Full text provided by www.sciencedirect.com

The genomic rate of adaptive evolution

Adam Eyre-Walker

National Evolutionary Synthesis Center, Durham, NC 27705, USA
Centre for the Study of Evolution, School of Life Sciences, University of Sussex, Brighton, UK, BN1 9QG

The role of positive darwinian selection in evolution at the molecular level has been keenly debated for many years, with little resolution. However, a recent increase in DNA sequence data and the development of new methods of analysis have finally made this question tractable. Here, I review the current state-of-play of the field. Initial estimates in *Drosophila* suggest that ~50% of all amino acid substitutions, and a substantial fraction of substitutions in non-coding DNA, have been fixed as a consequence of adaptive evolution. Estimates in microorganisms are even higher. By contrast, there is little evidence of widespread adaptive evolution in our own species.

Introduction

In this post-genomic era, it is easy to forget just how big many genomes are and, as a consequence, how many DNA-sequence differences separate even closely related species. For example, if we consider ourselves, our closest relative, the chimpanzee, is only 1% divergent from us, but this amounts to ~34 million single nucleotide differences. Even if we concentrate our attention on the protein-coding complement of the genome, there are still ~60 000 amino acid differences between us and chimpanzees. So what proportion of these differences have been fixed because they were beneficial and enabled us and chimpanzees to adapt to our respective environments, and how many were simply fixed by random genetic drift? For over 30 years, evolutionary biologists have debated this question as one part of the neutralist–selectionist debate in relation to all organisms, including humans. However, the question has remained unresolved, largely through a lack of data. But now, with the publication of many genome sequences and the increasingly large amounts of DNA sequence diversity data, we are beginning to answer this question.

Here, I examine the methods that have been used to quantify the rate of adaptive evolution. I then describe the work that has been done on this question before briefly discussing some of the questions that the results raise.

Methods

The McDonald–Kreitman test

That we can potentially estimate the proportion of substitutions that are adaptive is due, in large part, to John McDonald and Marty Kreitman, who devised what is now known as the McDonald–Kreitman (MK) test (Box 1) of neutral evolution [1], which was based on the Hudson–Kreitman–Aguade test [2]. In the MK test, the numbers of polymorphisms (see Glossary) are contrasted to the number of substitutions at two classes of sites; within a protein-coding sequence, these sites might be synonymous and non-synonymous sites, or within a regulatory region they might be protein binding and non-binding sites. Here, I assume for simplicity that the two classes of site are synonymous and non-synonymous: the polymorphisms are generally single nucleotide polymorphisms (SNPs) segregating within a set of alleles and substitutions are fixed differences between species. I denote the number of non-synonymous substitutions as D_n, the number of synonymous substitutions as D_s, the number of non-synonymous polymorphisms as P_n and the number of synonymous polymorphisms as P_s. If all mutations are either strongly deleterious or neutral, then D_n/D_s is expected to roughly equal P_n/P_s. This is the basis of the MK test of molecular evolution. By contrast, if synonymous mutations are neutral, and some of the non-synonymous substitutions have been fixed by positive adaptive evolution, then D_n/D_s should be greater, on average, than P_n/P_s. This is because adaptive mutations contribute relatively more to substitution than to polymorphism, when compared with neutral mutations.

Glossary

α: the proportion of substitutions driven by positive selection.
Background selection: the process by which the removal of deleterious mutations reduces the effective population size, and hence the level of genetic variation, within a genomic region. Regions of low recombination are particularly prone to this process.
D_n: the number of non-synonymous substitutions per gene.
d_n: the number of non-synonymous substitutions per site. An alternative symbol is K_a.
D_s: the number of synonymous substitutions per gene.
d_s: the number of synonymous substitutions per site. An alternative symbol is K_s.
Hitch-hiking: the process by which an advantageous substitution removes genetic variation from the population. Regions of low recombination are particularly prone to this process.
K_a: the number of non-synonymous substitutions per site. An alternative symbol is d_n.
K_s: the number of synonymous substitutions per site. An alternative symbol is d_s.
Negative selection: selection against a deleterious mutation.
Non-coding DNA: DNA that does not encode a protein-coding or RNA gene. This includes introns and intergenic DNA.
Non-synonymous: a mutation or substitution that changes an amino acid, sometimes confusingly called a replacement.
P_n: the number of non-synonymous polymorphisms per gene.
P_s: the number of synonymous polymorphisms per gene.
Polymorphism: a genetic variant found segregating within a population. Single nucleotide polymorphisms are the most common polymorphism in most DNA sequences.
Positive selection: selection in favour of an advantageous mutation.
Selective sweep: synonymous with hitch-hiking.
Substitution: a fixed difference between two species.
Synonymous: a mutation, in a protein-coding sequence, which leaves the amino acid sequence unaltered.

