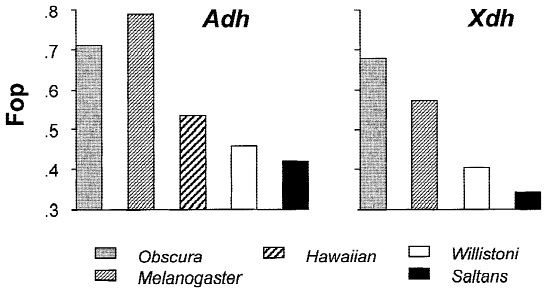

Fig. 3. Frequency of optimal codons [*Fop* values (Ikemura 1985), obtained with the set of major codons defined by Akashi (1995)] for the species groups of this study.

Figure 4 plots the proportions of high-GC (group I), intermediate-GC (group II), and low-GC (group III) amino acids against the GC content at fourfold degenerate sites in the *Xdh* region for the various species. As expected, group I amino acids are less used by species with a low GC content (the *saltans* and *willistoni* groups), while the opposite is the case for group III amino acids and, less so, for group II amino acids. The correlation between the group I contrast, *f*(I), and the GC_4 contrast is highly significant ($r = 0.80$, $p = 0.002$), the correlation of the GC_4 and *f*(III) contrasts is $r = -0.63$ ($p = 0.02$), and the correlation between the GC_4 and the *f*(II) contrasts is not significant ($r = -0.30$, $p = 0.34$). For the more conserved *Adh* region, an association between amino acid composition and GC content cannot be detected.

Previous analyses with the relative rate test on a reduced data set have shown that the above differences in amino acid composition between the *saltans* and the *melanogaster* and *obscura* species arise as the result of an increased rate of amino acid replacement in the *saltans* lineage (Rodríguez-Trelles et al. 1999b). This is apparent in Fig. 1, where we see that, since their split from a common ancestor, the number of amino acid substitutions that have occurred in the *saltans* and the closely related *willistoni* lineages is conspicuously larger than in the *obscura* and *melanogaster* lineages; a likelihood-ratio test against a model without the restriction of rate constancy among lineages clearly leads to the rejection of the molecular clock hypothesis ($2\log\Lambda = 27.10$, $p \sim 0.004$, 11 df).

The GC% compositional statistics for the ancestor of the species in our study [obtained by the maximum-likelihood RateAncestor method of PAML 1.4 (Yang 1997)] are 52.6, 39.2, 71.8, and 72.9 and 61.3, 41.5, 73.1, and 76.9, respectively, for positions 1, 2, 3, and GC_4 of

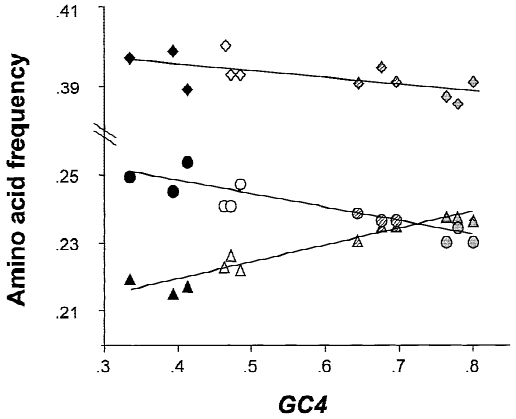

Fig. 4. Regression of the frequencies of amino acid groups I (*triangles*), II (*diamonds*), and III (*circles*) (high, medium, and low GC content, respectively) on the frequency of fourfold degenerate codons, GC_4, for *Xdh* (species groups represented as in Fig. 2).

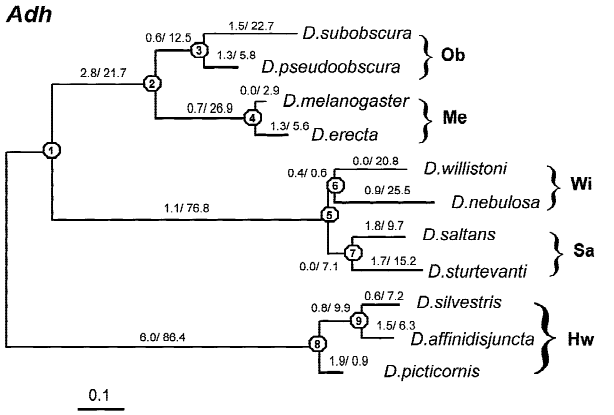

Fig. 5. The maximum-likelihood tree of a subset of the *Adh* nucleotide sequences used in this study; the tree is obtained using a simplified version of the codon-based model of Goldman and Yang [(1994; Yang 1998); program PAML 1.4 (Yang 1997)]. *Numbers above branches* represent the estimated number of nonsynonymous/synonymous changes per 100 sites of each class. Enclosed node numbers are used for reference in the text and Table 1.

