Quantifying the Slightly Deleterious Mutation Model of Molecular Evolution

Adam Eyre-Walker,* Peter D. Keightley,† Nick G. C. Smith,*‡ and Daniel Gaffney†

*Centre for the Study of Evolution & School of Biological Sciences, University of Sussex; †Institute of Cell, Animal and Population Biology, University of Edinburgh; and ‡Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University

We have attempted to quantify the frequency and effects of slightly deleterious mutations (SDMs), those that have selective effects close to the reciprocal of the effective population size of a species, by comparing the level of selective constraint in protein-coding genes of related species that have different present-day effective population sizes. In our two comparisons, the species with the smaller effective population size showed lower constraint, implying that SDMs had become fixed. The fixation of SDMs was supported by the observation of a higher fraction of radical to conservative amino acid substitutions in species with smaller effective population sizes. The fraction of strongly deleterious mutations (which rarely become fixed) is >70% in most species. Only ~10% or fewer of mutations seem to behave as SDMs, but SDMs could comprise a substantial fraction of mutations in protein-coding genes that have a chance of becoming fixed between species.

Introduction

In his seminal 1968 paper, Kimura used the term “nearly neutral” to describe mutations whose selective effects are sufficiently small that their fate is largely determined by random genetic drift (Kimura 1968). Ohta and Kimura (1971) later invoked nearly neutral mutations as an explanation for one of the salient observations of molecular evolution, the constancy of the molecular clock. Under a purely neutral model, a constant rate of molecular evolution per unit time across diverse taxa is only expected if the mutation rate per year is constant; yet, it seems more likely that the mutation rate should be constant per generation. Ohta and Kimura (1971) suggested a way in which this problem could be resolved under the neutral theory. They reasoned that the rate of evolution would be constant if many mutations were slightly deleterious, and there was a negative correlation between generation time and effective population size. The reason is as follows: deleterious mutations with selection coefficients lower than $1/N_e$, the reciprocal of the effective population size, are subject to genetic drift and behave as effectively neutral, whereas those with selection coefficients greater than $1/N_e$ are selected against. Thus, organisms with large effective population sizes have a smaller proportion of effectively neutral mutations than organisms with small effective population sizes; if they also have higher mutation rates per unit time because they have shorter generation times, the resultant rate of evolution might be roughly constant per year across taxa. Ohta went on to champion the nearly neutral theory in a series of papers (Ohta 1972b, 1973, 1976, 1977, 1992), but until recently there was little empirical evidence for the existence of nearly neutral mutations.

There are two lines of evidence which suggest that a significant proportion of mutations are slightly deleterious, that is, deleterious mutations with selective effects close to $1/N_e$. The first comes from several studies showing that the level of selective constraint in protein-coding sequences is positively correlated to population size or to correlates of population size. Constraint is usually calculated as one minus the ratio of the rate of nonsynonymous (or amino acid) substitution to the rate of synonymous (or silent) substitution. Under a model in which synonymous mutations are neutral and nonsynonymous mutations are either neutral or deleterious, constraint is the proportion of amino acid mutations which are deleterious and removed by natural selection. A correlation between constraint and a correlate of generation time was first demonstrated by Ohta (1972a), who showed that the ratio of DNA sequence divergence to protein sequence divergence is negatively correlated to generation time across a broad range of animal taxa (mammals and Drosophilids). Because generation time and population size appear to be negatively correlated (Chao and Carr 1993), this study suggested that constraint is positively correlated to the population size of a species. Ohta’s result was corroborated by those of Li, Tanimura, and Sharp (1987) and Ohta (1995), who showed that the ratio of nonsynonymous to synonymous substitution rates is greater in primates and artiodactyls than in rodents (rodents are thought to have larger population sizes than primates and artiodactyls), and by that of Keightley and Eyre-Walker (2000), who found a negative correlation between constraint and generation time over a broad range of animal taxa (mammals, birds, and Drosophilids). Studies of island species have also yielded evidence of slightly deleterious mutations; in both Hawaiian Drosophila (Ohta 1976, 1993) and species of birds restricted to islands (Johnson and Seger 2001), levels of constraint are lower than those in continental species.