Corresponding author: Eyre-Walker, A. (a.c.eyre-walker@sussex.ac.uk).
Available online 3 July 2006

www.sciencedirect.com 0169-5347/$ – see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.tree.2006.06.015

Box 1. The McDonald–Kreitman test

Derivation

In 1991, McDonald and Kreitman (MK) proposed a test of the neutral theory of molecular evolution, which has become the basis of several methods to estimate the proportion of substitutions that are fixed by positive selection rather than by genetic drift. The test compares the amount of variation within a species to the divergence between species at two types of site, which I assume, for simplicity, are synonymous and non-synonymous sites. Let us assume that all synonymous mutations are neutral and that non-synonymous mutations are strongly deleterious, neutral, or strongly advantageous. Furthermore, let us assume that advantageous mutations are rare.

Under these simplifying assumptions, the expected number of synonymous polymorphisms (P_s) segregating within a collection of n alleles is kL_sN_eu, where N_e is the effective population size, u is nucleotide mutation rate, L_s is the number of synonymous sites and k is a constant that depends upon several factors, including the number of alleles sampled, the sampling strategy and the population history. For example, in a panmictic diploid population of stationary size $k = 4\Sigma 1/i$ for $i = 1$ to $n - 1$. The expected number of non-synonymous polymorphisms (P_n) is kL_nN_euf, where f is the proportion of mutations that are neutral and L_n is the number of non-synonymous sites. The number of synonymous substitutions (D_s) is $2L_sut$, where t is the time since the two species split; the number of non-synonymous substitutions (D_n) is $2L_nutf/(1-\alpha)$, where α is the proportion of non-synonymous substitutions that are adaptive.

Adaptive mutations contribute to D_n but not substantially to P_n because they contribute relatively little to polymorphism, but they can contribute substantially to divergence. For example, if just 1% of mutations, that are either neutral or advantageous, had a selected advantage of $N_es = 100$, they would account for ~2% of the heterozygosity but very nearly 100% of the substitutions.

It is evident that, if $\alpha = 0$, then D_n/D_s is expected to equal P_n/P_s. This is the basis of the MK test. It is also evident that it is possible to estimate the proportion of adaptive substitutions as Equation I:

$$\alpha = 1 - \frac{D_sP_a}{D_nP_s} \qquad \text{[Eqn I]}$$

Several methods have been proposed to estimate the average value of α across genes. These vary from simply summing D_n, D_s, P_n and P_s across genes [4] to a maximum likelihood (ML) method [19]. Recent simulations suggest that the ML method is the best method available, however in practice, the ML and the simple summation methods generally agree [56]. Software to run these methods is available at the author's web-site (http://www.lifesci.sussex.ac.uk/CSE/members/aeyrewalker/aeyrewalker.htm.) Sawyer and colleagues have proposed an alternative parameterisation of the MK test, which assumes weak selection [46]. A recent development of this method enables one to estimate the proportion of substitutions that are adaptive [16].

Example

To illustrate the use of the MK-type data to infer the number of adaptive substitutions, I use data from 115 genes for which there are multiple alleles from *D. simulans* and a single allele from *D. yakuba* [56]. The total number of synonymous and non-synonymous polymorphisms and substitutions are shown in Table I. The number of substitutions has been corrected for multiple hits.

As one can see, $D_n/D_s > P_n/P_s$, which suggests that there has been some adaptive substitution. Application of Equation I suggests that ~49% of the amino acid substitutions between these two species were driven by adaptive evolution, which is similar to the estimate of ~41% from the ML method [56]. To obtain confidence intervals we can bootstrap the data by gene for the summation method (0.35, 0.61) or use standard maximum likelihood techniques (0.33, 0.48).

