ELSEVIER

Gene 238 (1999) 39–51

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Within- and between-species DNA sequence variation and the 'footprint' of natural selection

Hiroshi Akashi *

Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS 66045-2201, USA

Received 11 March 1999; received in revised form 17 June 1999; accepted 6 July 1999; Received by G. Bernardi

Abstract

Extensive DNA data emerging from genome-sequencing projects have revitalized interest in the mechanisms of molecular evolution. Although the contribution of natural selection at the molecular level has been debated for over 30 years, the relevant data and appropriate statistical methods to address this issue have only begun to emerge. This paper will first present the predominant models of neutral, nearly neutral, and adaptive molecular evolution. Then, a method to identify the role of natural selection in molecular evolution by comparing within- and between-species DNA sequence variation will be presented. Computer simulations show that such methods are powerful for detecting even very weak selection. Examination of DNA variation data within and between Drosophila species suggests that 'silent' sites evolve under a balance between weak selection and genetic drift. Simulated data also show that sequence comparisons are a powerful method to detect adaptive protein evolution, even when selection is weak or affects a small fraction of nucleotide sites. In the Drosophila data examined, positive selection appears to be a predominant force in protein evolution.

Keywords: Adaptive evolution; Molecular evolution; Nearly neutral theory; Neutral theory; Population genetics

1. Introduction

Current theories of molecular evolution can be classified into three broad categories. The neutral theory (Kimura, 1968,1983; King and Jukes, 1969) proposes that mutations fall primarily into two fitness classes. A fraction of DNA changes are strongly deleterious and are quickly eliminated from populations by natural selection. The vast majority of non-deleterious mutations have little or no effect on an organism's fitness (they are redundant with respect to physiology and function) and their evolutionary dynamics are governed solely by genetic drift. The 'nearly neutral' model (Ohta, 1973, 1992) is a derivative of the neutral theory and posits that a large fraction of mutations have selection coefficients near the reciprocal of the species effective population size. For such mutations, evolution proceeds under a balance among the forces of mutation pressure, natural selection, and genetic drift. Finally, adaptive theories of molecular evolution allow for a large fraction of mutations with deleterious effects, but, of the remaining mutations, a relatively large proportion confer a fitness advantage to the organism. Under this model, positive selection plays an important role in DNA evolution.

The models discussed above differ in their predicted distributions of the fitness effects of mutations. The critical parameter in the models is the product of the effective population size and the selection coefficients of mutations, $N_e s$. Fig. 1**a** shows an example of a probability density of fitness effects of newly arising mutations under the neutral model. Almost all mutations are either strongly deleterious or have no fitness consequence, but there is a very small density of adaptive mutations. In this example, two-thirds of mutations are deleterious and, of the remaining mutations, the ratio of neutral to adaptive changes is 999:1. Given the probability of fixation for mutations of a given fitness effect (Kimura, 1962), this distribution can be transformed into the density of the fitness effects of DNA substitutions that accumulate between species. Fig. 1**b** shows such a density under an assumption of constant N_e. Under the neutral model, selection effectively prevents deleterious mutations from going to fixation. Neutral mutations, however, may go to fixation by

* Tel.: +1-785-864-3727; fax: +1-785-864-5321.
E-mail address: hiroshi@falcon.cc.ukans.edu (H. Akashi)

0378-1119/99/$ – see front matter
PII: S0378-1119(99)00294-2

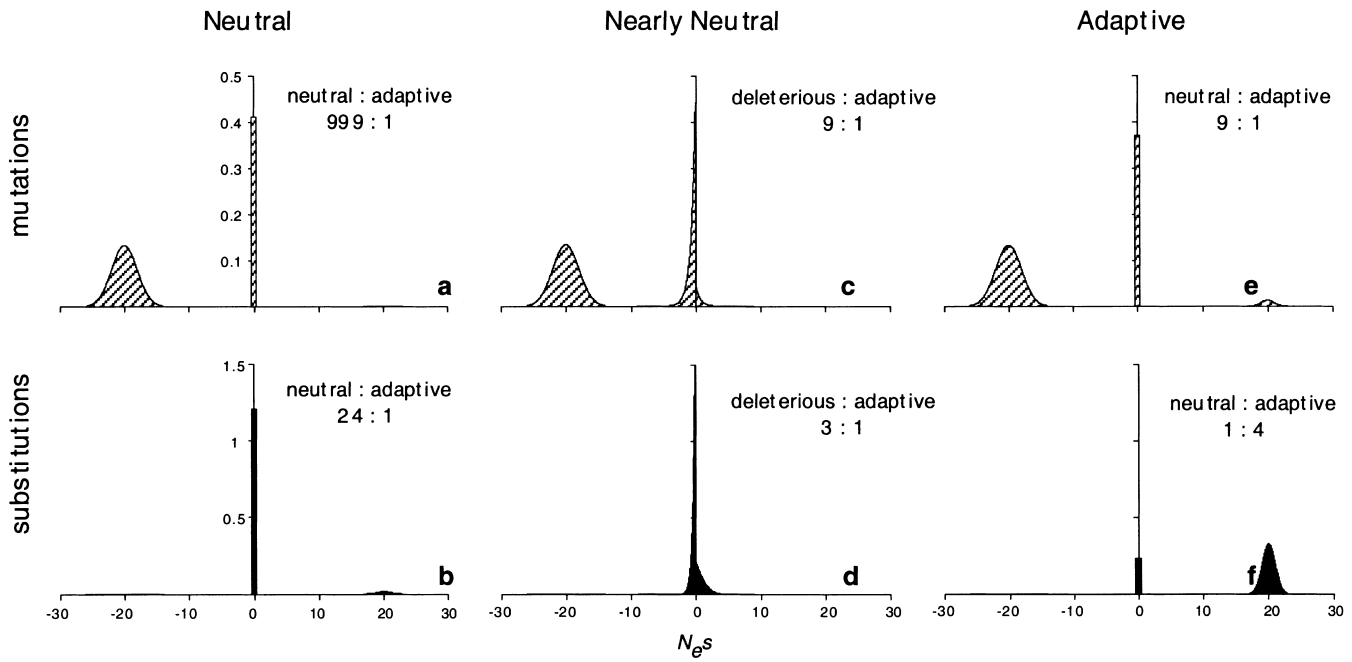

Fig. 1. Distributions of fitness effects under neutral, nearly neutral, and adaptive models of molecular evolution. One possible density of fitness effects of mutations is shown for each model (the area under the curves sum to one in each graph). The bar at $N_e s = 0$, neutral mutations, is broadened for visualization. The graphs in **a**, **c**, and **e** are examples of distributions of fitness effects of *mutations* that arise within a population under the neutral, nearly neutral, and adaptive models of molecular evolution respectively. The graphs **b**, **d**, and **f**, show the distribution of the fitness effects of *fixed differences* given the density in graphs **a**, **b**, and **c** respectively. The sampling formulae of Sawyer and Hartl (1992) were used to transform the density of fitness effects of mutations to that of fixations. These results assume that the density of fitness effects remains constant over evolutionary time and that evolution is independent at all sites (free recombination and no epistasis).

genetic drift, and the small proportion of positively selected substitutions explains phenotypes that are clearly adaptive (such as wing morphology or the compound eye). Note that the probability of fixation is strongly dependent on $N_e s$; the ratio of neutral to adaptive *fixations* (Fig. 1**b**) is considerably lower than the ratio of neutral to adaptive *mutations* (Fig. 1**a**).

Fig. 1**c** shows an example of the fitness effects of mutations under Ohta's nearly neutral model (Ohta, 1973, 1992). This model allows a fraction of mutations to fall into the strongly deleterious class, but, in addition, a substantial proportion of mutations confers fitness effects in the neighborhood of neutrality. Such mutations evolve under roughly equal magnitudes of selection and drift, so that nucleotide differences between species reflect a combination of slightly deleterious and weakly advantageous substitutions. An important property of weak selection models is the fixation of mutations with deleterious fitness consequences. However, even for weakly selected mutations, the sieve of natural selection decreases the ratio of deleterious to advantageous fixations relative to the same ratio for mutations (Fig. 1**d**).

