Adaptive protein evolution in *Drosophila*

Nick G. C. Smith*† & Adam Eyre-Walker*

** Centre for the Study of Evolution and School of Biological Sciences, University of Sussex, Brighton BN1 9QG, UK*

For over 30 years a central question in molecular evolution has been whether natural selection plays a substantial role in evolution at the DNA sequence level[1,2]. Evidence has accumulated over the last decade that adaptive evolution does occur at the protein level[3,4], but it has remained unclear how prevalent adaptive evolution is. Here we present a simple method by which the number of adaptive substitutions can be estimated and apply it to data from *Drosophila simulans* and *D. yakuba*. We estimate that 45% of all amino-acid substitutions have been fixed by natural selection, and that on average one adaptive substitution occurs every 45 years in these species.

Mutations can spread through a population either by random genetic drift or by the action of natural selection. The relative contributions of these two processes to evolution at the DNA level is one of the oldest and most keenly debated questions in molecular evolution[1,2]. Yet despite extensive analysis[3], no consensus has been reached. With the great increase in single nucleotide polymorphism data, however, we are now in a position to tackle this question.

The proportion of adaptive mutations can be estimated by a simple extension of the McDonald–Kreitman test[5–7]. Let us begin by assuming that all synonymous mutations are neutral, and that all non-synonymous mutations are either strongly deleterious, neutral or strongly advantageous. Under this model the numbers of synonymous (P_s) and non-synonymous (P_n) polymorphisms segregating in a sample of sequences from a population are equal to $4N_e uL_s k$ and $4N_e ufL_n k$ respectively for an autosomal locus, where N_e is the effective population size[2], u is the nucleotide mutation rate, f is the proportion of amino-acid mutations which are neutral, L_s and L_n are the numbers of synonymous and non-synonymous sites respectively and k is a constant reflecting the probability of observing a neutral variant. The constant k is dependent upon a number of factors including the number of sequences sampled, the sampling strategy, demography, background selection[8] and genetic hitch-hiking[9]; however, because synonymous and non-synonymous sites are interspersed, the value of k is the same for neutral mutations at each type of site. We assume that advantageous mutations contribute little to polymorphism, although they may contribute substantially to the divergence between species. This is not an unrealistic assumption; at most an advantageous mutation will contribute twice as much heterozygosity during its lifetime as a neutral variant[2]. For example, if advantageous mutations, with an advantage of $N_e s = 25$ (where s is the strength of selection) occur at one-hundredth the rate of neutral mutations, they will account for 50% of substitutions, but account for just 2% of the heterozygosity. The numbers of synonymous (D_s) and non-synonymous (D_n) substitutions are $2utL_s$, and $2utfL_n + a$, where t is the time of divergence between the two species being considered (strictly the average time to coalescence of the genealogies of the sites being considered), and a is the number of adaptive substitutions. It is not difficult to show from these equations that the number of adaptive substitutions in a gene can be estimated by

$$a = D_n - D_s \frac{P_n}{P_s} \quad (1)$$

† Present address: Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden.

Table 1 Proportion of amino-acid substitutions driven by positive selection

Data set	x	Number of genes	$\overline{\alpha}$ (95% C.I.)	Proportion of $\overline{\alpha} < 0$
All genes				
	0	35	0.24 (–0.11, 0.49)	0.083
	5	30	0.43 (0.21, 0.61)	0.001
	10	25	0.48 (0.30, 0.65)	0
	20	12	0.52 (0.26, 0.74)	0
Excluding *mth*, *Zw* and *Hex-t1*				
	5	27	0.35 (0.08, 0.53)	0.011

Genes with $P_s \leq x$ were excluded. The last column gives the proportion of bootstrap replicates in which $\overline{\alpha}$ was estimated to be less than zero. C.I., confidence interval.

So dividing this expression by D_n gives an estimate of the proportion of amino-acid substitutions driven by positive selection

$$\alpha = 1 - \frac{D_s P_n}{D_n P_s} \quad (2)$$

To estimate the average proportion of amino-acid substitutions which are driven by adaptive evolution we need to combine data across genes. Unfortunately, both equations (1) and (2) are undefined if $P_s = 0$, and equation (2) is undefined if $D_n = 0$. Furthermore, caution must be exercised in summing the values of D_n, D_s, P_n and P_s across genes as this will give an overestimate of adaptive substitution if N_e and f are negatively correlated, as we might expect them to be: N_e is thought to vary across the genome in *Drosophila*[10] owing to processes such as genetic hitch-hiking[9] and background selection[8], and this variation, in association with slightly deleterious mutations, will generate a negative correlation between N_e and f. We have therefore estimated the average proportion of amino-acid substitutions driven by positive selection by the expression