The second line of evidence comes from studies of within-population variation. It has been observed that the ratio of polymorphism to substitution is greater for nonsynonymous than for synonymous changes in many mitochondrial DNA data sets (Rand and Kann 1996; Nachman 1998), in nuclear genes of Arabidopsis thali-
Slightly Deleterious Mutations

We estimated the recent effective population size of a species by combining data on DNA sequence polymorphism with data on rates of molecular evolution. The nucleotide diversity (π) in a neutral sequence is expected to be equal to $4N_e \mu$ ($3N_e \mu$ for X-linked sites), where $u$ is the nucleotide mutation rate per site per generation. The level of divergence, $k$, for the sequence between two species is equal to $2tg_u$, where $g$ is the number of generations per year and $t$ is the time of divergence between sequences in the species in years: i.e., $t = t_r + 2N_e/g$, where $t_r$ is the time when the species physically diverged and $N_e$ is the ancestral population size. Serendipitously, the time of divergence in each of our comparisons was estimated using a locally calibrated clock and the time estimated by such methods is $t$, and not $t_r$, so no correction is required for ancestral polymorphism (this effect would have been small anyway because levels of nucleotide diversity are an order of magnitude smaller than divergences in each of our comparisons). An estimate for $N_e$ is therefore $\pi g t / 2k$ ($2\pi g t / 3k$ for X-linked sequences). Estimates of $g$ and $t$ were obtained from the literature or were estimated by ourselves.

We used either introns or intergenic regions to estimate $N_e$, because for the most part these regions are thought to be free of selection. We calculated average estimates for $N_e$, weighting by sequence length. We have used published estimates of $\pi$ where possible but have otherwise estimated the values using data retrieved from GenBank or provided by the authors of the papers cited. To estimate divergence, we randomly chose one of the sequences used to estimate nucleotide diversity and a single out-group sequence.

To estimate effective population sizes in primates, we used the 10-kb noncoding sequences from 1q24 (Yu et al. 2001) and 22q11 (Zhao et al. 2000) in humans and intron sequences from HoxB6 and ApoB in chimpanzees (Deinard and Kidd 2000). We ignored the data from Xq13.3 in humans and chimpanzees because these are from a low-recombination region, so the diversity may be unduly influenced by background selection and genetic hitchhiking (Kaessman, Wiebe, and Paabo 1999; Kaessman et al. 1999). For humans we only considered diversity in African sequences because humans are thought to have expanded out of Africa; the African sequences are therefore likely to better reflect the effective population size of humans. For humans and chimpanzees we assumed a generation time of 25 years because studies of natural human and chimpanzee populations suggest generation times in excess of 27 and 23 years, respectively (see references in Eyre-Walker and Keightley [1999]). We assumed a divergence time of 6 Myr (Goodman et al. 1998).

For rodents, we used diversity data from intron (plus short adjoining lengths of exon) sequences from two X-linked genes surveyed in Mus domesticus, with M. caroli used as an out-group for the divergence data (Nachman 1997). We only included two of the four genes surveyed because there is a correlation between nucleotide diversity and recombination rate in mice (Nachman 1997), and two of the genes that were surveyed came from regions of low recombination. The two intron sequences came from the Gira2 and Ang genes. We inferred an evolutionary divergence date from a local molecular clock calibrated to the date of the Mus-Rattus divergence. But there is substantial uncertainty over the dates of rodent divergences, so we used two alternatives, the first from fossil evidence, which implies an age of 13 Myr for the Mus-Rattus divergence (Jaeger, Tong, and Denys 1986), and the second from a recent molecular analysis, which implies a date of 23 Myr (Adkins et al. 2001). We also assume that mice have 2 generations per year (see references in Keightley and Eyre-Walker [2000]).