Finally, Fig. 1**e** shows an example of a distribution of selection coefficients under a predominantly adaptive mode of molecular evolution. This model also accepts that a large fraction of mutations is deleterious and does not contribute to molecular polymorphism or divergence. Adaptive models can also accommodate a considerable fraction of neutral mutations. The critical difference between adaptive and neutral/nearly neutral models is the relative frequency of beneficial mutations. In Fig. 1**c**, two-thirds of new mutations are strongly deleterious (identical to the neutral model in Fig. 1**a**), but, among the remaining mutations, the ratio of neutral to adaptive mutations is 9:1. Although the density of the fitness effects of new *mutations* is skewed toward neutral changes, the effect of the selective sieve is dramatic; among mutations that substitute between species in this example, approximately 80% are adaptive.

Two features of the distribution of fitness effects of mutations distinguish among these models: the density of mutations in the neighborhood of neutrality and the frequency of beneficial changes. Section 2 will discuss how patterns of DNA sequence variation within and between closely related species can reveal such features in the distribution of the fitness effects of mutations.

2. Evolutionary configurations and the fitness effects of mutations

Population genetics theory shows that even very weak selection can have a substantial effect on evolution within and between closely related species (Kimura,

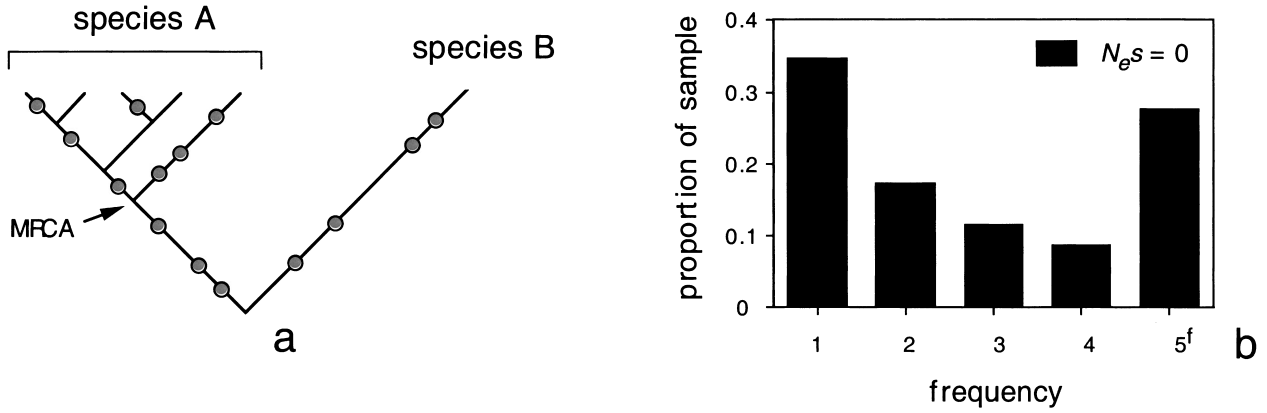

Fig. 2. Within- and between-species DNA sequence variation. (**a**) An example of a gene genealogy among five sequences from within species A and one sequence of species B. Dots represent mutations that have accumulated between the sequences and the arrow points to the most recent common ancestor in the within-species sample from species A. (**b**) A histogram showing the expected proportion of segregating and fixed neutral mutations in frequency classes $1<r<m$ for a sample of $m=5$ sequences. Frequency classes $1<r<m$ are polymorphic in the within-species sample (mutations that have occurred since the MRCA). Frequency class $r=m$ represents fixations in the lineage (mutations that have accumulated prior to the MRCA) and is denoted by the superscript 'f'. The proportion of mutations in the $r=m$ class depends on the length of the lineage examined. Here, an estimate for the time since divergence in the *D. simulans* lineage since its split with *D. melanogaster*, $t_{div}=0.6$, is used [see Akashi and Schaeffer (1997)], so that expectations can be compared with data from *D. simulans*.

1983). Fig. 2**a** represents the genealogical relationships among five sequences of a given gene sampled from within species A and one sequence sampled from a closely related outgroup, species B. The dots on the tree represent mutations that have occurred some time in the history connecting these sequences. The arrow points to the most recent common ancestor (MRCA) of the sequences sampled from species A; mutations that have occurred after this point are 'polymorphic' in this species. Polymorphic mutations have a frequency in the within-species sample depending on where they occurred in the genealogy. More recent mutations tend to be found in only one or two of the sequences and mutations deeper in the tree (closer to the MRCA) tend to be found at higher frequencies. Mutations that occurred in the genealogy prior to the MRCA are 'fixed' within the species A sample (they are shared among the five within-species sequences).

Under the neutral model, mutations accumulate at a constant rate along each branch of the genealogy. The expected number of mutations on each branch of the genealogy (both within and between species) is proportional to the length of the branch and is independent of its location in the tree. Natural selection, however, can affect both the expected shape of the genealogy of the sequences and the likelihood of finding mutations at different depths in the tree. Positive selection will move mutations deeper in the genealogy, whereas negative selection will cause mutations to be found in the tips of the branches. Positive and negative selection will also increase and decrease respectively the expected total number of mutations on the genealogy.

The frequency distribution of observed mutations contains information regarding the location of mutations in a genealogy. The histogram in Fig. 2**b** shows the expected proportion of mutations in each frequency class under a stationary, neutral model. The x-axis represents the frequency r, the number of sequences in which a new mutation is found. r varies from unity to m, where m is the number of sequences, or alleles, sampled from a given species. Mutations at frequencies $1<r<m$ are polymorphic and the frequency class $r=m$ represents mutations that are fixed in the sample. I will refer to such a histogram as the 'configuration' of mutations. In Fig. 2**b**, roughly one-third of the mutations are found in only one sequence and fewer are found deeper in the within-species part of the tree. The proportion of mutations in the fixed class ($r=5$) depends on the length of time between the MRCA and the outgroup examined. The divergence time of 0.6 (scaled to effective population size) used in Fig. 2**b** is that estimated for the *Drosophila simulans* data that will be examined in the following sections (Akashi, 1995).

The histograms in Fig. 3 show the effect of natural selection on the expected configurations of mutations. Fig. 3**a** shows the proportion of mutations in each frequency class under negative selection. Deleterious mutations show a unilateral skew toward rare variants (lower values of r). Negative selection also dramatically reduces the expected *number* of mutations in the sample (Fig. 3**b**). Fig. 3**c** shows the equivalent histograms under adaptive evolution. Even very weak positive selection skews mutations to higher values of r, reflecting mutations accumulating deeper in the genealogy. Adaptive evolution also increases the expected number of mutations on the tree (Fig. 3**d**). These patterns illustrate the sensitivity of configuration patterns to both selection in