Adh and *Xdh*, assuming the phylogenies in Figs. 5 and 6. These values are much closer to those of the *melanogaster* and *obscura* groups than to those of the *willistoni* and *saltans* species (Fig. 2). This observation supports the inference that GC content has evolved faster in the *willistoni* and *saltans* group species than in the *Sophophora.* Furthermore, the fact that the *willistoni* and *saltans* species exhibit a quite similar base composition, which is very different from that of their ancestor, indicates that most of the compositional change in the lineage took place before the differentiation of these two species groups, along the branch marked by the arrow in Fig. 1. As mutation and selection have different effects on synonymous and nonsynonymous substitutions, comparison of synonymous and nonsynonymous rates in this branch with those in other branches in the tree may help to elucidate the mechanisms responsible for the switch in base composition that occurred in the *willistoni* and *saltans* lineages. Hence, this is the branch of interest in later analyses.

The Ratio of Nonsynonymous-to-Synonymous Substitution Rates (d_N/d_S) Among Lineages. For practical reasons, because maximum-likelihood methods are computationally highly demanding, the following analyses are conducted on a subset of the sequences identified in Fig. 1 (i.e., we have eliminated one species of each group, generally one of the two most closely species in the group). This is not a serious handicap if we take into account that we are interested in lineage effects, rather than in effects associated with particular species. In addition, we have considered *Adh* data from the three picture-winged Hawaiian representatives. Previous analyses of the *Adh* gene in these species have shown an excess of nonsynonymous over synonymous substitutions, which was interpreted to be a consequence of reduced population sizes known for this Hawaiian lineage (DeSalle and Templeton 1988; Ohta 1993). These species provide an invaluable ground to evaluate the d_N/d_S ratios of other lineages.

Figures 5 and 6 show, respectively, the *Adh* and *Xdh* phylogenies of the major species groups of *Sophophora* (and the Hawaiian species for *Adh*) inferred under the codon-based model, allowing independent d_N/d_S ratios for branches (the "free-ratio" model). Above the branches are shown the numbers of nonsynonymous-to-synonymous substitutions, each per 100 sites of its class. Reference node numbers are shown in circles. According to the codon-based model, the *Adh* region has 101.15 synonymous and 303.85 nonsynonymous sites (with a total of 405 sites or 135 codons); for *Xdh* these numbers become, respectively, 534.78 and 1550.22 (with 2085 sites or 695 codons). We can see that synonymous substitutions occur more often than nonsynonymous substitutions, indicating that *Adh* and *Xdh* have spent a majority of time under purifying selection during the evolution of *Sophophora*. Constraining the ratio to be the same for all branches (the "one-ratio" model) we obtain d_N/d_S

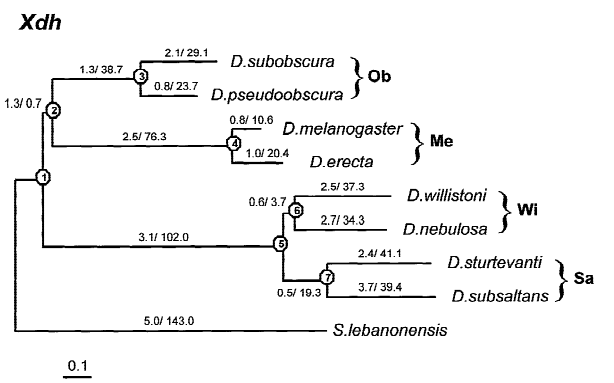

Fig. 6. The maximum-likelihood tree of a subset of *Xdh* nucleotide sequences used in this study; methods and labels are as in Fig. 5.