For D. melanogaster and D. simulans we used a recent compilation of noncoding sequences from African flies (Andolfatto 2001). The noncoding sequences came from anon1A3, anon1E9, anon1G5, eve, per, vermillion, yp2, and zeste. We restricted our analysis to African lines of D. melanogaster and D. simulans because it is thought that non-African populations have gone through a recent population bottleneck; the African population is therefore likely to give a better estimate of the long-term $N_e$. We assume that there are 10 generations per year in Drosophila and an evolutionary divergence date of 2.5 Myr for D. melanogaster—D. simulans (Powell and Desalle 1995). To estimate constraint we aligned D. melanogaster, D. simulans, and D. yakuba sequences using the D. yakuba sequence as an out-group.

Calculation of Constraint

We calculated levels of constraint using methods based on those described previously (Eyre-Walker and Rand 2000), and in Escherichia coli (N. G. C. Smith and A. Eyre-Walker, unpublished data). This pattern is consistent with the segregation of slightly deleterious amino acid mutations, which contribute to polymorphism but rarely become fixed (Kimura 1983, p. 44). This conjecture gains support from the observation that nonsynonymous mutations tend to segregate at lower frequencies than synonymous mutations in several mitochondrial DNA data sets, which show an excess of nonsynonymous polymorphism (Nielsen and Weinreich 1999).

So far, however, there have been few attempts to quantify the fraction of amino acid mutations that are slightly deleterious and to estimate the strength of selection acting against them (Fay et al. 2001). In this study we attempt to estimate the fraction of mutations which are slightly deleterious by examining the level of constraint in nuclear protein-coding genes in primates, rodents, and Drosophilids, three groups of organisms for which we can also estimate the average recent effective population size. We also examine the nature of any changes in constraint by examining the ratio of radical to conservative amino acid substitutions.

Materials and Methods

Effective Population Size

We estimated the recent effective population size of a species by combining data on DNA sequence polymorphism with data on rates of molecular evolution. The nucleotide diversity ($\pi$) in a neutral sequence is expected to be equal to $4N_e \mu$ ($3N_e \mu$ for X-linked sites), where $u$ is the nucleotide mutation rate per site per generation. The level of divergence, $k$, for the sequence between two species is equal to $2tg_u$, where $g$ is the number of generations per year and $t$ is the time of divergence between sequences in the species in years: i.e., $t = t_r + 2N_e/g$, where $t_r$ is the time when the species physically diverged and $N_e$ is the ancestral population size. Serendipitously, the time of divergence in each of our comparisons was estimated using a locally calibrated clock and the time estimated by such methods is $t$, and not $t_r$, so no correction is required for ancestral polymorphism (this effect would have been small anyway because levels of nucleotide diversity are an order of magnitude smaller than divergences in each of our comparisons). An estimate for $N_e$ is therefore $\pi g t / 2k$ ($2\pi g t / 3k$ for X-linked sequences). Estimates of $g$ and $t$ were obtained from the literature or were estimated by ourselves.

We used either introns or intergenic regions to estimate $N_e$, because for the most part these regions are thought to be free of selection. We calculated average estimates for $N_e$, weighting by sequence length. We have used published estimates of $\pi$ where possible but have otherwise estimated the values using data retrieved from GenBank or provided by the authors of the papers cited. To estimate divergence, we randomly chose one of the sequences used to estimate nucleotide diversity and a single out-group sequence.
Gene Sequence Data for Estimation of Constraint

We estimated levels of evolutionary constraint in protein-coding genes in samples of protein-coding gene sequences extracted from GenBank using ENTREZ. In the comparisons involving mammals, we controlled for differences in the level of evolutionary constraint induced by the specific properties of the samples of genes by compiling “four-way” sets involving mammal-A, mammal-B, mouse, and rat. We computed estimates of constraint between mammal-A and mammal-B and between the rodents; the rodent estimate acts as a control for effects specific to the gene sample. The coding sequences of homologous genes were aligned using CLUSTALX (Thompson, Higgins, and Gibson 1994) and adjusted manually. Sequences corresponding to gaps or insertions were deleted to exclude nonhomologous gene segments from the analysis. For the comparison between human and chimpanzee, we included a set of genes compiled previously (Eyre-Walker and Keightley 1999; Keightley and Eyre-Walker 2000) for which the orthologous mouse and rat gene sequences were also available. These data were augmented by human-chimpanzee-mouse-rat homologues deposited in GenBank during 2000–2001.