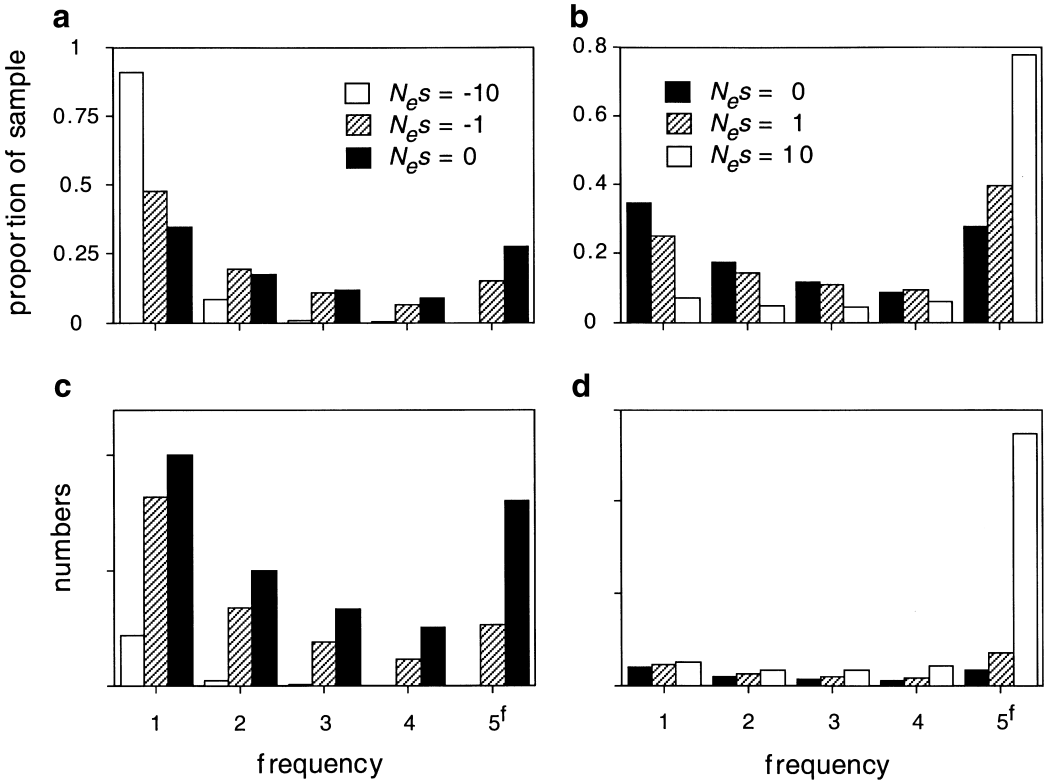

Fig. 3. Expected configurations under negative and positive selection. The expected numbers of newly arisen mutations at frequency classes $r=1$ to m in a sample of sequences were calculated according to Sawyer and Hartl (1992) and Hartl et al. (1994). Data are shown for $m=5$ sequences and $t_{div}=0.6$. The graphs **a** and **b** show the expected proportion of *variable sites* in the sample at different frequencies under negative and positive selection, respectively. The graphs **c** and **d** show the proportion of *mutable sites* at which variants are expected to be segregating at different frequencies or fixed in the sample under negative and positive selection respectively. Note that the scales for the *y*-axes for **c** and **d** depend on mutation rates and are unlabeled. For a given mutation rate, the scales will differ for **c** and **d**.

the neighborhood of neutrality and to adaptive evolution.

Fig. 2 shows the expected configuration of mutations under a model of free recombination among nucleotide sites and an equilibrium frequency distribution of mutations (Sawyer and Hartl, 1992; Hartl et al., 1994). The configuration of mutations for DNA sequences from a natural population, however, can be affected by population level phenomena, such as changes in population size or limited migration between local populations, as well as balancing and directional selection at genetically linked regions. Golding et al. (1986) and Sawyer et al. (1987) first suggested that *differences* in fitness effects between two categories of mutations (such as replacement and silent changes) could be identified by comparing their configurations. If the classes of mutations are randomly interspersed within a region of DNA, population history and selection at linked sites will have a roughly equivalent impact on the two classes. Only the direct action of natural selection on the mutations examined will cause differences in the frequency distributions of polymorphic mutations (Sawyer et al., 1987) or in the ratios of polymorphic and fixed mutations (McDonald and Kreitman, 1991; Templeton, 1996; Akashi, 1997a). A growing number of claims of adaptive (McDonald and Kreitman, 1991; Eanes et al., 1993; Long and Langley, 1993; Karotam et al., 1995; King, 1998), deleterious (Sawyer et al., 1987; Ballard and Kreitman, 1994; Nachman et al., 1994, 1996; Rand et al., 1994; Akashi, 1996; Templeton, 1996; Wise et al., 1998), and balancing (Wayne et al., 1996) selection on amino acid variants, mutation-selection-drift at silent sites (Ballard and Kreitman, 1994; Akashi, 1995, 1997a; Akashi and Schaeffer, 1997), and deleterious effects of transposable element insertions (Golding et al., 1986) rely on differences in the observed configurations of mutations. The following analyses will examine how configuration comparisons can be employed to distinguish among the models of evolution discussed above.

3. Tests of weak selection at silent sites in DNA

Configuration comparisons require within- and between-species DNA sequence data and two classes of mutations, preferably putative fitness classes. Under a model of weak selection, referred to as 'major codon preference', silent mutations fall naturally into weakly

deleterious changes from translationally preferred to unpreferred codons and slightly advantageous mutations in the opposite direction. This section will review evidence motivating this model, demonstrate the statistical power of configuration comparisons to detect weak selection, and apply such tests to DNA sequence data from Drosophila.

Several lines of evidence suggest that natural selection discriminates among synonymous codons to enhance the efficiency and/or accuracy of protein synthesis *Escherichia coli*, *Saccharomyces cerevisiae*, *D. melanogaster*, and a number of other organisms [for reviews see: Ikemura (1985), Andersson and Kurland (1990), Sharp et al. (1995), and Akashi and Eyre-Walker (1998)]. In these organisms, genome-wide patterns of synonymous codon usage are biased toward a subset of codons, called 'major codons', for each amino acid. Such codons tend to be recognized by abundant tRNAs and experimental evidence in *E. coli* has shown that major codons can enhance translational elongation rates [reviewed in Andersson and Kurland (1990)] and reduce misincorporations [reviewed in Parker (1989)]. Codon usage bias varies considerably among genes, and, in *E. coli* and yeast, the degree to which a gene is biased is a positive function of its expression level. Such patterns support the notion that mutations at silent sites affect an organism's fitness through their effect on protein synthesis (Sharp and Li, 1986; Li, 1987; Bulmer, 1988, 1991).

Li (1987) and Bulmer (1991) have modeled the dynamics of molecular evolution under major codon preference. The simplest case, of twofold redundant codons in a haploid organism, is depicted below. Mutations occur at rates v from non-major codons to major codons and u in the opposite direction. Major codons confer selective advantage, s.

$$\underset{+s}{\text{major}} \overset{u}{\underset{v}{\rightleftarrows}} \underset{-s}{\text{minor}}$$

At a 'locus', or protein-coding gene, consisting of a number of such sites, the expected proportion of major codons is determined by u/v, the ratio of the mutation rates, and $N_e s$, the product of effective population size and selection coefficient. Under relatively constant parameter values, the proportion of major codons at the locus will reach a steady-state at which the numbers of forward and backward substitutions will be equal.

Major codon preference can be tested using configuration comparison because the model predicts two fitness classes of silent mutations, 'preferred' mutations from non-major to major codons and 'unpreferred' mutations in the opposite direction (Akashi, 1995). If selection is sufficiently close to zero ($N_e s \ll 1$), then codon bias will be maintained by a combination of differences in the forward and backward mutation rates and genetic drift (Freese, 1962; Sueoka, 1962, 1988), and the evolutionary configurations of mutations in the two directions will not differ (both are neutral).

Fig. 4 shows the expected configurations of preferred and unpreferred mutations under weak selection. Under $N_e s = \pm 1$, roughly the selection intensity required to maintain codon bias at levels observed in the *D. simulans* genes of Table 1, the configurations of unpreferred and preferred mutations show unilateral shifts toward lower and higher frequencies respectively. This pattern can be regarded as the signal, or the 'footprint', of major codon preference in DNA sequence variation. However, the patterns shown in Fig. 3 reflect average configurations

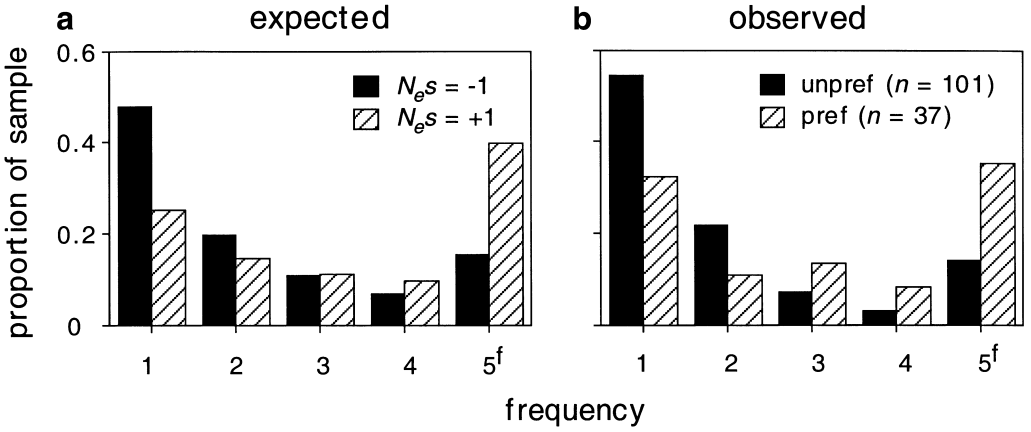

Fig. 4. Expected configurations under major codon preference and observed configurations of silent mutations in *D. simulans*. (**a**) The expected numbers of newly arisen mutations at frequency classes $r=1$ to m in a sample of sequences were calculated according to Sawyer and Hartl (1992) and Hartl et al. (1994). Data are shown for $m=5$ sequences, $t_{div}=0.6$, and $N_e s=1$ and -1 for preferred and unpreferred mutations respectively. (**b**) The proportions of 101 unpreferred (black) and 37 preferred (striped) mutations segregating at the given frequencies or fixed in the sample are shown. Pooled data from eight *D. simulans* genes from Table 1. Methods to identify major codons and infer ancestral and derived states for silent mutations are given in Akashi (1995). Equivalent data from *D. melanogaster* are not shown because other lines of evidence indicate a reduction in the efficacy of selection at silent sites in this lineage (Akashi, 1995, 1996).