$$\overline{\alpha} = 1 - \frac{\overline{D}_s}{\overline{D}_n}\overline{\left(\frac{P_n}{P_s + 1}\right)} \quad (3)$$

where all averages are across genes. We use the average of $P_n/(P_s + 1)$ to estimate the mean value of $L_n f/L_s$, rather than the average of P_n/P_s, because $P_n/(P_s + 1)$ is defined for all genes and is less biased; it is essentially unbiased if $P_s > 5$. The rationale behind equation (3) is given the Methods section. The confidence interval of $\overline{\alpha}$ was obtained by bootstrapping the data by randomly selecting genes with replacement.

We have used our method to estimate the proportion of amino-acid substitutions which have been driven by positive selection in the divergence between *D. simulans* and *D. yakuba* by using polymorphism data from *D. simulans* (see Supplementary Information for values of D_n, D_s, P_n and P_s). To do this we compiled a data set of 43 genes for which we had multiple sequences from *D. simulans* and an orthologous *D. yakuba* sequence. From this data set we

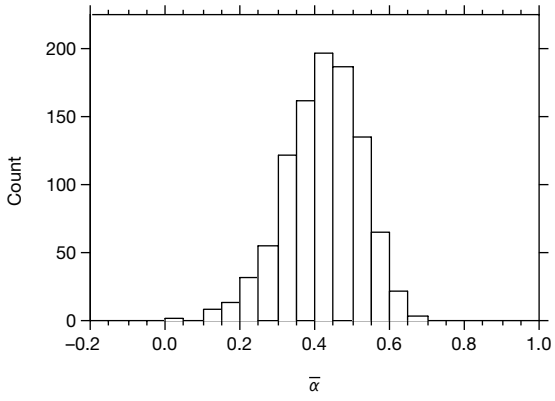

Figure 1 The distribution of 1,000 bootstrap values of $\overline{\alpha}$ for the divergence between *Drosophila simulans* and *D. yakuba* for genes in which $P_s > 5$. $\overline{\alpha}$ is the average proportion of amino-acid substitutions driven by positive selection.

Table 2 Numbers of synonymous and non-synonymous substitutions

Lineage	$\overline{D}_n$	$\overline{D}_s$	$\overline{D}_n/\overline{D}_s$
D. simulans	6.15	14.40	0.43
D. yakuba	23.97	56.71	0.42
D. melanogaster	10.91	18.26	0.60

Substitutions are shown along the lineages leading to *D. simulans*, *D. yakuba* and *D. melanogaster* from the node which connects them. The ratio $\overline{D}_n/\overline{D}_s$ is not significantly different between lineages. $\overline{D}_n$ and $\overline{D}_s$ are the average numbers of non-synonymous and synonymous substitutions.

discarded four genes that had been sequenced because they were thought to be likely targets of adaptive evolution, and four other genes because they contained no polymorphism in the *D. simulans* alleles. Using the remaining 35 genes we estimate that 24% of all amino-acid substitutions between *D. simulans* and *D. yakuba* have been driven by positive adaptive evolution (Table 1). This estimate is not significantly different from zero. However, some genes have very little polymorphism and therefore contribute a substantial amount of variance to the estimate of $\overline{\alpha}$. If we remove genes which have five or fewer synonymous polymorphisms the estimate of $\overline{\alpha}$ is increased to 43%, which is significantly greater than zero (Table 1, Fig. 1). Removal of genes with more synonymous polymorphisms increases the estimate of $\overline{\alpha}$ slightly to about 50%. We therefore estimate that $\overline{\alpha}$ is about 45%. This is similar to an estimate recently obtained in primates: $\overline{\alpha} = 35\%$ (ref. 6). However, the removal of genes with low P_s values has the potential to bias the estimate of $\overline{\alpha}$ upwards; we therefore conducted a series of simulations which suggest that removing genes with low P_s values is not a problem in this data set (see Methods).

We noticed that our estimate of $\overline{\alpha}$ is reduced only slightly if we remove three genes (*mth*, *Zw* and *Hex-t1*) that individually show evidence of adaptive substitution when we conduct a McDonald–Kreitman test[7] (at $P = 0.05$, not correcting for multiple tests) (Table 1). Thus our method reveals evidence of adaptive evolution even in genes which do not individually show any evidence of positive selection.