Nature of Amino Acid Substitutions

To investigate the nature of changes in constraint, we used parsimony to count the number of conservative and radical amino acid replacements ($N_c$ and $N_R$) in our interspecies comparisons. Amino acid changes were classified as conservative or radical according to a classification by polarity and volume (Zhang 2000), which divides the 20 amino acids into six groups (special: C; neutral and small: A, G, P, S, T; polar and relatively small: N, D, Q, E; polar and relatively large: R, H, K; nonpolar and relatively small: I, L, M, V; nonpolar and relatively large: F, W, Y). Changes within an amino acid group are termed conservative, whereas amino acid changes between groups are termed radical.

We also estimated the rates of substitution at conservative and radical amino acid sites ($D_C$ and $D_R$), as well as at synonymous sites ($D_\nu$), using Zhang’s method (Zhang 2000), which accounts for biases in the relative rates of transition and transversion (the transition/transversion ratio was assumed to be 2 in Drosophila species and 3 in mammals), and in which multiple-hits correction is performed using the Jukes-Cantor formula. For the comparisons between Drosophila species, substitution rates down lineages were calculated using the least squares method (Zhang 2000). The estimation of conservative and radical substitution rates allows measures of conservative and radical constraint: $C_C = 1 - D_C/D_S$ and $C_R = 1 - D_R/D_S$.

Results

Effective Population Size

Estimates of silent site nucleotide diversity ($\pi$) are compiled for humans, chimpanzees, mice, D. melanogaster, and D. simulans in table 1. Levels of diversity are similar across the mammalian species but are an order of magnitude higher in the Drosophilids. The table also gives estimates of the number of substitutions per site in the sequences surveyed for nucleotide diversity, from which it is possible to estimate the nucleotide mutation rate per generation (shown as averages across pairs of lineages), by using the divergence dates and generation times listed in the table (see also Materials and Methods). By combining polymorphism and divergence estimates, we estimate that the recent effective population size of hominids to be around 10,000, whereas the effective population size in mice is an order of magnitude higher. In contrast the two Drosophila species differ by about twofold. Table 1 suggests that within the mammals and Drosophilids there are contrasts in $N_e$ between species which should lead to differences in constraint if slightly deleterious mutations are common.

Constraint

Table 2 shows estimates of constraint in protein-coding genes for human-chimpanzee, mouse-rat, and two lineages of Drosophila. As expected, constraint is significantly lower in hominids than in rodents and in D. melanogaster than in D. simulans.

Constraint in Humans

As we have noted previously, the low constraint in the human-chimpanzee comparison partly reflects the properties of the genes which have been sequenced in these species because the corresponding mouse-rat orthologues have a lower average level of constraint than mouse-rat genes in general (Eyre-Walker and Keightley...
Table 1
Estimates of $N_e$, the Effective Population Size

<table>
<thead>
<tr>
<th>Species Pair</th>
<th>$\pi$</th>
<th>Species</th>
<th>$K_s$</th>
<th>Divergence Time (Myr)</th>
<th>Generation Time (Years)</th>
<th>$N_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens.................</td>
<td>0.00080</td>
<td>Homo-Pan</td>
<td>0.0099</td>
<td>6</td>
<td>25</td>
<td>9,700</td>
</tr>
<tr>
<td>Pan troglodytes verus............</td>
<td>0.00070</td>
<td>Homo-Pan</td>
<td>0.0084</td>
<td>6</td>
<td>25</td>
<td>10,000</td>
</tr>
<tr>
<td>Pan troglodytes troglodytes......</td>
<td>0.0208</td>
<td>Homo-Pan</td>
<td>0.0125</td>
<td>6</td>
<td>25</td>
<td>20,000</td>
</tr>
<tr>
<td>Pan paniscus....................</td>
<td>0.0096</td>
<td>Homo-Pan</td>
<td>0.0130</td>
<td>6</td>
<td>25</td>
<td>8,860</td>
</tr>
<tr>
<td>Mus domesticus..................</td>
<td>0.00144</td>
<td>M. domesticus–M. caroli</td>
<td>0.0274</td>
<td>2.3</td>
<td>0.5</td>
<td>161,000</td>
</tr>
<tr>
<td>Drosophila melanogaster.........</td>
<td>0.0148</td>
<td>D. melanogaster–D. simulans</td>
<td>0.128</td>
<td>2.5</td>
<td>0.1</td>
<td>1,450,000</td>
</tr>
<tr>
<td>Drosophila simulans.............</td>
<td>0.0263</td>
<td>D. melanogaster–D. simulans</td>
<td>0.128</td>
<td>2.5</td>
<td>0.1</td>
<td>2,580,000</td>
</tr>
</tbody>
</table>