Table I. MK data from *Drosophila*

	Divergence	Polymorphism
Non-synonymous	3648	439
Synonymous	7365	1741

As Charlesworth [3] first pointed out, it is possible to estimate the proportion of non-synonymous substitutions that have been fixed by adaptive evolution, a statistic I refer to as α, using the data from an MK test in a simple manner. The method, its derivation and an example are described in Box 1.

Although, there are several assumptions behind the MK methods of estimating the level of adaptive evolution, it is generally robust to violations of most of them. The exception is the segregation of slightly deleterious non-synonymous mutations, because these can bias the estimate of α either upwards or downwards depending on the demography of the population. If the population size has been relatively stable, the estimate of α is an underestimate, because slightly deleterious mutations tend to contribute relatively more to polymorphism than they do to divergence when compared with neutral mutations. These slightly deleterious mutations can be controlled for by removing low-frequency polymorphisms from the analysis, because such mutations tend to segregate at lower frequencies than do neutral mutations [3,4]. However, slightly deleterious mutations can lead to an overestimate of α if population sizes have expanded, because mutations that might have been fixed in the past, when the population size was small, no longer segregate as polymorphisms. Even fairly modest increases in population size can create artifactual evidence of adaptive evolution [5].

Although methods based on the MK test are the only methods that can, in principle, yield an unbiased estimate of the amount of evolution that has been driven by positive selection, they are susceptible to demography and require large amounts of data. As a consequence, several other methods have been used to investigate the role of adaptive processes in evolution at the molecular level.

The K_a/K_s *or* d_n/d_s *test*

The first of these alternative methods is the K_a/K_s or d_n/d_s test [6] (Box 2). In this test, the rate of non-synonymous substitution, K_a or d_n, is compared to the rate of silent substitution, K_s or d_s (unfortunately, there are two sets of symbols used to denote substitution rates). If we assume that silent substitutions are neutral, then we can infer that the gene has undergone adaptive evolution if d_n is significantly greater than d_s, because advantageous mutations have a higher probability of spreading through a population than do neutral mutations. In its simplest form, this test is conservative because most non-synonymous mutations are expected to be deleterious; hence d_n tends to be much lower than d_s. Only if there is rampant adaptive evolution will d_n exceed d_s (Box 2). However, there are versions of this test in which the d_n/d_s test is performed for each codon [7,8]. These have not yet been used extensively in genomic analyses because they require data from many species, or from many alleles from within a species. However, they have the potential to be more powerful, and they can yield an estimate of the number of codons that have undergone adaptive evolution.

Box 2. The d_n/d_s and K_a/K_s tests

The d_n/d_s or K_a/K_s test is one of the most frequently used tests of adaptive evolution. It relies on the simple fact that, only if there is adaptive evolution can the rate of evolution exceed that of a neutral sequence. So if we assume for simplicity that synonymous mutations are neutral and non-synonymous mutations are either deleterious, neutral or advantageous then, using the symbols from Box 1, $d_s = 2ut$ and $d_n = 2utf/(1-\alpha)$.

From these expressions, it is evident that d_n can only exceed d_s if $(1-\alpha)<f$; that is, if there is adaptive evolution. Furthermore, it is evident that adaptive evolution needs to be frequent. Values of f are typically <0.3, as judged by average d_n/d_s values, so the proportion of substitutions that are adaptive needs to be >0.7 for adaptive evolution to be detectable.

Genomic surveys of genetic diversity

The third general technique that has been used to quantify adaptive evolution is genomic surveys of genetic diversity. When an advantageous mutation spreads to fixation, it is expected to reduce the genetic diversity in the surrounding genomic region to generate a skew towards rare alleles and to increase linkage disequilibrium [9]. This is a process known as a selective sweep or a genetic hitch-hiking event. If the advantageous mutation only spreads through some populations, then there will be an increase in population differentiation [9]. One could therefore identify regions of the genome that have undergone adaptive evolution by scanning the genome for regions that show one or more of these signatures. Unfortunately, this technique has several limitations.

First, it can only detect very recent adaptive events; typically, only strongly selected [10] mutations that have spread to appreciable frequency within the last 0.1 N_e generations can be detected [11], where N_e is the effective population size. This means, for example, that, in humans, we would only be able to detect a selective sweep that had occurred within the past 25 000 years. Second, genomic scans can only tell us the general location of an adaptive substitution; they do not tell us which mutation or even which gene was responsible for the event. And finally, demography can often mimic the effects of genetic hitch-hiking [12,13]. However, genomic scans have been the subject of great interest because they have the potential to inform us about very recent adaptive evolution, and this is of particular interest for understanding the evolution of our own species.