However, the method makes two critical assumptions: (1) that the average proportion of amino-acid mutations which are neutral, $\overline{f}$, is constant through time; and (2) that the mutations segregating within a species are neutral. To test whether a change in $\overline{f}$ could be responsible for the high estimate of $\overline{\alpha}$ we used *D. melanogaster* to partition the values of D_n and D_s between the lineages leading to *D. simulans* and *D. yakuba* (see Methods). If there has been little or no adaptive evolution (that is, $\overline{\alpha} = 0$) but $\overline{f}$ was different in the past, then we expect the ratio $\overline{D}_n/\overline{D}_s$ to differ between the lineages leading to *D. simulans* and *D. yakuba*. The ratio, however, is almost identical in those lineages, whereas it is substantially higher along the lineage leading to *D. melanogaster* (Table 2).

The method also makes the assumption that the mutations segregating as polymorphisms within the sample of sequences are neutral, whereas there is evidence that selection is acting upon synonymous mutations segregating in *D. simulans*[11,12]. However, our estimate of $\overline{\alpha}$ will only be an overestimate if the synonymous mutations which are segregating are more deleterious on average than the non-synonymous mutations which are segregating. To investigate this, we calculated the difference between the average frequency of synonymous and the average frequency of non-synonymous polymorphisms segregating in each gene, taking the frequency of each polymorphism as the frequency of the minor allele. There is no apparent difference between the frequencies of two types of polymorphism in our sample (mean difference in frequencies is –0.004, $P > 0.1$ in a paired *t*-test), which suggests that synonymous and non-synonymous polymorphisms are subject to similar levels of selection, and that our estimate of $\overline{\alpha}$ is unbiased.

We have estimated that approximately 45% of all amino-acid substitutions between *D. simulans* and *D. yakuba* have been adaptive. There are about 13,600 genes in the *Drosophila* genome, of average length of 590 codons[13], and the average number of amino-acid substitutions separating *D. simulans* and *D. yakuba* is 0.074 per codon (calculated from our data); we therefore estimate that there have been approximately 270,000 positively selected amino-acid substitutions in the evolution of *D. simulans* and *D. yakuba*. Given that *D. simulans* and *D. melanogaster* are thought, on the basis of biogeographic data, to have diverged 2.5 Myr ago[14], we estimate from the data in Table 2 that *D. simulans* and *D. yakuba* diverged ~6 Myr ago. This implies that these two species have undergone one adaptive substitution every 45 years, or one substitution every 450 generations if *Drosophila* undergoes ten generations a year. This is consistent with Haldane's cost of natural selection[15], the problem which first motivated Kimura to propose the neutral theory of molecular evolution[16]. □

Methods

Data

Collections of sequence data were assembled by searching the literature and sequence databases. If DNA sequence alignments were unavailable, we obtained our DNA alignments on the basis of protein alignments generated using the default parameters of ClustalW[17] as implemented in DAMBE version 4.0.31 (ref. 18). From our initial data set of 43 genes we excluded four genes because they segregated no polymorphisms in the sequences surveyed, and four genes that had been sequenced because they were thought to have undergone adaptive substitution; these latter genes were three reproductive genes (*janA*, *janB* and *ocn*) and one immune system gene (*relish*). We did not exclude genes that were sequenced because they segregated balanced polymorphisms or genes sequenced because they had high rates of non-synonymous substitution.

Analysis

Polymorphisms were counted using DnaSP[19]. To determine the number of substitutions separating *D. simulans*, *D. yakuba* and *D. melanogaster* we aligned a single sequence from each species; if multiple alleles were available we randomly selected one. The number of synonymous and non-synonymous differences were calculated using the program *codeml*, as implemented in the PAML package[20] with codon frequencies estimated from the nucleotide frequencies at the three codon positions and the ratio of K_a/K_s free to vary down each lineage. (See the Supplementary Information for a list of genes along with their D_n, D_s, P_n and P_s values.)

To maintain consistency with the other analyses, McDonald–Kreitman tests were performed using the numbers of synonymous and non-synonymous differences calculated using PAML. The significance of the McDonald–Kreitman test was assessed by a Fisher's exact test; all genes which were significant at 5% showed an excess of non-synonymous substitution, rather than non-synonymous polymorphism.