Note.—The estimates of silent sit diversity are averages weighted by length, over the compilation of sequences available for the species, and values for X-linked sequences are multiplied by 4/3.

Table 2
Constraint Estimates for Human-Chimpanzee, Mouse-Rat, and Drosophila

<table>
<thead>
<tr>
<th>Species or Species Pair</th>
<th>Number of Genes</th>
<th>Mean Constraint (SE)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drosophila simulansa....</td>
<td>44</td>
<td>0.88 (0.04)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>44</td>
<td>0.80 (0.06)</td>
<td></td>
</tr>
<tr>
<td>Mouse-rat</td>
<td>52</td>
<td>0.76 (0.03)</td>
<td>0.036</td>
</tr>
<tr>
<td>Human-chimpanzee</td>
<td>52</td>
<td>0.62 (0.09)</td>
<td></td>
</tr>
<tr>
<td>Mouse-ratb</td>
<td>432</td>
<td>0.84 (0.008)</td>
<td>—</td>
</tr>
<tr>
<td>Human-chimpanzee</td>
<td>52</td>
<td>0.69 (0.09)</td>
<td></td>
</tr>
</tbody>
</table>

a Large data set of mouse-rat orthologues, which are not paired to human-chimpanzee orthologues.

b Level of constraint corrected using mouse-rat orthologues and large data set of mouse-rat genes.

In the data set of human-chimpanzee genes we have compiled in this study, the constraint in human-chimpanzee genes which have a mouse-rat homologue is 0.62, which is lower than the constraint in the orthologous mouse-rat genes ($C = 0.76, P = 0.036$, one-tail test). But these mouse-rat genes also have significantly lower constraint than mouse-rat genes in general ($C = 0.76$ vs. $C = 0.84$, as found in a data set of 432 mouse-rat orthologues [Makalowski and Boguski 1998]). It is possible to correct the level of constraint in the human-chimpanzee data set using the distribution of constraint values in the large ($n = 432$) data set of mouse-rat orthologues (fig. 1). If $f_x$ is the proportion of mouse-rat genes in the large data set with constraint values between $x - 0.05$ and $x + 0.05$, and $f_x'$ is the proportion in the data set of mouse-rat genes with human-chimpanzee orthologues, then a corrected estimate of constraint in human-chimpanzee is

$$C_{corr} = \frac{\sum U_i F(C_i)}{\sum M_i F(C_i)}$$

where $U_i$ and $M_i$ are the deleterious and amino acid mutation rate estimates, respectively, for human-chimpanzee gene $i$, $C_i$ is the constraint in the homologous mouse-rat gene, and $F(C_i) = f_x f_x'$ for $x - 0.05 < C_i < x + 0.05$. The method works by weighting genes such that the distribution of constraint values in the mouse-rat genes, for which we have human-chimpanzee orthologues, is transformed in such a way that it is identical to the distribution of constraint values in the large data set of mouse-rat genes. Using this method the corrected constraint estimate for human-chimpanzee is $C = 0.69 (0.09)$. This is somewhat lower than the estimate of $\sim80\%$ obtained by Fay, Wycoff, and Wu (2001), but
Table 3