Table 1
Evolutionary configurations of DNA mutations in *D. simulans*[a]

r	n_r		
	unpref	pref	rep
1	55	12	9
2	22	4	1
3	7	5	0
4	3	3	0
5	14	13	12

[a] The number of non-ancestral mutations n_r segregating at frequency r in samples of five *D. simulans* sequences are shown for unpreferred (unpref) and preferred (pref) silent changes, and for replacement (rep) mutations. Data were pooled across eight genes: *Adh*, *Adhr*, *boss*, *Mlc1*, *Rh3*, *per*, *Pgi*, and *Zw*. See Akashi (1997a,b) for GenBank accession numbers or references for these data.

that will be approached over a large number of independent trials of evolution. In order to apply and interpret configuration tests to silent DNA changes, it is important to estimate the statistical power of the approach. If the model is correct, what is the probability of observing differences in the configurations of mutations in a sample of sequences drawn from a single realization of a stochastic process?

Patterns of DNA sequence variation under major codon preference can be generated by computer simulation under assumptions of free recombination and a stationary frequency distribution of mutations. Such a scenario is achieved when population sizes and mutation rates have remained relatively constant, and when recombination rates have been high. Under these conditions, each nucleotide site has an independent genealogical history. Li (1987) and Bulmer (1991) give expressions for the steady-state proportion of major codons in a given gene and Sawyer and Hartl (1992) and Hartl et al. (1994) provide formulae for the expected numbers of each category of mutations (in this case, preferred and unpreferred changes) in each frequency class. Under the assumptions of the model, the numbers of mutations in each frequency class are independent Poisson random variables. Their expected values are a function of $N_e s$, m, l, the number of samples nucleotide sites, t_{div}, the time of divergence on the lineage examined (scaled to effective population size), and u and v. Given this information, simulated sequence data can be generated under major codon preference and the probability of rejecting a null hypothesis of equivalent configurations can be determined.

Fig. 5 shows the statistical power to detect major codon preference under the scenario described above. Parameters for effective population size and mutation rates were set to estimates in Drosophila and the selection coefficients, and numbers of alleles and sites were varied over a range of interest (see figure legend for parameter values). For each set of parameter values, 1000 independent data sets (configurations for preferred and unpreferred mutations) were generated and four statistical tests were employed to detect differences in the configurations. These tests examine different parts of the configuration of mutations and are sensitive to different departures from homogeneity of configurations.

Sawyer et al. (1987), in the first configuration test, compared the frequency distributions of silent and replacement DNA polymorphisms. A departure from a null hypothesis of homogeneity between 'singletons' and intermediate frequency polymorphisms was interpreted as evidence for differences in the fitness effects of the mutations. The approach does not require an outgroup sequence (none was available for their analyses), but restricts the analyses to polymorphism data with unknown ancestral and derived states. In the simulations described here, only results from a Mann–Whitney U test comparing the frequency distributions of polymorphic mutations are shown (fdMWU tests). This test was generally found to be at least as powerful as the 2×2 contingency table comparisons suggested by Sawyer et al. (1987).

McDonald and Kreitman (1991) expanded configuration comparisons to include between-species variation in a test of homogeneity between silent and replacement mutations. Their 2×2 contingency table compares the numbers of polymorphic mutations, pooled across frequency classes, and the numbers of fixed differences. This approach is likely to add statistical power to the analyses because directional selection has a strong impact on the fixed differences class (Fig. 3**a** and **c**), but this gain in power may be mitigated by pooling all polymorphic mutations into a single category, thus sacrificing information from the frequency distribution of mutations that have accumulated since the MRCA. Here, Monte Carlo analogs of Fisher's exact test were used to compare ratios of polymorphism to divergence (pdF tests). See Akashi (1999) for details of the analyses.

Templeton (1996) combined the approaches described above by testing homogeneity across three frequency classes. The numbers of singleton polymorphisms, polymorphisms at intermediate frequencies, and fixed differences were compared between silent and replacement mutations. Although the statistical test examines information from both the frequency distribution of segregating mutations and the numbers of fixed differences, some information is lost by pooling all polymorphic mutations segregating at frequencies greater than one. Monte Carlo analogs of Fisher's exact test were used to test homogeneity in 2×3 contingency tables (sidF tests).

Finally, configuration comparisons can be extended so that each frequency class ($1 < r < m$) is treated as a distinct category (Akashi, 1997a). In the following analyses, Mann–Whitney U tests were employed to test