Quantifying adaptive evolution

Protein-coding sequences: Drosophila

The first attempt to determine the extent of adaptive evolution was made by Brookfield and Sharp [14], who reviewed the results of MK tests in seven *Drosophila* genes; in three of these, there was a significant excess of non-synonymous substitutions, suggesting that adaptive evolution was widespread. Confirming this initial result, Smith and Eyre-Walker [15], using an extension of the MK test, estimated that ~45% of amino acid substitutions in *Drosophila* were driven by positive selection. Although, they examined relatively few genes, more extensive analyses by several authors have confirmed these results (Table 1). The one exception is the analysis by Sawyer *et al.* [16] who used a slightly different parameterisation of the MK test (Box 1) to the previous authors and concluded

Table 1. Estimates of adaptive evolution

Species 1[a]	Species 2[b]	Locus type	Analysis type	Number of loci	α or % of loci adaptively evolving[c]	Refs
Drosophila simulans	*Drosophila yakuba*	Protein	MK	35	45	[15]
		Protein	MK	115	41	[56]
	D. melanogaster	Protein	MK	75	43	[19]
		Protein	MK	56	94	[16]
Drosophila melanogaster	*D. simulans*	Protein	MK	44	45	[19]
		5′ UTR	MK	18	61	[36]
		3′ UTR	MK	13	53	[36]
		Intron	MK	72	19	[36]
		Intergenic	MK	50	15	[36]
Human *Homo sapiens*	Mouse	Protein	MK	330	*0*	[21]
	Old-world monkey	Protein	MK	149	*0*	[21]
		Protein	MK	182/106[d]	35	[4]
	Chimpanzee	Protein	d_n/d_s	8079	0.4	[10]
		Protein	MK	13 500	0–9	[20]
		Protein	MK	289	*20*	[21]
		5′ flank	MK	305	*0.11*[e]	[37]
		3′ flank	MK	305	*0.14*[e]	[37]
		Protein	MK	4916	6	[26]
	Chimpanzee–mouse	Protein	d_n/d_s	7645[f]	0.08	[22]
Escherichia coli	*Salmonella enterica*	Protein	MK	410	>56	[29]
HIV		Protein	MK	1	50	[30]
		Protein	d_n/d_s	1	75[g]	[31]
Influenza		Protein	d_n/d_s	1	85[g]	[31]
Arabidopsis thaliana	*Arabidopsis lyrata*	Protein	MK	12	0	[27]
	A. lyrata	Protein	d_n/d_s	304	5	[28]

[a] Species 1 is the species for which there is polymorphism data for MK-type analyses.

[b] Species 2 is the outgroup or outgroups.

[c] Figures in italics indicate that the estimate was not significantly different from zero.

[d] Numbers of genes differ for divergence (182) and polymorphism (106).

[e] Average of the estimates given for bases 1–500 and 501–1000.

[f] Only the results from Model I are presented.

[g] Proportion of codons showing evidence of adaptive evolution.

that ~94% of amino acid substitutions were driven by positive selection in *Drosophila*, but that these mutations were very weakly selected ($N_e s \approx 5$, where s is the strength of selection). The difference between these analyses probably arises because the method used by Sawyer *et al.* effectively assumes that adaptive mutations are weakly selected, whereas the other methods assume adaptive mutations are strongly selected.

All the analyses discussed above were done using polymorphism data from *Drosophila simulans* and *Drosophila melanogaster*, two species that are believed to have spread out of Africa; thus, it is possible that the current N_e is larger than the ancestral one, and that the estimate of adaptive evolution is therefore an overestimate. This is particularly pertinent in *D. melanogaster* because there is evidence that some non-synonymous mutations are slightly deleterious; non-synonymous polymorphisms segregate, on average, at lower frequencies than do synonymous mutations [17]. However, several lines of evidence suggest that α has not been overestimated because of a population size increase. First, if anything, *D. melanogaster* appears to have gone through a population size decrease [18]. Second, estimates using polymorphism data from either *D. simulans* or *D. melanogaster* are very similar [19]; it is difficult to see how the bias could be the same given that the two species have different N_e [18].