<table>
<thead>
<tr>
<th>Species Pair</th>
<th>Number of Genes</th>
<th>Constraint (SE)</th>
<th>Constraint in Mouse-Rat (SE)</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-chimpanzee</td>
<td>52</td>
<td>0.62 (0.9)</td>
<td>0.76 (0.03)</td>
<td>0.036</td>
</tr>
<tr>
<td>Human-orangutan</td>
<td>22</td>
<td>0.73 (0.06)</td>
<td>0.79 (0.03)</td>
<td>0.11</td>
</tr>
<tr>
<td>Human-macaque</td>
<td>60</td>
<td>0.73 (0.06)</td>
<td>0.79 (0.03)</td>
<td>0.20</td>
</tr>
<tr>
<td>Macaque-African green monkey</td>
<td>15</td>
<td>0.72 (0.10)</td>
<td>0.79 (0.06)</td>
<td>0.18</td>
</tr>
<tr>
<td>Sheep-cow</td>
<td>67</td>
<td>0.69 (0.03)</td>
<td>0.80 (0.02)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sheep-goat</td>
<td>15</td>
<td>0.73 (0.06)</td>
<td>0.71 (0.04)</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Note.—The probability value is from a one-tail test of whether the constraint is lower in the species pair than in mouse-rat.

Further Constraint Comparisons

The earlier mentioned analyses suggest that the level of evolutionary constraint of protein-coding genes is similar in *Drosophila* and rodents, whereas primates have a somewhat lower mean level of constraint. To investigate whether primates are unusual in this respect, we compiled data sets for several other mammalian pairs for which we had orthologous genes in mouse and rat. The resulting estimates of constraint are shown in table 3. Although few of the differences between the constraint of the species pair and mouse-rat are significant, a clear picture emerges: constraint is \( \approx 70\% \) in most mammalian species, which is similar to the corrected level of constraint in human-chimpanzee, whereas it is \( \approx 80\% \) in almost all of the data sets for mouse-rat. In particular, it is worthwhile noting that constraint in human-macaque and human-orangutan is \( \approx 73\% \). These comparisons could be deceptive if macaques and orangutans had higher constraint than humans. To investigate this we compiled a data set of 80 human-macaque homologues with an artiodactyl out-group sequence. This analysis yielded constraint estimates of \( 0.76 \pm 0.03 \) for both the macaque and ape lineages. Therefore, it seems that the constraint in humans and chimpanzees is typical of most mammalian species and that mice and rats are rather atypical.

Table 4

<table>
<thead>
<tr>
<th>Species or Species Pair</th>
<th>( N_R/N_C )</th>
<th>( P )</th>
<th>( C_C )</th>
<th>( P )</th>
<th>( C_R )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Drosophila simulans</em></td>
<td>0.61</td>
<td>&lt;0.001</td>
<td>0.76</td>
<td>0.088</td>
<td>0.89</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>0.23</td>
<td>0.71</td>
<td>0.56</td>
<td>0.61</td>
<td>0.79</td>
<td>0.15</td>
</tr>
<tr>
<td>Mouse-rat</td>
<td>1.12</td>
<td>0.029</td>
<td>0.58</td>
<td>0.58</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>Human-chimpanzee</td>
<td>1.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—The probability value is from a one-tail bootstrap-by-gene test of whether the statistic is higher in the species with smaller current effective population size. \( N_R/N_C \) is the ratio of the numbers of radical and conservative amino acid changes, whereas \( C_C \) and \( C_R \) are measures of constraint at conservative and radical sites, respectively (see Materials and Methods).

The Nature of Changes in Constraint

In both of our comparisons we see a significant difference in the level of overall amino acid constraint between species, or pairs of species, with different current effective population sizes (table 2). Table 4 shows that these differences in constraint are mirrored by differences in the ratio of the numbers of radical and conservative amino acid substitutions; in both the hominid-rodent and *D. melanogaster–D. simulans* comparisons, the ratio of radical to conservative amino acid changes is significantly higher for the species with the lower effective population size.

We have also considered separate measures of constraint in terms of conservative and radical amino acid changes. In both comparisons (hominid-rodent and *D. melanogaster–D. simulans*), there are more marked differences in the level of constraint in radical constraint than in conservative constraint (see table 4).