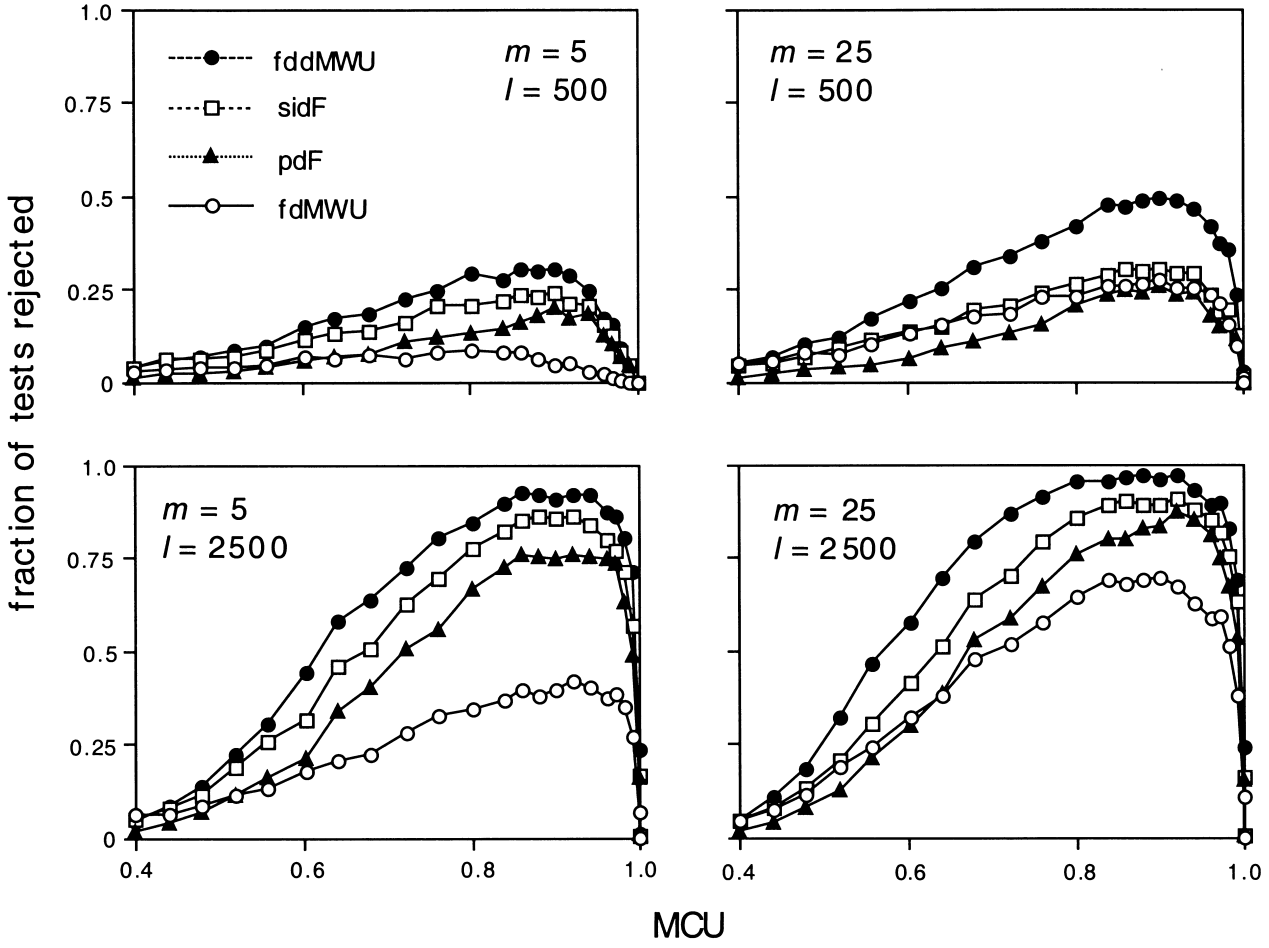

Fig. 5. Statistical power to detect major codon preference through configuration comparisons. The *y*-axis plots the proportion of tests that reject fitness equivalence, $P<0.05$, among 1000 simulated data sets for each value of MCU, the proportion of major codons at silent sites. Because the direction of the deviations in the configurations of preferred and unpreferred mutations are predicted under major codon preference, one-tailed probabilities were calculated for these tests. The four tests are described in the text. The four graphs show results for sample sizes, $m=5$ and 25 sequences from a population and for sequence lengths, $l=500$ and 2500 mutable sites. Each data point is based on simulations of mutation-selection-drift with parameters $N_e=5\times10^6$, $t_{div}=0.6$, u (per site) $=2.4\times10^{-9}$, and v (per site) $=1.6\times10^{-9}$. $u/v=1.5$ gives an equilibrium mutational base composition of 60%AT, the average base composition of *D. melanogaster* introns (Shields et al., 1988; Moriyama and Hartl, 1993).

for differences in patterns of frequency distributions and divergence between classes of DNA variation (fddMWU tests).

Fig. 5 compares the statistical power of these four methods to detect mutation-selection-drift. For all the tests, the power to detect selection increases initially with $N_e s$ but falls off as selection intensity increases. This is because as major codon usage reaches 100%, the per locus unpreferred mutation rate (from non-major to major codons) decreases to zero. Each of the statistical methods shows some power to detect weak selection. The frequency distribution, fdMWU, test is generally less powerful than tests that include divergence data. Among the latter category, 2×3 sidF tests of independence are considerably more powerful than 2×2 pdF tests. Overall, however, the fddMWU test is either indistinguishable from, or more powerful than, all the other tests over the parameter ranges considered. The gain in power is greatest when the number of sampled alleles is large (Fig. 5, $m=25$). For the same total number of aligned nucleotides, increasing the number of sampled sites has a greater impact on statistical power than increasing the numbers of alleles.

These power analyses suggest that configuration comparisons, given enough mutations, can detect natural selection near its limit of efficacy. Fig. 4**b** shows the configurations of preferred and unpreferred mutations pooled from five alleles from each of eight *D. simulans* genes found in the literature or in GenBank (Table 1). Although only five alleles were analyzed in *D. simulans*, close to 2500 silent sites were examined across the eight genes. If codon bias is maintained under mutation-selection-drift in this lineage, then comparing frequency distribution and divergence data should have a high

probability of rejecting fitness equivalence between preferred and unpreferred mutations (Fig. 5, $m=5$, $l=2500$).

In *D. simulans*, the configurations of preferred and unpreferred mutations are very similar to those expected under weak selection (Fig. 4**b**). The 37 preferred mutations are segregating at higher frequencies and are more often fixed in the lineage than the 101 unpreferred changes (Mann–Whitney U test, $z=3.12$, $P=0.0009$, one-tailed). Although the power tests described above require some stringent assumptions and a number of parameter estimates, comparisons of configurations between classes of mutations appear to be remarkably free of these assumptions (Sawyer et al., 1987; Hudson, 1993). Genetic linkage among mutations and/or changes in effective population size may alter the power to detect selection (results from simulations of evolution under these conditions will be published elsewhere) but do not appear to account for differences in the configurations of mutations [although see Eyre-Walker (1997) and Akashi (1997b) for discussion of mutation rate changes and configuration comparisons]. It is difficult to explain these patterns in the absence of major codon preference (Akashi, 1997a).

4. Detecting deleterious and adaptive protein evolution

The above analyses demonstrate the power of configuration comparisons to detect weak evolutionary forces at silent sites in Drosophila. Unfortunately, the fitness effects of particular amino acid mutations are generally more difficult to predict [for notable counter-examples see Hughes and Nei (1988), Shaw et al. (1993) and Yokoyama (1997)]. Deleterious amino acid changes will show configurations skewed toward low frequency variants, whereas adaptive evolution results in an excess of high frequency polymorphisms and fixed differences. However, in the absence of a method to categorize protein changes into putative fitness classes, we are likely to be examining pooled fitness classes of amino acid variation. The statistical power to detect a distribution of selection coefficients is examined below.

The original model by Ohta (1973) proposed that slightly deleterious mutations constitute a large fraction of both polymorphic and fixed amino acid variation. To determine the sensitivity of configuration comparisons to detect such a scenario, simulations were conducted under which a neutral class of variation is compared with a second class that includes strongly deleterious mutations, weakly deleterious mutations, and some fraction of neutrally evolving sites. Simulations were conducted for a total of $l=10\,000$ mutable sites, roughly corresponding to the size of the *D. simulans* data from Table 1. 2500 of the sites fell into the neutral class and for the remaining 7500 sites (representing replacement mutations) the fractions of strongly deleterious, slightly deleterious and neutral sites were varied over a range of interest. The strongly deleterious fraction does not contribute to polymorphism or divergence between species and essentially lowers the number of mutable sites. Either 1/3 or 9/10 of replacement changes were assumed to be strongly deleterious. At the remaining sites, both the proportion of neutral and deleterious changes and the fitness effects of the latter category were varied. Fig. 6 shows the statistical power to detect a mixture of three fitness classes of mutations (the x-axis plots the proportions of neutral and weakly deleterious mutations).

The power curves in Fig. 6 show that configuration comparisons can reveal the effects of a relatively small fraction of weakly deleterious replacement mutations. The power to detect $N_e s=-2$ increases steadily as the fraction of the mutations increases and the fddMWU test appears to be uniformly most powerful for detecting a combination of strongly deleterious, weakly deleterious, and neutral protein evolution. However, for more strongly deleterious mutations, the power to detect selection decreases as a function of the strength of selection (Akashi, 1999) because, although the configuration of mutations shows strong skews toward rare variants, the expected number of mutations on the genealogy declines rapidly to zero (Fig. 3**b**). For the scenario examined here, configuration tests can have substantial power to detect $N_e s=-10$, but the power declines for stronger deleterious mutations (data not shown).

Although Fig. 6 demonstrates that deleterious protein evolution can be identified, it also raises some issues with defining 'slightly deleterious' fitness effects. An important feature of Ohta's model (Ohta, 1973, 1992) is that deleterious mutations substitute at a substantial rate. Selection has a stronger effect on the probability of fixation of mutations than it does on the probability of segregating within populations (Kimura, 1983, p. 44). For a range of selection coefficients, the fixation probability is essentially zero, but the probability of segregating within populations remains substantial (i.e. for $N_e s=-10$, the probability of fixation is 1×10^{-5} relative to that of neutral mutations, but the probability of segregating within population samples is 0.15). It is important to note that a configuration of amino acid variants skewed toward rare variants suggests deleterious evolution (if the comparison class is indeed neutral), but such patterns do not necessarily imply that deleterious mutations have gone to fixation in the lineage examined. Given a distribution of selection coefficients, mutations in the fixed difference class could reflect neutral or even adaptive changes. An excess of rare variants suggests the existence of a pool of deleterious mutations with the *potential* to go to fixation in the