Protein-coding sequences: primates

Whereas the estimates of adaptive evolution in *Drosophila* protein-coding sequences appear to be converging around a value of ~50%, the pattern in the great apes is very different. In their landmark paper, Fay *et al.* [4], the first group to use the MK approach to calculate α, estimated that ~35% of the amino acid differences between humans and old-world monkeys had been driven by adaptive evolution. However, two more recent papers, also using the MK approach, have placed the estimate close to zero within hominids [20,21]. The discrepancy between these studies could be for any one of several reasons, because the analyses involved different species, genes and different sampling strategies of the SNPs.

However, there was one short-coming in the study by Fay *et al.* [4]; because of limited data at the time, the authors had to combine polymorphism data from one set of genes with substitution data from another set. If the genes were a random selection in each case, then this would have been of little consequence, but the polymorphism data came from a set of genes involved in cardiovascular disease, endocrinology and neuropsychiatry. If these genes tended to be more conserved, on average, than the genes used for the substitution data, then you would get artifactual evidence of adaptive evolution.

Analyses using d_n/d_s tests have also failed to find much evidence of widespread adaptive evolution in primates. Clark *et al.* [22] using alignments of human, chimpanzee and mouse, inferred that >11% of genes in hominids showed evidence of adaptive evolution using their model 2 test. However, in this test, it is assumed that the relative branch lengths leading to human, chimpanzee and mouse are the same for all genes [22]; genes with an excessively long human branch can therefore appear to be subject to adaptive evolution. But an excessively long human branch could also be a consequence of relaxed natural selection, rather than of adaptive evolution. In fact, relaxed constraint is consistent with one of their more puzzling results; they inferred that a large proportion of olfactory receptor genes were undergoing adaptive evolution, but these genes appear to be subject to little natural selection in humans, because many of the copies have become pseudogenes [23,24]. Using a more conventional d_n/d_s test, the authors found that only 0.08% of genes showed evidence of adaptive evolution. This is similar to the proportion of genes that Nielsen *et al.* [25] detected as significant with a d_n/d_s test using just humans and chimpanzees (0.4%).

By contrast, Bustamante *et al.* [26], using a multi-locus MK test, detected adaptive evolution in ~6% of genes. This proportion is somewhat higher than that detected by Clark *et al.* [22] and Nielsen *et al.* [25], which could be due to the increased power that the MK test can have over the d_n/d_s test. However, the proportion is still fairly low. Thus, on balance, there appears to relatively little evidence of widespread adaptive evolution in hominids, although we have clearly adapted.

Protein-coding sequences: other species

Although *Drosophila* and hominids have been the focus of most attempts to infer overall levels of adaptive evolution, estimates are starting to be produced for a range of other organisms. Bustamante *et al.* [27] applied a multi-locus variant of the MK test to 12 genes from *Arabidopsis thaliana* and could find no evidence of adaptive evolution. However, Barrier *et al.* [28] found that $d_n > d_s$ in ~5% of the 304 *A. thaliana* genes that they surveyed, suggesting that at least some genes were undergoing adaptive evolution.

The level of adaptive evolution has also been estimated in some viruses and bacteria. Using 410 genes from six *Escherichia coli* and six *Salmonella enterica* strains, Charlesworth and Eyre-Walker [29] estimated, using an MK-type analysis, that at least 56% of amino acid substitutions had been fixed by adaptive evolution. For viruses, Williamson [30], also using an MK type analysis, estimated that ~50% of the substitutions within the *env* gene of HIV-1 had been driven by positive selection within a patient, whereas Nielsen and Yang [31] estimated the proportion to be slightly higher, at 75%, by considering the distribution of d_n/d_s across codons. Nielsen and Yang [31] also estimated that ~85% of amino acid substitutions in the hemagluttinin gene of the human influenza virus had been driven by adaptive evolution. Thus, estimates of adaptive evolution appear to be high in viruses and bacteria, although the genes considered in the viruses are involved in the interaction between the virus and the immune system. They might therefore be particularly prone to adaptive evolution.