Rationale for equation (3)

Equation (3) can be expanded to a form which reflects the rationale better: $\overline{\alpha} = (\overline{D}_n - \overline{D}_s(\overline{P_n/(P_s+1)}))/\overline{D}_n$. The numerator estimates the average number of amino-acid substitutions driven by positive selection per gene, and the expression $\overline{(P_n/(P_s+1))}$ estimates the mean of L_nf/L_s. We use $\overline{(P_n/(P_s+1))}$ rather than $\overline{(P_n/P_s)}$ because the former is less biased—it is essentially unbiased if $P_s > 5$—and is defined for all values of P_s. The numerator is expected to be unbiased since $E(D_n - D_sz) = E(D_n) - E(D_s)E(z)$ if D_s and $z = L_nf/L_s$ are uncorrelated ($E(y)$ is the expected value of y); we do not expect D_s and z to be strongly correlated and there is no correlation in our data sets between D_s and $P_n/(P_n+1)$ either for all genes ($r^2 = 0.006$, $P > 0.1$) or for genes for which $P_s > 5$ ($r^2 = 0.05$, $P > 0.1$).

Simulations

To investigate the bias associated with estimating $\overline{\alpha}$ we conducted a simulation in which we used the observed values of P_s and D_s to generate new data sets. Values of D_n, D_s, P_n and P_s were randomly generated from Poisson distributions with means of $\hat{D}_sz$, $\hat{D}_s$, $\hat{P}_sz$ and $\hat{P}_s$ respectively where $\hat{D}_s$ and $\hat{P}_s$ are the observed values of D_s and P_s and $z = L_nf/L$. The value of $\overline{\alpha}$ was calculated for each replicate data set, removing genes with $P_s < x$ after randomly generating a data set to simulate the effect of removing genes with low P_s. We investigated three models: (1) a single value of z was applied to all genes, (2) z was allowed to vary randomly, and (3) z was negatively correlated to P_s. The value of α was set to zero in each gene. For each set of parameters (that is, values of z and the value of P_s below which genes are discarded) 100 replicate data sets were produced. We calculated the mean of $\overline{\alpha}$ and the 95% percentiles. In simulations with $P_s > 5$ the mean of $\overline{\alpha}$ was less than 5% in all simulations with the 95% percentile being below 13% in all simulations (see the Supplementary Information for additional simulation results).

Received 15 May; accepted 19 December 2001.

1. Gillespie, J. H. *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford, 1991).
2. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
3. Kreitman, M. & Akashi, H. Molecular evidence for natural selection. *Annu. Rev. Ecol. Syst.* **26,** 403–422 (1995).
4. Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15,** 496–503 (2000).

5. Charlesworth, B. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.* **63,** 213–227 (1994).
6. Fay, J., Wycoff, G. J. & Wu, C.-I. Positive and negative selection on the human genome. *Genetics* **158,** 1227–1234 (2001).
7. McDonald, J. H. & Kreitman, M. Adaptive evolution at the *Adh* locus in *Drosophila*. *Nature* **351,** 652–654 (1991).
8. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134,** 1289–1303 (1993).
9. Maynard Smith, J. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23,** 23–35 (1974).
10. Begun, D. J. & Aquadro, C. F. levels of naturally occuring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356,** 519–520 (1992).
11. Begun, D. The frequency distribution of nucleotide variation in *Drosophila simulans*. *Mol. Biol. Evol.* **18,** 1343–1352 (2001).
12. Kliman, R. Recent selection on synonymous codon usage in *Drosophila*. *J. Mol. Evol.* **49,** 343–351 (1999).
13. Adams, M. D. *et al*. The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
14. Powell, J. R. & DeSalle, R. Drosophila molecular phylogenies and their uses. *Evol. Biol.* **28,** 87–138 (1995).
15. Haldane, J. B. S. The cost of natural selection. *J. Genet.* **55,** 511–524 (1957).
16. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217,** 624–626 (1968).
17. Thompson, J. D., Higgins, D. G. & Gibson, T. J. ClustalW—improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22,** 4673–4680 (1994).
18. Xia, X. *Data Analysis in Molecular Biology and Evolution* (Kluwer Academic, London, 2000).
19. Rozas, J. & Rozas, R. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15,** 174–175 (1999).
20. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555–556 (1997).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com).

Acknowledgements

We thank B. Charlesworth, C.-I. Wu, S. Otto, M. Whitlock, T. Johnson, P. Awadalla, J. Gillespie, G. McVean and P. Keightley for helpful discussions, and E. Moriyama for help with data collection. N.G.C.S. was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) and A.E.-W. is funded by the Royal Society and the BBSRC.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.E.-W. (e-mail: a.c.eyre-walker@sussex.ac.uk).