Table 1. Estimates of d_N/d_S rate ratio for *Adh* and *Xdh* in several *Drosophila* lineages[a]

Gene region	Lineage	$d_N/d_S \pm$ SE
Adh	Node 1 · Me & Ob	0.076 ± 0.019
	Node 1 · node 5	0.018 ± 0.012
	Node 5 · Wi & Sa	0.062 ± 0.018
	Node 1 · node 8	0.070 ± 0.029
	Node 8 · Hawaiian	0.197 ± 0.069
Xdh	Node 1 · Me & Ob	0.046 ± 0.005
	Node 1 · node 5	0.036 ± 0.009
	Node 5 · Wi & Sa	0.070 ± 0.006
	Node 1 · S. lebanonensis	0.039 ± 0.008

[a] d_N/d_S estimates are obtained with a five (for *Adh*)- and a four (for *Xdh*)-ratio codon-based model. These two models render a significant improvement of the log-likelihood over the corresponding one-ratio models ($2\log\Lambda = 12.64$, $p \sim 0.013$, 4 df, and $2\log\Lambda = 16.64$, $p < 0.001$, 3 df, respectively). Estimates for *Adh* are based on 135 codons and the tree shown in Fig. 5. For *Xdh* we use 695 codons and the tree shown in Fig. 6. Node numbers are those shown in Figs. 5 and 6.

$= 0.072 \pm 0.010$ for *Adh*, and $d_N/d_S = 0.053 \pm 0.003$ for *Xdh*, which can be interpreted as average d_N/d_S ratios over all the branches. The difference in log likelihoods between the free-ratio and the one-ratio models is significant for both, *Adh* ($2\log\Lambda = 36.18$, $p \sim 0.006$, 18 df), and *Xdh* ($2\log\Lambda = 31.42$, $p \sim 0.004$, 14 df), which means that the differences in d_N/d_S ratios among lineages are, anyhow, too large to be accounted for by a strictly neutral model.

For the purpose of this study, it is possible to extract the main trends contained in Figs. 5 and 6 using simplified models, intermediate between the simplest (one-ratio) and the most general (free-ratio) models discussed above. For instance, for *Adh* we chose a five-d_N/d_S ratio model, which includes one ratio for the species of the *willistoni* and *saltans* groups (i.e., from node 5 to the tips; see Fig. 5) and a second one for the branch ancestral to them (between node 1 and node 5), a third ratio for the *melanogaster–obscura* lineage (from node 1 to the tips), and a fourth and fifth ratio for the Hawaiian species (node 8 to the tips) and the branch ancestral to them (nodes 1 to 8). Similarly, for *Xdh* we use a four-d_N/d_S ratio model. The results are presented in Table 1. First, we can see that the highest d_N/d_S ratio in the table is, by and large, that exhibited by the Hawaiian species in the *Adh* region (0.197 ± 0.069), which agrees with Ohta's (1993) results and is consistent with a hypothesis of relaxed constraints driven by reduced population numbers. Second, the branch ancestral to the *willistoni–saltans* lineage shows the lowest d_N/d_S ratio in both *Adh* (0.018 ± 0.012) and *Xdh* (0.036 ± 0.009), which is, on average, about eight times lower than the ratio shown by the Hawaiian species. Despite the low d_N/d_S ratio, this branch is the largest of *Sophophora*, i.e., has accumulated many changes, but these appear to have been overwhelmingly synonymous.

We can compare statistically these d_N/d_S ratios with the most common situation for the lineages of this study, if we assume that the latter situation is reflected by the background d_N/d_S ratio, i.e., the ratio taken over all branches but the one of interest. This is a two-ratio model. Accordingly, the average d_N/d_S ratio of the Hawaiian species is significantly higher than the background ratio for *Adh* ($2\log\Lambda = 12.50$, $p \sim 0.004$, 1 df). In contrast, the d_N/d_S ratio of the branch ancestral to the *willistoni–saltans* lineage is significantly lower than the background ratio for *Adh* ($2\log\Lambda = 4.58$, $p \sim 0.032$, 1 df), and the difference is marginally significant for *Xdh* ($2\log\Lambda = 3.22$, $p \sim 0.070$, 1 df). Given that the most common scenarios of positive selection and relaxed constraints predict an increase in the d_N/d_S ratio, rather than a decrease, it seems that none of these two processes has been a factor for the evolution of base composition in the *willistoni* and *saltans* lineages.

The Average Rate of Evolution of Xdh *and Its Degree of Among-Site Rate Variation.* We have shown that the *Xdh* region evolves faster in the *willistoni–saltans* lineage than in the lineage consisting of the *melanogaster* and *obscura* species (also see Rodríguez-Trelles et al. 1999b). Now we are interested in ascertaining whether this difference in the average evolutionary rate among lineages is reflected in the pattern of among-site rate variation within the gene. This approach differs from previous attempts at exploring this question, which have focused on several genes across a single lineage. Indeed, previous studies have found that within a lineage fast-evolving proteins exhibit a lower degree of among-site rate variation than slowly evolving proteins (Kumar 1996; Zang and Gu 1998).