Non-coding DNA: *Drosophila*

Whereas much work has focussed on the evolution of protein-coding sequences, relatively little attention has been paid to adaptive evolution in the regulation of gene control. Prior to 2004, adaptive evolution had only been tested for in the regulatory elements of just three genes [32–34] compared with the many-tens of studies in which it had been investigated in protein-coding sequences. This started to change when Kohn *et al.* [35] published an analysis of adaptive evolution in the 5′ region of eight genes in *Drosophila*; using the MK approach, the authors estimated that ∼50% of all substitutions in the 700 bp upstream of the genes had been fixed by positive selection. This result has been dramatically confirmed by Andolfatto [36] who not only estimated that ∼60% of substitutions in the untranslated regions (UTR) have been fixed by positive selection, but that ∼20% of all substitutions within introns had also been the subject of adaptive evolution and that a similar proportion might have been affected in intergenic DNA. If Andolfatto is correct, his results radically change our perspective on both the nature of non-coding DNA and adaptive evolution.

Non-coding DNA: hominids

The situation appears to be different in hominids. Using an MK-type analysis, Keightley *et al.* [37] found no evidence of adaptive evolution either upstream or downstream of protein-coding genes. This might reflect a genuine difference in the level of adaptive evolution between drosophilids and hominids, as appears to be the case for protein-coding sequences, a lack of power in hominids owing to low polymorphism levels or the fact that gene control elements might be more dispersed in the larger hominid genome.

Genomic scans

A few years ago, there was great optimism that, by looking for the signature of selective sweeps, it might be possible to use genomic scans of DNA diversity to identify regions of the genome that had recently undergone adaptive evolution. Unfortunately, this program of research has not proved as fruitful as had been hoped. Recent large surveys in humans [10,38–40], *Drosophila* [13,41], *Plasmodium* [42] and mice [43] have yielded just a handful of potential adaptive events. The reason that genomic scans have generally failed to find evidence of adaptive evolution is probably because ’traditional’ genomic scans are not very powerful and adaptive evolution might not be either frequent enough or strong enough to generate a signal. In these respects, it is interesting that only two studies in which widespread evidence of adaptive evolution has been detected were an analysis of adaptive evolution during the domestication of maize (*Zea mays* spp. *mays*) [44], in which selection was probably very strong, and an analysis of partial selective sweeps in humans [45]. In this latter analysis, Voight *et al.* used a method based on linkage disequilibrium. It might be that linkage disequilibrium methods are more powerful at detecting adaptive evolution than are methods based on the level or skew of the genetic diversity, although they are only likely to be good at detecting partial selective sweeps. Unfortunately, Voight *et al.* [45] did not quantify the number of loci that had undergone a partial selective sweep in humans, but it appeared to be substantial.

Inferring the rate and strength of adaptive evolution

Genomic rates

An estimate of the proportion of substitutions that are adaptive enables one to estimate the genomic rate of adaptation at the molecular level. If 45% of amino acid substitutions are adaptive, then *Drosophila* species undergo approximately one substitution every 45 years or 450 generations [15]. This means that they undergo ∼22 000 amino acid substitutions per million years. The estimates from the UTR and intron regions of *Drosophila* are somewhat higher; Andolfatto [36] estimates that there has been an adaptive substitution in a UTR intron every 26 years and in intron regions perhaps every seven years. The sampling of intergenic DNA is not sufficient to draw any firm conclusions, but the combined rate of adaptive evolution might approach one substitution every few years. *Drosophila* is currently the only species for which we can make these calculations, because it is the only species for which we have all the necessary information.

The strength of selection

In principle, it is possible to estimate not only the number of adaptive substitutions that have occurred in the evolution of a species, but also the average strength of selection that has acted upon them. There are two approaches to this.

The first approach is to assume that adaptive mutations are weakly selected; under this assumption, one can fit a weak selection model to MK-type data. In the simplest case, all non-synonymous mutations that contribute to polymorphism and divergence are assumed to be subject to the same strength of selection, which can then be estimated [26,27,46]; this model has been extended to enable the strength of selection to be drawn from a normal distribution [16]. In both cases, the strength of selection in favour of advantageous mutations is estimated to be such that $1 < N_e s < 10$ [16,26,27].