Testing the neutral theory of molecular evolution with genomic data from *Drosophila*

Justin C. Fay*†, Gerald J. Wyckoff*† & Chung-I Wu*‡

** Committee on Genetics, University of Chicago, Chicago, Illinois 60637, USA*
‡ Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA

Although positive selection has been detected in many genes, its overall contribution to protein evolution is debatable[1]. If the bulk of molecular evolution is neutral, then the ratio of amino-acid (*A*) to synonymous (*S*) polymorphism should, on average, equal that of divergence[2]. A comparison of the *A*/*S* ratio of polymorphism in *Drosophila melanogaster* with that of divergence from *Drosophila simulans* shows that the *A*/*S* ratio of divergence is twice as high—a difference that is often attributed to positive selection. But an increase in selective constraint owing to an increase in effective population size could also explain this observation, and, if so, all genes should be affected similarly. Here we show that the difference between polymorphism and divergence is limited to only a fraction of the genes, which are also evolving more rapidly, and this implies that positive selection is responsible. A higher *A*/*S* ratio of divergence than of polymorphism is also observed in other species, which suggests a rate of adaptive evolution that is far higher than permitted by the neutral theory of molecular evolution.

The neutral theory holds that the bulk of DNA divergence between species is driven by mutation and drift, rather than by positive darwinian selection[3]. But because the effect of positive selection is often masked by negative selection[4], detecting positive selection is a challenging task. A rate of amino-acid substitution greater than that of synonymous substitution can be explained only by positive selection[5], but such a criterion is very stringent as negative selection lowers the rate of amino-acid substitution. A high rate of amino-acid substitution is limited mostly to genes that are involved in resistance to disease or in sexual reproduction, where there is continual room for improvement[6,7].

The McDonald–Kreitman test can detect positive selection even in the presence of negative selection through a ratio of amino-acid divergence to synonymous divergence greater than that of polymorphism[2]. The *A*/*S* ratio of divergence is inflated above polymorphism by advantageous amino-acid mutations, which quickly sweep through a population but have a cumulative effect on divergence. The McDonald–Kreitman test has been applied to many genes individually, but only a few have yielded a significant excess of amino-acid divergence (*Drosophila* genes are reviewed in refs 8, 9). This may in part be caused by a lack of power in detecting positive selection in individual genes unless a large number of adaptive substitutions have occurred.

For those genes that have yielded a significant McDonald–Kreitman test result, the *A*/*S* ratio of divergence is more than twice as great as polymorphism[10–12]. The effects of positive selection may also be obscured by slightly deleterious amino-acid mutations that inflate the *A*/*S* ratio of polymorphism but not divergence. The effects of slightly deleterious mutations can be removed by comparing common polymorphism with divergence, because deleterious amino-acid mutations are kept at low frequency in the population[4]. This can only be done when the data from a large number of genes are combined; individual genes rarely contain more than a few common amino-acid polymorphisms.

An important but rarely appreciated assumption of the McDonald–Kreitman test is that the selective constraint on a gene remains constant over time. The selective constraint on a gene is determined by the proportion of amino-acid mutations that are deleterious[3], $2Ns < -1$, so both a change in the selection coefficient (s) and a change in effective population size (N) can result in a change in selective constraint. Although it is well known that selective constraint is not static across phylogenetic lineages[13,14], this assumption is rarely justified in applications of the McDonald–Kreitman test. Whereas the strength of selection on each gene might fluctuate over time depending on the genetic or environmental background, a genome-wide change in constraint, such as that caused by a change in effective population size, should produce a consistent increase or decrease in the *A*/*S* ratio across all genes. Alternatively, under positive selection each gene might be affected to a different degree and some genes might not be affected at all.

To compare genomic patterns of amino-acid and synonymous

Table 1 Polymorphisms in *D. melanogaster* and divergence from *D. simulans*

Gene*	Class	Amino-acid polymorphism, *A*	Synonymous polymorphism, *S*	*A*/*S*
X-linked	Rare (≤12.5%)	4	67	0.06
	Common (>12.5%)	6	46	0.13
	Divergence	42	189	0.22
Autosomal	Rare	79	126	0.63
	Common	44	118	0.37
	Divergence	421	521	0.81

* There are 5 X-linked and 31 autosomal genes with a sample size of eight or greater (see text for the data from all 45 genes).

† Present addresses: Department of Genome Sciences, Lawrence Berkeley National Laboratory, Berkeley, California 94720 (J.C.F.); Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA (G.J.W).