We have focused on both the degree of among-site rate variation (the subject of previous analyses) and the rate correlation over adjacent sites. The variation and dependence of substitution rates over sites are described with the autodiscrete gamma model of Yang (1995), which is based on a serially correlated gamma distribution (setting eight equal-probability categories of rates). The gamma distribution involves a shape parameter, α (>0), which is inversely related to the extent of rate variation among sites. For nucleotide data the substitution pattern is accommodated by the general reversible model (Yang 1994), which for *Xdh* produces a better fit than simpler models (Rodríguez-Trelles et al. 1999b). For amino acids we use the JTT-F empirical model (Jones et al. 1992; Yang et al. 1988). In a coding sequence, rates at sites three nucleotides apart are highly correlated, so that the correlation between rates for two adjacent sites is weakened (Yang 1995). Therefore, in order to estimate the correlation among neighboring sites from nucleotides (ρ), we have first removed the large-scale variation allowing different rate parameters for codon positions in the model. Model parameters are estimated by maximum likelihood, and their relevance for describing the data is evaluated by means of likelihood-ratio tests.

Table 2. *Xdh* among-site rate variation (α) and correlation (ρ) in different lineages[a]

Model	Parameter	Lineage		
		Wi & Sa	Me & Ob	All species
Among nucleotide sites	α	0.425 ± 0.039	0.312 ± 0.030	0.401 ± 0.023
	ρ	0.439 ± 0.113	0.669 ± 0.113	0.460 ± 0.070
Among amino acid sites	α	0.426 ± 0.101	0.186 ± 0.048	0.433 ± 0.056
	ρ	0.450 ± 0.113	0.551 ± 0.142	0.485 ± 0.080

[a] All parameters contribute significantly to describe the variation of the sequences, as inferred from the comparison of the models with the corresponding nested submodels by means of likelihood-ratio tests (1-df tests; results not shown). Estimates are obtained with the general reversible model (Yang 1994) for nucleotides and the empirical model of Jones et al. (1992) (as modified by Yang et al. 1998) for amino acids. Among-site rate variation is assumed to follow an autodiscrete gamma process with eight categories of rates. Different rate parameters for codon positions are allowed in nucleotide models for ρ.

Estimates of the among-site rate variation (α) and the rate correlation among adjacent sites (ρ) for lineages are shown in Table 2. As expected from the differences in the average rate of evolution of *Xdh* among lineages, α values are largest, meaning that the degree of among-site rate variation is lowest, in the fast-evolving *willistoni–saltans* lineage at both the nucleotide (0.425 ± 0.039 vs 0.312 ± 0.03) and the amino acid (0.426 ± 0.101 vs 0.186 ± 0.048) levels; on the other hand, ρ shows the opposite pattern: it is largest in the comparatively slowly evolving *melanogaster–obscura* lineage (0.669 ± 0.113 vs 0.439 ± 0.113 and 0.551 ± 0.142 vs 0.450 ± 0.113, at the nucleotide and amino acid levels, respectively). From normal-deviate tests [e.g., $z = (\alpha_1 - \alpha_0)/\mathrm{SQRT}(\mathrm{SE}_1^2 + \mathrm{SE}_0^2)$, using standard errors (SEs) computed by the curvature method in the likelihood analysis (see Yang 1997)], the difference in α values between the two lineages is significant ($p \sim 0.02$ and $p \sim 0.03$, for nucleotides and amino acids, respectively); for ρ the standard errors of the estimates are larger, and the differences are not significant ($p \sim 0.11$ and $p \sim 0.58$, for nucleotides and amino acids, respectively). While the tree topology is the same for the two species sets (i.e., *willistoni–saltans* and *melanogaster–obscura*), it can be argued that our results might be affected by tree length differences among lineages, attributable to chronological differences. In fact, α tends to be underestimated for closely related sequences and overestimated for highly diverged sequences (Zhang and Gu 1998). This asymptotic bias, however, occurs when the number of sequences is small [four or less (Zhang and Gu, 1998)]. Moreover, the *willistoni* and *saltans* lineages are likely to be more closely related to each other than *melanogaster* is to *obscura* (Throckmorton 1975), which would generate a bias in the opposite direction to the observed α values. Finally, enlarging the time scale by combining the two data sets yields estimates that are intermediate between those for the two lineages separately (see Table 2).