The second method is based on the observed correlation between levels of nucleotide diversity and the recombination rate, which exists in several species, including *Drosophila* [13,47]. This correlation has two explanations; first it could be due to adaptive evolution and genetic hitch-hiking [48]. Second, it could be due to the removal of deleterious mutations through a process called background selection [49]. Which of these is the cause still remains unresolved. However, if we assume that the correlation is caused by genetic hitch-hiking, then we can infer the strength of selection if we know the number of adaptive substitutions [50,51]. Stephan [50] has estimated that, to explain the correlation between diversity and recombination rate in *Drosophila,* the product $v\gamma$ (where v is the rate of adaptive evolution per nucleotide per generation and $\gamma = 2N_e s$, s being the strength of selection in favour of the mutation) must be between 10^{-7} and 10^{-8}. Andolfatto [36] gives estimates of v. He estimates that *D. melanogaster* and *D. simulans* differ from one another by ∼1.3 million adaptive changes, even if we ignore adaptive substitutions in intergenic regions, for which the estimates are currently not very reliable. This equates to 1.4×10^{-10} adaptive substitutions per nucleotide per generation, which implies that $350 < N_e s < 3500$. Unfortunately, it is currently not clear how much intergenic DNA is under selection and how much adaptive evolution occurs in it; at most, the number of adaptive substitutions will be doubled and the strength of selection halved.

It is evident that these two methods of estimating the average strength of selection acting upon advantageous mutations give very different estimates. Both methods make several questionable assumptions; the first method effectively assumes that advantageous mutations are weakly selected, whereas the second method assumes that the correlation between diversity and recombination rate is solely due to genetic hitch-hiking. However, it is difficult to reconcile the estimates from the first approach with ecological estimates of selection. For example, Hoekstra *et al.* [52] estimate that the strength of selection acting upon coat colour in pocket mice is between 0.5% and 5%, with the coat colour in this population being determined by a single gene [53]. Given that mice probably have N_e of at least 100 000, [54] this means that N_es is at least 500, in keeping with the estimates from the second method but not the first. Similarly, a recent analysis of positive selection on a non-sense mutation in the Caspase-12 gene in humans gives $N_es \approx 100$ [55]. However, further work is clearly needed.

Future questions

A correlation with population size?

Although the data are limited, there does appear to be a possible correlation between the level of adaptive evolution and population size: hominids appear to have undergone very little adaptive evolution, compared with *Drosophila*, bacteria and viruses. This is to be expected given that large populations generate more mutations and selection is effective on a greater proportion of mutations. This might mean that species with small population sizes are much less able to adapt to their environment. Or it might be that most adaptation precedes by mutations with strong effects, which can be fixed in species of all population sizes, and that many of the adaptive mutations fixed in species with large population sizes are merely fine-tuning.

What do all these substitutions do?

Drosophila melanogaster and *D. simulans* are estimated to differ from one another by at least 1.3 million adaptive differences [36]; even if we focus on the protein-coding complement of the genome they appear to have ~110 000 adaptive amino acid differences (calculated from [15]). And yet these species are almost identical morphologically. What do all these adaptive differences do? It might be that many of them are involved in the physiology and ecology of these species, something that we know remarkably little about. And it might be that some of the adaptive substitutions are a consequence of arms races between host and parasite. But it is still difficult to comprehend how so much adaptive evolution can be going on. It might be that we just have no idea how complex the environment really is and how it is constantly changing in ways that always challenge an organism to adapt.

Conclusion

For over 30 years, scientists have debated, often bitterly, the relative contributions of genetic drift and adaptation to evolution at the molecular level. Although we are still some way from resolving the controversy, it is clear that we should know the answer within the next few years, at least in terms of the differences that differentiate species. What proportion of the variation within a species is maintained by balancing selection is still far from clear and would appear to be a much harder problem to solve.

Acknowledgements

I am grateful to Peter Andolfatto, Jane Charlesworth, John Welch and four anonymous referees for comments, and to the National Evolutionary Synthesis Center and the BBSRC for funding.