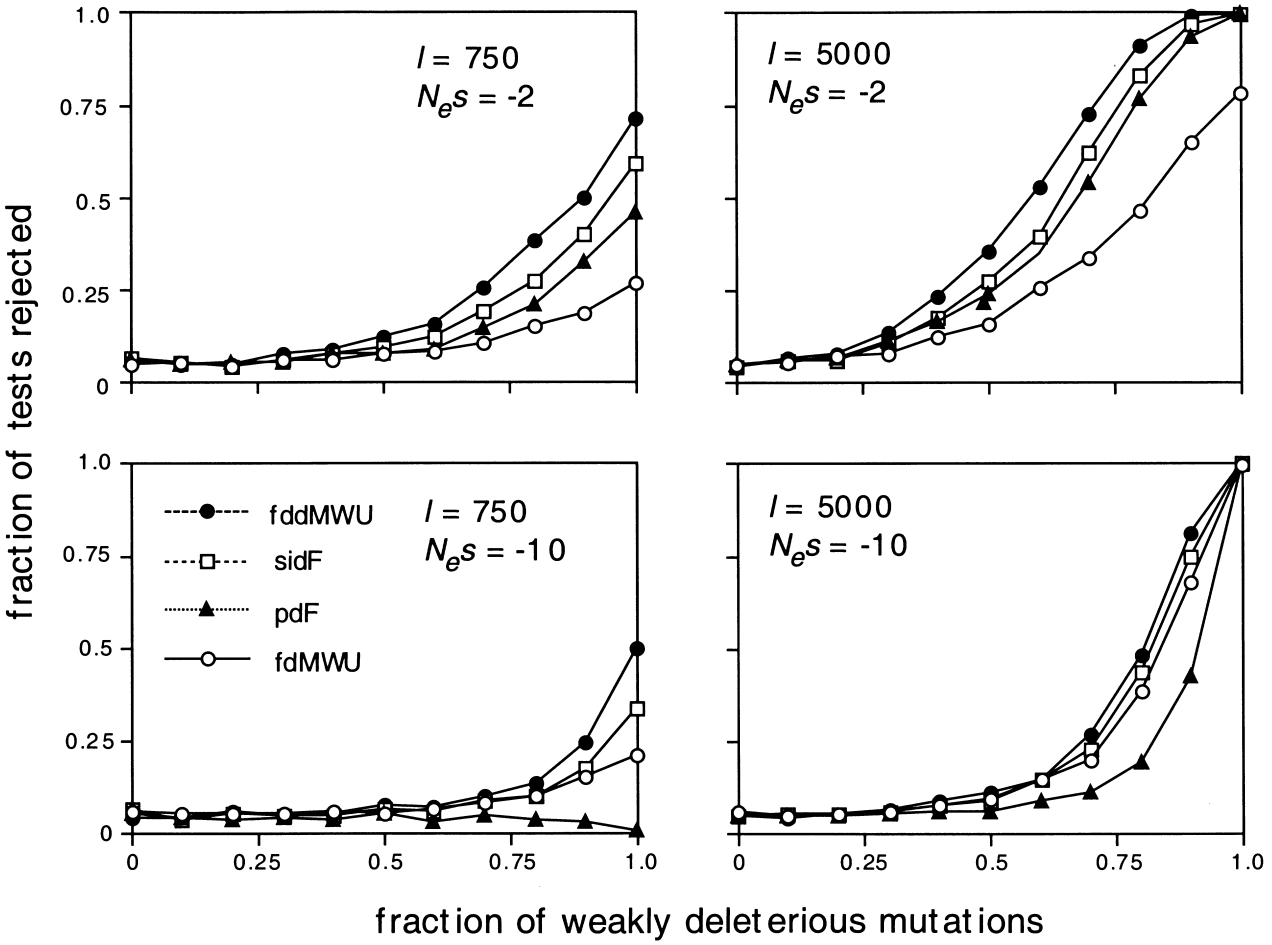

Fig. 6. Statistical power to detect deleterious mutations through configuration comparisons. The *y*-axis plots the proportion of tests that reject fitness equivalence, $P<0.05$, among 1000 simulated data sets for each set of fitness effects of $N_e s$. Two-tailed probabilities were calculated for the four tests described in the text. Data were simulated for $m=5$, $N_e=5\times10^6$, $t_{div}=0.6$, u (per site)$=2\times10^{-9}$ for both classes of mutations. The numbers of mutable sites l were 2500 for the neutral class and 7500 sites in the comparison class. Of the sites in the comparison class, either 90% or 33% mutated to strongly deleterious mutations (these sites did not contribute to the configurations). Of the remaining mutations (shown in the figure as l), the fraction of sites giving rise to neutral and weakly deleterious is plotted on the *x*-axis, and the fitness effects of deleterious mutations are shown on the graphs.

lineage (if effective population sizes decrease such that $N_e s$ falls in a range close to -1).

The role of adaptive fixations in protein evolution also remains a contentious issue. The neutral theory posits that a very small fraction of substitutions have been caused by natural selection. Ohta and Kimura (1971) estimated that about 10% of amino acid difference between species may be driven by positive selection. Power tests were conducted to determine the ability of configuration tests to detect adaptive evolution when it might occur at only a fraction of sites. A neutral class of variation is compared with a second class consisting of strongly deleterious, neutral, and adaptive mutations. Of $l=10\,000$ sites, 2500 are neutral and constitute one class of mutations. In the second class, either 1/3 or 9/10 of sites were strongly deleterious and did not produce mutations that contribute to polymorphism or divergence. Of the remaining mutations in the second class, the fraction of adaptive fixations was varied between 0 and 100% of mutations, and their fitness effects were varied between $N_e s=5$ and 100 (see Fig. 6 legend). Fig. 7 shows the power of five different statistical tests to detect such scenarios of adaptive evolution. The tests include the four discussed above, as well as a comparison of substitution rates between the two classes (this comparison is limited to fixed differences, $r=m$). Faster rates of amino acid than silent evolution have been argued as evidence for adaptive evolution for a number of genes [reviewed in Vacquier and Lee (1993) and Endo et al. (1996)]. In these analyses, Monte Carlo analogs of Fisher's exact test were performed on 2×2 contingency table comparisons of the numbers of neutral and non-neutral sites that have undergone fixations and the numbers of sites at which substitutions have not occurred (KaKsF tests).

Fig. 7 shows that small proportions of adaptive mutations can be detected through configuration compari-