Discussion

GC-Content Differences: Fluctuating Mutation Bias vs Natural Selection. Interspecific differences in nucleotide composition and codon usage bias among lineages within the *Sophophora* subgenus similar to those investigated herein for *Adh* and *Xdh* are also found in the *Sod, Per,* and *28SrRNA* gene regions (Rodríguez-Trelles et al. 1999b); they appear to be the case for other sequenced regions as well [e.g., the *Ddc* and *amd* regions (Andrei Tatarenkov personal communication)]. The occurrence of such compositional patterns across several unlinked regions, scattered throughout the genome (see Rodriguez-Trelles et al. 1999b), suggests that these patterns reflect genomewide differences between lineages. Extensive variation in GC content is not unique to *Drosophila;* large differences in nucleotide composition have been long known among the genomes of bacterial species (Lee et al. 1956) and between regions of the mammalian genome [the so-called isochores, discovered by Bernardi et al. (1985; Bernardi 1995)].

In both respects, nucleotide composition and codon usage bias, the species of the *willistoni* group, which had not been previously investigated, show a pattern similar (but less extreme) to the pattern exhibited by the *saltans* species, as might be expected because these two species groups are more closely related to each other than to the *melanogaster* and *obscura* groups (Throckmorton 1975). The changes in GC composition can be attributed to an AT increase in the lineage ancestral to the *willistoni* and *saltans* groups after its split from the lineage that gave rise to the *melanogaster* and *obscura* groups [≈55 Mya (Fitch and Ayala 1994; Tatarenkov et al. 1999)]. The shift in nucleotide composition is associated with an accelerated change in the amino acid composition of relatively low-constrained proteins such us *Xdh*.

At least three competing hypothesis may account for these observations: (i) positive selection triggered, for example, by a change in protein function or a shift in metabolic efficiency; (ii) a reduction in the effectiveness of selection caused by either a reduction in population numbers, a reduction in the rate of recombination, or a relaxation of constraints; and (iii) a shift in the pattern of point mutation. According to the first two hypotheses we should observe an increase in the nonsynonymous-to-synonymous substitution rate ratio in the ancestor of the *willistoni* and *saltans* groups; specifically, if positive selection had been responsible for the GC changes, we would expect a d_N/d_S ratio significantly greater than one, because natural selection would have favored the fixation of nonsynonymous mutations over the neutral prediction; on the other hand, if the GC changes had been

caused by a reduction in the effectiveness of selection, we should observe a d_N/d_S ratio less than or equal to one but still greater than the average ratio observed in other *Drosophila* lineages. On the contrary, the branch ancestral to *willistoni* and *saltans* exhibits the lowest d_N/d_S ratio in both the *Adh* and the *Xdh* trees (Figs. 5 and 6, Table 1). This is what one would expect if a shift in the pattern of point mutation had actually occurred some time after the split of that branch ancestral to the *willistoni*/*saltans* lineage, because mutation bias would affect more the least constrained parts of the genome (i.e., synonymous sites) than the functionally significant parts (i.e., nonsynonymous sites) (Sueoka 1988). That the observed AT increase is not due to reductions in population size is further corroborated by noting that the d_N/d_S ratio obtained for the Hawaiian *Drosophila* in the *Adh* region is ~11 times higher than for the *willistoni*/*saltans* lineage (Table 1). High d_N/d_S ratios have previously been noticed for Hawaiian species (Ohta 1993), which were attributed to diminished effectiveness of natural selection due to their known history of severe bottlenecks (DeSalle and Templeton 1988; Ohta 1993). However, it seems reasonable to infer that if reduced efficiency of natural selection had been the cause of the even larger change in GC content observed in the *willistoni*/*saltans* lineages, the d_N/d_S ratio for this lineage should not be far lower than for the Hawaiian species, as is actually the case.