Discussion

Frequency of Slightly Deleterious Mutations and Their Effects

We have attempted to estimate the proportion of mutations which are slightly deleterious by comparing levels of selective constraint in protein-coding genes between pairs of species that have different present-day effective population sizes. In our two contrasts, we find evidence of a significant difference in constraint that is in line with the expectation that a significant fraction of mutations are slightly deleterious and that these mutations are fixed more readily in species with small effec-
tive population sizes. Interestingly, the lower level of constraint in hominids and D. melanogaster seems to be largely due to an increase in the number of radical amino acid substitutions, relative to conservative replacements.

We can use the differences in constraint in the two data sets, which show a significant difference in constraint, to make inferences about the shape of the distribution of fitness effects in these species (fig. 2). Because the corrected level of constraint in humans and chimpanzees is 69%, this implies that ~69% of all amino acid mutations have selection coefficients more negative than \(1/4N_e(p)\), where \(N_e(p)\) is the long-term effective population size of humans and chimpanzees (note that when \(s = -1/4N_e\), the fixation probability of a deleterious mutation is approximately one-half of the fixation probability of a neutral mutation). The difference in constraint between mouse-rat and human-chimpanzee (table 2), therefore, implies that ~15% of mutations have \(-1/4N_e(p) < s < -1/4N_e(r)\), where \(N_e(r)\) is the long-term effective population size of mouse-rat. If we take \(N_e(p) = 15,000\) and \(N_e(r) = 220,000\) (table 1), the distribution of fitness effects is as shown in figure 2a.

FIG. 2.—Distribution of fitness effects inferred from the comparison of constraint in (a) human-chimpanzee and mouse-rat, and (b) D. melanogaster and D. simulans.

We emphasize that these estimates are crude, relying as they do on several simplifying assumptions and the fact that we only have an estimate of the recent effective population size of our species and not their long-term \(N_e\) (see below). But they provide the first approximate estimate of how prevalent slightly deleterious mutations are.

Limitations of the Analysis

We have not incorporated two major complications into our analysis; these are adaptive substitutions and selection on synonymous codon use. Selection on synonymous codon use could lead to either an underestimation or an overestimation of the level of constraint, depending on whether selection is still in operation. If current selection on synonymous codon use is at a level similar to that in the past, then constraint is underestimated because selection will depress the rate of synonymous substitution and hence the estimate of \(M\), the amino acid mutation rate. In contrast, if selection on synonymous codon use has recently been relaxed, then the rate of amino acid mutation may be overestimated. There is good evidence for past selection on synonymous codon bias in Drosophila (Shields et al. 1988), but selection appears to have been relaxed, to the point of being absent, in D. melanogaster (Akashi 1996; McVean and Vieira 2001). In D. simulans there is evidence for a recent relaxation of selection because codon bias has declined in D. simulans since it diverged from D. melanogaster (Begun 2001; McVean and Vieira 2001); however, population genetic evidence suggests that selection on synonymous codon use is currently in operation (Akashi and Schafer 1997; Kliman 1999; Begun 2001). These considerations suggest that constraint may have been overestimated in D. melanogaster because the codons preferred by selection are all G and C ending in Drosophila, whereas the prevailing mutation bias is AT biased; hence, a relaxation of selection will lead to a mutation rate which is above that experienced at nonsynonymous sites. It seems likely that constraint has also been overestimated in D. simulans, but this overestimation will have been attenuated to some extent by recent selection on synonymous codons in this species. Thus, selection on synonymous codon usage may have reduced the differences in constraint between D. simulans and D. melanogaster that we have observed.