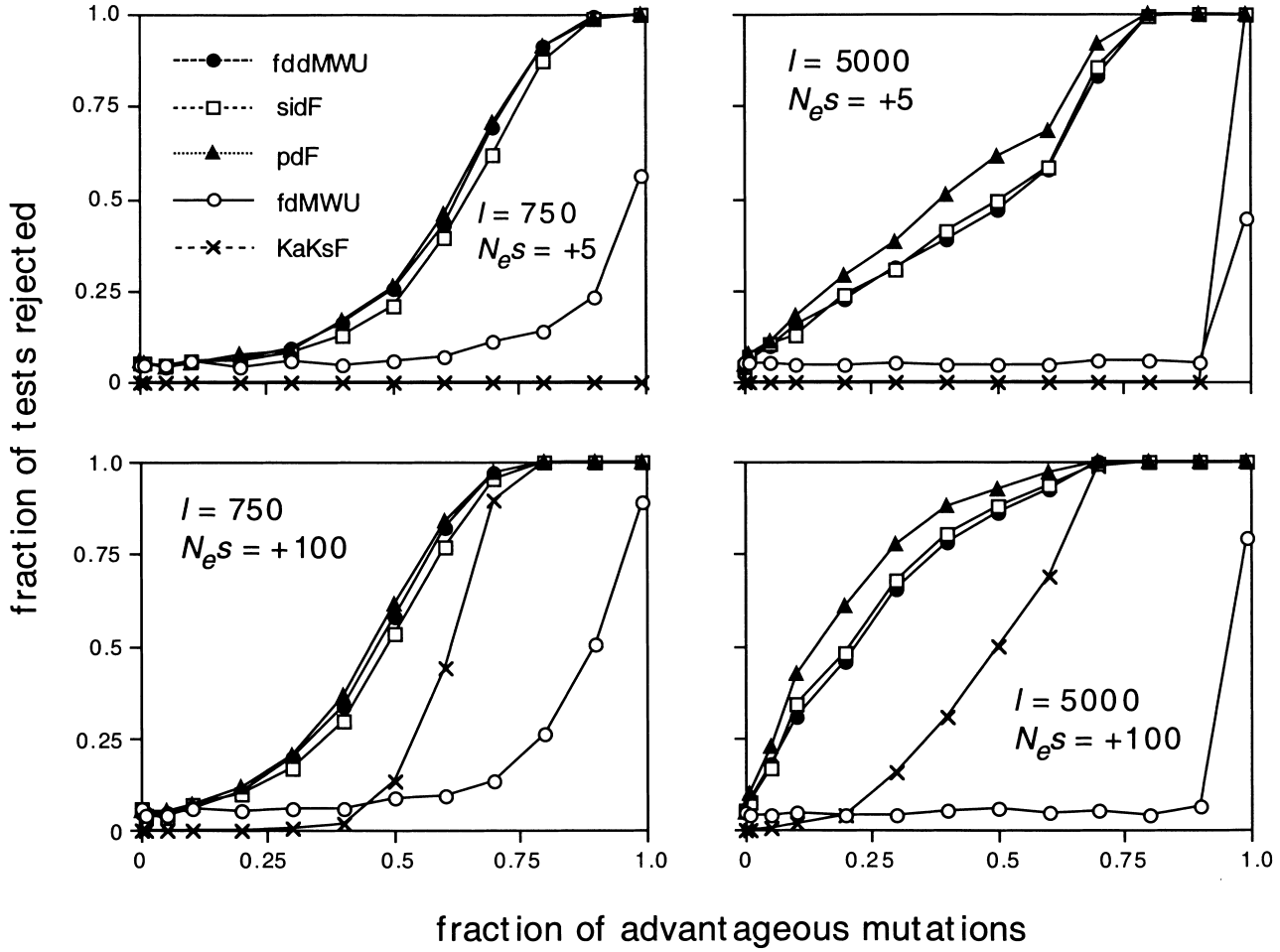

Fig. 7. Statistical power to detect adaptive evolution through configuration comparisons. The *y*-axis plots the proportion of tests that reject fitness equivalence, $P<0.05$, among 1000 simulated data sets for each distribution of selection coefficients. The tests include the four employed in Fig. 6, as well as a test restricted to the fixed difference class, KaKsF (see text). The parameter values are the same as those in the legend of Fig. 6. The numbers of mutable sites *l* were 2500 for the neutral class and 7500 sites in the comparison class. Of the sites in the comparison class, either 90% or 33% mutated to strongly deleterious mutations (these sites did not contribute to the configurations). Of the remaining mutations, the ratios of neutral to adaptive mutations were chosen so that the fraction of adaptive *substitutions* were 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.995 (the remaining substitutions were neutral). To obtain such proportions, the fraction of non-deleterious mutations with positive selection coefficients were set to 0.0010, 0.0020, 0.0060, 0.011, 0.025, 0.042, 0.063, 0.091, 0.13, 0.19, 0.29, 0.47, and 0.95 for $N_e s=5$, and 0.000 010, 0.0010, 0.0020, 0.0030, 0.0040, 0.0050, 0.0060, 0.0070, 0.0080, 0.012, 0.020, 0.044, and 0.50 for $N_e s=100$.

sons (see Fig. 7 legend). The power to detect a mixture of fitness effects increases as a function of both the proportion of adaptive mutations and their fitness advantage. When selection is relatively weak, or when the fraction of adaptive mutations is small, comparisons limited to the fixed difference class are considerably less powerful than methods that include polymorphism data. This is because the sites at which strongly deleterious mutations occur contribute substantial numbers of non-evolved sites in the contingency table and make the test a more conservative one; many cases of adaptive evolution can go undetected [see Nielsen and Yang (1998) for a method to compare evolutionary rates at individual nucleotide sites]. This method is more powerful when regions or sites in proteins expected to be evolving adaptively can be identified a priori (Hughes and Nei, 1988), and, under the assumptions employed here, all methods that include fixation data are more powerful when longer divergence times are examined.

The time of divergence examined in these simulations is small; at $t_{div}=0.6$, neutral fixations are expected to have occurred at about 2.5% of the sites. In practice, however, there are advantages to examining molecular evolution on short lineages. Low divergence allows more reliable estimates of the numbers of fixed differences. If the amount of divergence is sufficiently large, correction formulae [see Li (1996)] are required to transform the observed numbers of nucleotide differences to the number of substitutions that have occurred (this corrects for the numbers of multiple substitutions at a given site). Such a transformation requires an appropriate model of evolutionary change, and confidence intervals

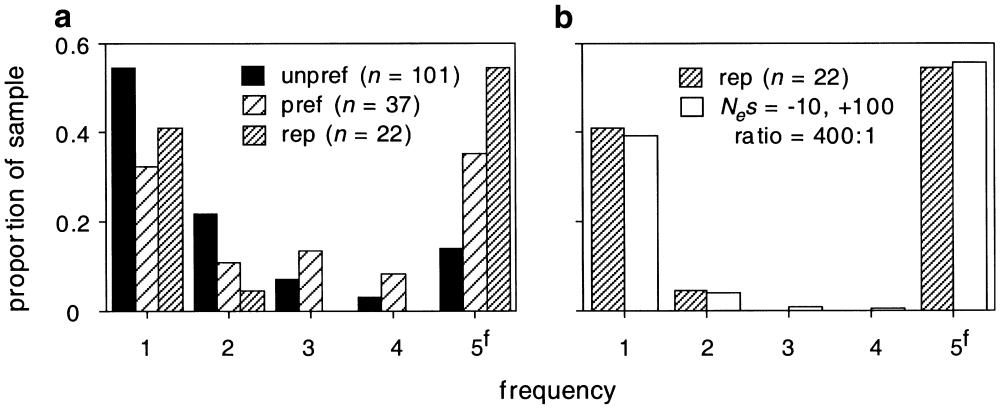

Fig. 8. The configurations of preferred, unpreferred, and replacement mutations in *D. simulans* and the expected configuration under negative and positive selection. (**a**) The proportions of 101 unpreferred (black), 37 preferred (gray), and 22 replacement (striped) mutations segregating at the given frequencies or fixed among five alleles of each of eight *D. simulans* genes are shown. Pooled data from eight *D. simulans* genes from Table 1. 11 of the 12 amino acid fixations in *D. simulans* have occurred in the *Zw* gene, whereas the singleton polymorphisms are distributed more evenly among the eight genes. This suggests that the distribution of selection coefficients may differ among genes. (**b**) The expected configuration of mutations under a combination of $N_e s = -10$ and $+100$ in a ratio of 400:1 is shown to fit the *D. simulans* data remarkably well (the fraction of strongly deleterious mutations was not estimated). The expectations were calculated according to Sawyer and Hartl (1992) and Hartl et al. (1994).

for estimates of the numbers of fixed differences can be large (the error increases as a function of divergence time). In the simulations examined here, the numbers of fixed differences are assumed to be counted without error (under the infinite sites model, all mutations occur at different sites). Another advantage of examining evolution on a relative short time scale is that adaptive evolution may be episodic (Gillespie, 1991). Evidence for short bursts of adaptive evolution may be diluted when examining evolution on a long lineage.

Finally, low divergence allows more accurate inference of ancestral and derived states at variable nucleotide sites. In the simulations discussed above, it is assumed that such inferences can be made with complete accuracy. However, if forward and backward mutation rates are unequal, or if the lineages examined are long, ancestral state inference in DNA sequence data can be error-prone and, under parsimony assumptions, highly biased (Collins et al., 1994; Frumhoff and Reeve, 1994; Yang et al., 1995; Schluter et al., 1997; Zhang and Nei, 1997; Eyre-Walker, 1998).