The previous inferences regarding the d_N/d_S ratios can potentially be confounded by several factors. First, our likelihood analysis assumes that nucleotide frequencies have remained more or less the same during the course of evolution (i.e., the stationarity assumption). This assumption is made by virtually all methods and models currently used for phylogenetic analysis; consequently, taxa with similar composition are clustered together regardless of their true phylogenetic relationships. Stationarity does not hold in our case (see also Rodríguez-Trelles et al. 1999b). It should be noted, however, that maximum-likelihood methods appear to be particularly robust to violations of the stationarity assumption and that the purpose of our analysis is not phylogenetic; rather, we start from phylogenetic relationships that are known in advance. Second, our likelihood analysis assumes that the d_N/d_S ratio is constant over all codon sites in *Adh* and *Xdh*. This assumption is expected to lead to underestimates of d_N, as nonsynonymous (or amino acid) substitution rates are variable among sites, and hence to lead to underestimates of the d_N/d_S ratio. From the analysis of the among-site rate variation in *Xdh*, however, rates are more variable in the *melanogaster*/*obscura* lineage ($\alpha = 0.312$ and $\alpha = 0.186$, from the nucleotide and amino acid sequences, respectively; see Table 2) than in the *willistoni*/*saltans* lineage ($\alpha = 0.425$ and $\alpha = 0.426$; see Table 2). Consequently, the underestimation problem should be larger in the former lineage than in the latter, which means that the observed differences in the d_N/d_S ratios among lineages might be even larger yet, in the direction predicted by the mutation shift hypothesis. And third, synonymous substitution rates could be underestimated due to multiple hits, which would lead to overestimates of the d_N/d_S ratios. However, this overestimation should be more acute in the *willistoni*/*saltans* ancestral lineage, because it is this branch that exhibits the largest compositional change. Again, this factor would increase the differences in the d_N/d_S ratios among lineages, precisely in the direction expected from a shift in the pattern of point mutation.

Our results of high d_N/d_S ratios for the Hawaiian species (relative to the *melanogaster* and *obscura* groups) in *Adh* coincides qualitatively with former analyses of this same gene by Ohta (1993). Nevertheless, ours and Ohta's (1993) estimates of the d_N/d_S ratios differ considerably from each other. For example, after correcting for multiple substitutions, Ohta (1993) obtains an average $d_N/d_S = 0.630$ for the Hawaiian *Drosophila*, whereas the corresponding estimate from our analysis is only 0.197. Discrepancies of this sort can be accounted for, in part, by disparities in the data sets (e.g., Ohta considers the entire *Adh* gene region, and she does not use *D. affinidisjuncta*). More importantly, however, might be the differences in the analytical procedures utilized. Ohta (1993) uses the approximate method for pairwise sequence comparisons of Nei and Gojobori (1986), unlike the codon-based likelihood models used here. But Yang and Nielsen (1998a) have shown that this method does not adequately accommodate transition/transversion biases and codon frequency biases. As a result, the Nei and Gojobori's (1986) method tends to underestimate d_S and overestimate (less so) d_N, yielding d_N/d_S ratios that are larger than those obtained by maximum likelihood (Yang and Nielsen 1998).

We have discussed the role of natural selection in determining base biases in terms of protein function. Yet positive selection might occur in a different way: a high GC content might be selected for a whole genome or DNA region. Corresponding replacements of amino acids in polypeptide chains might, then, be the result of such a process. In this case, it would seem difficult to predict what to expect concerning the nonsynonymous-to-synonymous rate ratio. For a given site, the ratio will depend on whether the DNA and protein levels interact synergistically or whether they are antagonistic with respect to the direction of selection. Be that as it may, it is far from obvious why natural selection should favor a lower GC genome content in the *saltans* and *willistoni* groups than in other *Drosophila* groups. Thermostable amino acids are encoded by GC-rich codons, and a high GC content in third codon positions and in introns and untranslated flanking regions would increase the thermal stability of the primary mRNA transcripts. Accordingly it has been suggested that an increasing need for protecting DNA and RNA from heat degradation would account

for the high GC content in the thermophilic bacteria (Kagawa et al. 1984) and, also, in isochores of warm-blooded vertebrates (Bernardi et al. 1985; 1988). However, a recent study has failed to find a correlation between the optimum growth temperature and the GC content of thermophilic bacteria (Galtier and Lobry 1997); also, a globin pseudogene from a GC-rich isochore was found to evolve faster than its paralogue from an AT-rich isochore, which is unexpected if thermal adaptation at the DNA level is governing the substitution process in the former (Francino and Ochman 1999). These observations bring into question whether thermal stability of the DNA is the evolutionary factor responsible for the GC-constant variation in bacteria and among the isochores of the mammalian genome. This hypothesis cannot account well for the large GC-content variation among *Drosophila* lineages either. In *Drosophila,* solar heating of necrotic fruit may expose larvae to temperatures above 45°C even in temperate latitudes (Feder 1996). However, differences in GC content do not fit the biogeography of the species we have investigated: the highest GC content occurs in the *obscura*-group species, which evolved in the cold and temperate climates of the Palearctic and Neartic regions (Powell 1997), whereas the lowest GC content is found in the *saltans*-group species, which evolved in tropical and subtropical regions (Powell 1997).