Unfortunately, the situation is less clear in primates and rodents. These two groups differ in their level of synonymous codon bias (Mouchiroud, Gautier, and Bernardi 1988), but we do not fully understand the basis of this difference (Eyre-Walker and Hurst 2001). It is gen-

not seem appropriate to combine the data from the mammalian and Drosophila data sets because it seems unlikely that the distribution of fitness effects would be the same in the two groups; in fact, one might legitimately argue that the distribution of fitness effects is likely to be different in rodents and primates because these groups differ markedly in their level of social interaction. Using the D. simulans and D. melanogaster data, we estimate the distribution to be that shown in figure 2b. We emphasize that these estimates are crude, relying as they do on several simplifying assumptions and the fact that we only have an estimate of the recent effective population size of our species and not their long-term \(N_e\) (see below). But they provide the first approximate estimate of how prevalent slightly deleterious mutations are.

This difference (Eyre-Walker and Hurst 2001). It is gen-
erally accepted that synonymous codon bias is declining in rodents (Mouchiroud, Gautier, and Bernardi 1988; Galtier and Mouchiroud 1998; Smith and Eyre-Walker 2002), and recent evidence suggests that this may be the case in primates (L. Duret, personal communication). If this is the case, then constraint is likely to have been overestimated in both rodents and primates, although this overestimation is probably small for most genes.

There are potentially two sources of advantageous mutations to consider. First, it seems likely that there are slightly advantageous mutations if there are deleterious mutations, and second, there may be strongly selected advantageous mutations contributing to adaptation. First, let us consider a model where there is a balance between weakly advantageous and deleterious mutations. If a slightly deleterious mutation A2 occurs at a site that was fixed for allele A1, and the strength of selection against it is \(-s\), then an A1 mutation will have an advantage of \(+s\) at a site which is fixed for A2. At equilibrium, this model has little effect on either the predictions of the slightly deleterious model used here (i.e., species with large effective population sizes should have high levels of constraint) or the estimation of the shape of the distribution of fitness effects (we can simply replace \(N_e s\) in figure 1 by \([N_e s]\)). But the nonequilibrium situation can be complicated because an increase in \(N_e\) can lead to a temporary decrease in constraint. This arises because sites at which selection has previously been ineffective will often be fixed for a deleterious mutation; when \(N_e\) increases, advantageous mutations can become fixed, leading to a temporary increase in the rate of evolution which will manifest itself as a decrease in constraint. But it seems likely that in each of the comparisons we have studied, the prevailing trend has been toward a decline in effective population size, rather than an increase.

It has been estimated that \(\sim 35\%\) and \(\sim 45\%\) of all amino acid substitutions are adaptive in humans (Fay et al. 2001) and Drosophila (Bustamante et al. 2002; Fay et al. 2002; Smith and Eyre-Walker 2002), respectively. The fixation of strongly advantageous mutations will reduce the level of constraint and thus lead to overestimation of the proportion of mutations which are slightly deleterious: if a proportion \(\alpha\) of the substitutions are advantageous, then the proportion of mutations which are effectively neutral is \((1 - \alpha) K_e/M = (1 - \alpha) (1 - C)\). For example, if we accept that \(45\%\) of substitutions in Drosophila are advantageous, then we estimate that \(89\%\) of mutations are more deleterious than \(1/4N_e(\text{mel})\), a proportion which is otherwise underestimated by \(2N_m u\) of this model. Rather more mutations will lie between two limits if the rate of adaptation is positively correlated to the effective population size. We might expect the rate of adaptation to be correlated to population size if the rate of adaptation is mutation limited because the rate of evolution under this model is equal to \(2N_m u\), where \(u\) is the mutation rate, hence the proportion of mutations which are advantageous, and \(s\) is the average strength of selection in favor of the advantageous mutations (where \(s < \frac{1}{16}\) and \(N_e s \gg 1\)).

**Summary**

Our results suggest that a large majority of amino acid mutations are strongly deleterious in all the species we investigated. Slightly deleterious mutations, those mutations with effects close to \(1/N_e\), could be a substantial fraction of fixed mutations because we found evidence for differences in selective constraint and the kinds of amino acid that are fixed between species with different recent effective population sizes.

**Acknowledgments**

We are very grateful to Peter Andolfatto for his help in compiling the Drosophila data. We also thank the Royal Society of London (A.E.W. and P.D.K.) and the BBSRC (N.G.C.S. and A.E.W.) for support provided.

**LITERATURE CITED**