Although only 22 replacement mutations were found in the *D. simulans* sample, the observed configuration is remarkable. Almost all amino acid changes are either rare polymorphisms or fixed differences. This configuration is significantly different from that of both unpreferred ($P<0.001$, two-tailed sidF test) and preferred ($P=0.028$) silent mutations and suggests a combination of negative and positive selection in protein evolution. In Fig. 8**b**, the configuration of amino acids is compared with that expected under a combination of $N_e s = -10$ and $+100$ in a mutational ratio of 399:1. Surprisingly, the estimated distribution of fitness effects requires no neutral amino acid variants (neutral evolution should produce a detectable density of polymorphic mutations at intermediate frequencies); the pattern suggests that both positive and negative selection operate on amino acid changes and, more surprisingly, that adaptive evolution may play a significant role in protein evolution. In the *D. simulans* data examined, 100% of amino acid fixations may have conferred a fitness advantage. It should be noted, however, that this estimation requires at least three parameters (two selection coefficients and the ratio of the numbers of mutations in each frequency class), and the number of mutations examined is too small to make any firm conclusions. In addition, 11 of the 12 amino acid substitutions in the *D. simulans* lineage have occurred in only one of the eight genes examined, *Zw*, which encodes glucose-6-phosphate dehydrogenase. This locus was chosen for study by Eanes et al. (1993) because it is a regulatory enzyme located at a branch point between metabolic pathways. It will be of great interest to determine how the distribution of selection coefficients in protein evolution differs across loci and over evolutionary time.

5. Conclusions

These analyses suggest that comparisons of DNA variation within and between closely related species can distinguish between neutral, nearly neutral, and adaptive processes. Configuration comparisons between functional classes of variation suggest that, in the *D. simulans* lineage, the nearly neutral model can account for the evolution of silent DNA mutations. Such statistical tests also appear to be a powerful method for detecting deleterious and adaptive protein evolution, even when advantageous amino acid changes are rare and/or their fitness advantages are small. The limited data for protein

evolution from *D. simulans* suggests a significant role of positive selection in protein evolution. However, this finding cannot be generalized without data for a larger number of genes in many more taxa; the relative contributions of deleterious, neutral, and adaptive protein evolution remains an open issue.

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., 1997a. Codon bias evolution in Drosophila: population genetics of mutation-selection-drift. Gene 205, 269–278.

Akashi, H., 1997b. Distinguishing the effects of mutational biases and natural selection on DNA sequence variation letter. Genetics 147, 1989–1991.

Akashi, H., 1999. Inferring the fitness effects of DNA mutations from patterns of polymorphism and divergence: statistical power to detect directional selection under stationarity and free recombination. Genetics 151, 221–238.

Akashi, H., Eyre-Walker, A., 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. 8, 688–693.

Akashi, H., Schaeffer, S.W., 1997. Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics 146, 295–307.

Andersson, S.G.E., Kurland, C.G., 1990. Codon preferences in free-living microorganisms. Microbiol. Rev. 54, 198–210.

Ballard, J.W.O., Kreitman, M., 1994. Unraveling selection in the mitochondrial genome of Drosophila. Genetics 138, 757–772.

Bulmer, M., 1988. Are codon usage patterns in unicellular organisms determined by selection-mutation balance. J. Evol. Biol. 1, 15–26.

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

Collins, T.M., Wimberger, P.H., Naylor, G.J.P., 1994. Compositional bias, character-state bias, and character-state reconstruction using parsimony. Syst. Biol. 43, 482–496.

Eanes, W.F., Kirchner, M., Yoon, J., 1993. Evidence for adaptive evolution of the *G6pd* gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages. Proc. Natl. Acad. Sci. USA. 90, 7475–7479.

Endo, T., Ikeo, K., Gojobori, T., 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. 13, 685–690.

Eyre-Walker, A., 1997. Differentiating between selection and mutational bias letter. Genetics 147, 1983–1987.

Eyre-Walker, A., 1998. Problems with parsimony in sequences of biased base composition. J. Mol. Evol. 47, 686–690.

Freese, E., 1962. On the evolution of base composition of DNA. J. Theor. Biol. 3, 82–101.

Frumhoff, P.C., Reeve, H.K., 1994. Using phylogenies to test hypotheses of adaptation: a critique of some current proposals. Evolution 48, 172–180.

Gillespie, J.H., 1991. The Causes of Molecular Evolution. Oxford University Press, New York.

Golding, G.B., Aquadro, C.F., Langley, C.H., 1986. Sequence evolution within populations under multiple types of mutation. Proc. Natl. Acad. Sci. USA. 83, 427–431.

Hartl, D.L., Moriyama, E.N., Sawyer, S., 1994. Selection intensity for codon bias. Genetics 138, 227–234.

Hudson, R.R., 1993. Levels of DNA polymorphism and divergence yield important insights into evolutionary processes comment. Proc. Natl. Acad. Sci. USA. 90, 7425–7426.

Hughes, A.L., Nei, M., 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335, 167–170.

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

Karotam, J., Boyce, T.M., Oakeshott, J.G., 1995. Nucleotide variation at the hypervariable esterase 6 isozyme locus of *Drosophila simulans*. Mol. Biol. Evol. 12, 113–122.

Kimura, M., 1962. On the probability of fixation of mutant genes in a population. Genetics 47, 713–719.

Kimura, M., 1968. Evolutionary rate at the molecular level. Nature 217, 624–626.

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

King, J.L., Jukes, T.H., 1969. Non-Darwinian evolution. Science 164, 788–798.

King, L.M., 1998. The role of gene conversion in determining sequence variation and divergence in the *Est-5* gene family in *Drosophila pseudoobscura*. Genetics 148, 305–315.

Li, W.-H., 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. 24, 337–345.

Li, W.-H., 1996. Molecular Evolution. Sinauer Associates, Massachusetts.

Long, M., Langley, C.H., 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in Drosophila. Science 260, 91–95.

McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351, 652–654.

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

Nachman, M.W., Boyer, S.N., Aquadro, C.F., 1994. Nonneutral evolution at the mitochondrial NADH dehydrogenase subunit 3 gene in mice. Proc. Natl. Acad. Sci. USA. 91, 6364–6368.

Nachman, M.W., Brown, W.M., Stoneking, M., Aquadro, C.F., 1996. Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics 142, 953–963.

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., 1973. Slightly deleterious mutant substitutions in evolution. Nature 246, 96–98.

Ohta, T., 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 23, 263–286.

Ohta, T., Kimura, M., 1971. On the constancy of the evolutionary rate of cistrons. J. Mol. Evol. 1, 18–25.

Parker, J., 1989. Errors and alternatives in reading the universal genetic code. Microbiol. Rev. 53, 273–298.

Rand, D.M., Dorfsman, M., Kann, L.M., 1994. Neutral and nonneutral evolution of Drosophila mitochondrial DNA. Genetics 138, 741–756.

Sawyer, S.A., Hartl, D.L., 1992. Population genetics of polymorphism and divergence. Genetics 132, 1161–1176.

Sawyer, S.A., Dykhuizen, D.E., Hartl, D.L., 1987. Confidence interval for the number of selectively neutral amino acid polymorphisms. Proc. Natl. Acad. Sci. USA. 84, 6225–6228.

Schluter, D., Price, T., Mooers, A.O., Ludwig, D., 1997. Likelihood of ancestral states in adaptive radiation. Evolution 51, 1699–1711.

Sharp, P.M., Li, W.-H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28–38.

Sharp, P.M., Averof, M., Lloyd, A.T., Matassi, G., Peden, J.F., 1995. DNA sequence evolution: the sounds of silence. Philos. Trans. R. Soc. London 349, 241–247.

Shaw, A., McRee, D.E., Vacquier, V.D., Stout, C.D., 1993. The crystal structure of lysin, a fertilization protein. Science 262, 1864–1867.

Shields, D.C., Sharp, P.M., Higgins, D.G., 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.

Templeton, A.R., 1996. Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochroms oxidase II gene in the Hominoid primates. Genetics 144, 1263–1270.

Vacquier, V.D., Lee, Y.-H., 1993. Abalone sperm lysin: unusual model of evolution of a gamete recognition protein. Zygote 1, 181–196.

Wayne, M.L., Contamine, D., Kreitman, M., 1996. Molecular population genetics of *ref(2)P*, a locus which confers viral resistance in Drosophila. Mol. Biol. Evol. 13, 191–199.

Wise, C.A., Sraml, M., Easteal, S., 1998. Departure from neutrality at the mitochondrial NADH dehydrogenase subunit 2 gene in humans, but not in chimpanzees. Genetics 148, 409–421.

Yang, Z., Kumar, S., Nei, M., 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141, 1641–1650.

Yokoyama, S., 1997. Molecular genetic basis of adaptive selection: examples from color vision in vertebrates. Annu. Rev. Genet. 31, 315–336.

Zhang, J., Nei, M., 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J. Mol. Evol. 44, S139–S146.