Mutation Pressure and the Degree of Among-Site Rate Heterogeneity. The profile of conservative and variable regions of any given gene reflects a unique interplay between variable functional constraints and mutation rates that can also change over short regions. As a consequence of this balance, substitution rates are variable from site to site and also correlated because sites within the same region share a broad causal background (Yang 1995). Former analyses (Kumar 1996) have shown that the extent of among-site rate variation along a given gene covaries with the average rate of evolution of the gene itself, with slowly evolving genes having a higher degree of among-site rate variation than fast-evolving genes. This pattern occurs independently of any assumption about the underlying distribution of the substitution rate (most frequently used is the gamma distribution) and appears to be a general feature of molecular evolution (Zhang and Gu 1998). These studies have focused on different genes within the same lineage, but it is not known whether this pattern also holds for a given gene across different lineages. Here we have provided evidence that this can actually be the case, at least for the *Xdh* gene.

In light of the negative correlation between the average evolutionary rate of a gene and its degree of among-site rate variation pointed out by other studies (Kumar 1996; Zhang and Gu 1998), the finding of a significantly lower degree of among-site rate variation of the *willistoni*/*saltans* lineage than in the *melanogaster*/*obscura* lineage provides additional support for the hypothesis that the former lineage is evolving faster. Both phenomena, i.e., variation in the rate of amino acid substitution among lineages and changes in the extent of the rate variation among sites in *Xdh,* can be explained as a consequence of fluctuating mutation bias (Sueoka 1962, 1988).

Sueoka (1993) postulated that for a set of nucleotides for which the pattern of point mutation is at equilibrium with base composition and selective constraints, the mutation frequency per gene will be much lower than the frequency observed immediately after a shift in the pattern of point mutation takes place. To illustrate this principle, we can imagine a hypothetical sequence with 100% G content and for which the pattern of spontaneous mutation is such that G never mutates toward A, T, or C. In this extreme scenario we would not observe variation even if all sites were neutral, i.e., sites would be invariant not because of functional constraints, but simply because variation cannot occur. It is worth noting that, under the assumption that mutations occur directionally at random, the previous reasoning would lead to overestimate functional constraints. However, a change in mutation bias (e.g., from G to A in our example) will provoke a burst of mutations in such a way that many unconstrained sites that were "frozen" with regard to variation will "unfreeze" until the new composition equilibrium is reached. It follows that the rate of replacement of quasi-neutral amino acids should increase, which is what we have observed for the amino acid replacements occurring in *Xdh* during the evolution of the *willistoni*/*saltans* lineage. Moreover, taking into account that the extent of among-site rate variation varies inversely with the proportion of neutral sites in a gene [i.e., the larger the proportion of neutral sites, the lower the extent of among-site rate variation (Zhang and Gu 1998)], the α value for *Xdh* should be larger in *willistoni*/*saltans* than in *melanogaster*/*obscura,* which is precisely what we found. In support of this interpretation of the α values, the estimated proportion of neutral sites [defined as those sites for which $d_N/d_S = 1$, obtained with the codon-based likelihood model of Nielsen and Yang (1998)] is significantly larger in the *willistoni*/*saltans* lineage (0.223 ± 0.017) than in *melanogaster*/*obscura* (0.159 ± 0.014) ($p \sim 0.003$, from the normal-deviate test).

Acknowledgments. We are indebted to Carlos Machado and Emile Zuckerkandl for valuable suggestions. F.R.-T. has received support from Ministerio de Educación y Cultura (Spain) (Contrato de Reincorporación and Grant PB96-1136 to A. Fontdevila). This research was supported by Grant GM42397 from the National Institutes of Health to F.J.A.

References

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in Drosophila DNA. Genetics 139:1067–1076

Akashi H (1996) Molecular evolution between *Drosophila melanogaster* and *D. simulans:* Reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics 144: 1297–1307

Akashi H, Schaeffer SW (1997) Natural selection and the frequency distributions of 'silent' DNA polymorphism in *Drosophila*. Genetics 146:295–307

Bernardi G (1995) The human genome: organization and evolutionary history. Annu Rev Genet 29:445–476

Bernardi G, Oloffsson B, Filipski J, et al. (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bernardi G, Mouchiroud C, Gautier C, Bernardi G (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–18

DeSalle R, Templeton AR (1988) Founder effects and the rate of mitochondrial DNA evolution in Hawaiian *Drosophila*. Evolution 42: 1076–1084

Feder ME (1996) Ecological and evolutionary physiology of stress proteins and the stress response: The *Drosophila melanogaster* model. In: Johnston IA, Bennett AF (eds) Animals and temperature. Cambridge University Press, Cambridge, pp 79–102

Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125:1–15

Felsenstein J (1993) PHYLIP—Phylogeny inference package, v. 3.5c. University of Washington, Seattle

Fitch WM, Ayala FJ (1994) The superoxide dismutase molecular clock revisited. Proc Natl Acad Sci USA 91:6802–6807

Francino MP, Ochman H (1999) Isochores result from mutation not selection. Nature 400:30–31

Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structure and optimal growth temperature in prokaryotes. J Mol Evol 44:632–636

Goldman N, Yang Y (1994) A codon-based model of nucleotide substitutions for protein-coding DNA sequences. Mol Biol Evol 11: 725–736

Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174

Ikemura T (1985) Codon usage and t-RNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

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

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. CABIOS 8:275–282

Kagawa YH, Nojima N, Nukiwa N, et al. (1984) High guanine plus cytosine content in the third letter of codons of an extreme thermophile. J Biol Chem 259:2956–2960

Kliman M, Hey J (1994) The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. Genetics 137:1049–1056

Lee KY, Wahl R, Barbu E (1956) Contenu en bases puriques et pyrimidiques des acids desoxyribonucleiques des bacteries. Ann Inst Pasteur 91:212–224

Li W-H (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitutions. J Mol Evol 36:96–99

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland, MA Moriyama EN, Hartl DL (1993) Codon usage bias and base composition of nuclear genes in *Drosophila*. Genetics 134:847–858

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

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Ohta T (1993) Amino acid substitution at the Adh locus of Drosophila is facilitated by small population size. Proc Natl Acad Sci USA 90:4548–4551

Patterson JT, Stone WS (1952) Evolution in the genus Drosophila. Macmillan, New York

Powell JR (1997) Progress and prospects in evolutionary biology: The Drosophila model. Oxford University Press, New York

Powell JR, DeSalle R (1995) *Drosophila* molecular phylogenies and their uses. Evol Biol 28:87–138

Petrov DA, Hartl DL (1999) Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. Proc Natl Acad Sci USA 96:1475–1479

Rodríguez-Trelles F, Tarrío R, Ayala FJ (1999a) Molecular evolution and phylogeny of the *Drosophila saltans* species group inferred from the *Xdh* gene. Mol Phylogenet Evol 13:110–121

Rodríguez-Trelles F, Tarrío R, Ayala FJ (1999b) Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. Genetics 153:339–350

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582–592

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657

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

Tarrío R, Rodríguez-Trelles F, Ayala FJ (1998) New *Drosophila* introns originate by duplication. Proc Natl Acad Sci USA 95:1658–1652

Tatarenkov A, Kwiatowski J, Skarecky D, Barrio E, Ayala JF (1999) On the evolution of *Dopa decarboxylase* (*Ddc*) and *Drosophila* systematics. J Mol Evol 48:445–462

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Throckmorton LH (1975) The phylogeny ecology and geography of *Drosophila*. In: King RC (ed) Handbook of genetics, Vol 3. Plenum Press, New York, pp 421–436

Yang Z (1994) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–111

Yang Z (1995) A space-time process model for the evolution of DNA sequences. Genetics 139:993–1005

Yang Z (1996) Maximum likelihood models for combined analyses of multiple sequence data. J Mol Evol 42:587–596

Yang Z (1997) PAML, a program package for phylogenetic analysis by maximum likelihood CABIOS 13:555–556

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568–573

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409–418

Yang Z, Nielsen R, Hasegawa M (1998) Models of amino acid substitution and applications to mitochondrial DNA evolution. Mol Biol Evol 15:1600–1611

Zhang J, Gu X (1998) Correlation between the substitution rate and rate variation among sites in protein evolution. Genetics 149:1615–1625