References

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

2 Hudson, R.R. *et al.* (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116, 153–159

3 Charlesworth, B. (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.* 63, 213–227

4 Fay, J. *et al.* (2001) Positive and negative selection on the human genome. *Genetics* 158, 1227–1234

5 Eyre-Walker, A. (2002) Changing effective population size and the McDonald–Kreitman test. *Genetics* 162, 2017–2024

6 Yang, Z. and Bielawski, J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503

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

8 Suzuki, Y. and Gojobori, T. (1999) A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 16, 1315–1328

9 Schlotterer, C. (2003) Hitch-hiking mapping - functional genomics from the population perspective. *Trends Genet.* 19, 32–38

10 Nielsen, R. *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Res.* 15, 1566–1575

11 Kim, Y. and Stephan, W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160, 765–777

12 Haddrill, P.R. *et al.* (2005) Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res.* 15, 790–799

13 Ometto, L. *et al.* (2005) Inferring the effects of demography and selection on *Drosophila melanogaster* populations from chromosome-wide scan of DNA variation. *Mol. Biol. Evol.* 22, 2119–2130

14 Brookfield, J.F. and Sharp, P.M. (1994) Neutralism and selectionism face up to DNA data. *Trends Genet.* 10, 109–111

15 Smith, N.G.C. and Eyre-Walker, A. (2002) Adaptive protein evolution in *Drosophila*. *Nature* 415, 1022–1024

16 Sawyer, S. *et al.* (2003) Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* 57, S154–S164

17 Fay, J. *et al.* (2002) Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* 415, 1024–1026

18 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

19 Bierne, N. and Eyre-Walker, A. (2004) Genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol. Biol. Evol.* 21, 1350–1360

20 Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87

21 Zhang, L. and Li, W.-H. (2005) Human SNPs reveal no evidence of frequent positive selection. *Mol. Biol. Evol.* 22, 2504–2507

22 Clark, A.G. *et al.* (2003) Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. *Science* 302, 1960–1963

23 Gilad, Y. *et al.* (2005) A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res.* 15, 224–230

24 Gilad, Y. *et al.* (2003) Human specific loss of olfactory receptor genes. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3324–3327

25 Nielsen, R. *et al.* (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 3, e170

26 Bustamante, C.D. *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature* 437, 1153–1157

27 Bustamante, C.D. *et al.* (2002) The cost of inbreeding in *Arabidopsis*. *Nature* 416, 531–534

28 Barrier, M. *et al.* (2003) Selection on rapidly evolving proteins in the *Arabidopsis*. *Genetics* 163, 723–733

29 Charlesworth, J. and Eyre-Walker, A. The rate of adaptive evolution in enteric bacteria. *Mol. Biol. Evol.* (in press)

30 Williamson, S. (2003) Adaptation in the *env* gene of HIV-1 and evolutionary theories of disease progression. *Mol. Biol. Evol.* 20, 1318–1325

31 Nielsen, R. and Yang, Z. (2003) Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol. Biol. Evol.* 20, 1231–1239

32 Jenkins, D.L. *et al.* (1995) A test for adaptive change in DNA sequences controlling transcription. *Proc. Biol. Sci.* 261, 203–207

33 Ludwig, M.Z. and Kreitman, M. (1995) Evolutionary dynamics of the enhancer region of *even-skipped* in *Drosophila*. *Mol. Biol. Evol.* 12, 1002–1011

34 Tautz, D. and Nigro, L. (1998) Microevolutionary divergence pattern of the segmentation gene *hunchback* in *Drosophila*. *Mol. Biol. Evol.* 15, 1403–1411

35 Kohn, M.H. *et al.* (2003) Inference of positive and negative selection on the 5′ regulatory regions of *Drosophila* genes. *Mol. Biol. Evol.* 21, 374–383

36 Andolfatto, P. (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437, 1149–1152

37 Keightley, P.D. *et al.* (2005) Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* 3, e42

38 Akey, J.M. *et al.* (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2, e286

39 Payseur, B.A. *et al.* (2002) Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol.* 19, 1143–1153

40 Storz, J.F. *et al.* (2004) Genome scans of DNA variability in humans reveal evidence for selective sweeps outside Africa. *Mol. Biol. Evol.* 21, 1800–1811

41 Kauer, M.O. *et al.* (2003) A microsatellite variability screen for positive selection associated with the ‘out of Africa’ habitat expansion of *Drosophila melanogaster*. *Genetics* 165, 1137–1148

42 Wootton, J.C. *et al.* (2002) Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* 418, 320–323

43 Ihle, S. *et al.* (2006) An analysis of signatures of selective sweeps in natural populations of the house mouse. *Mol. Biol. Evol.* 23, 790–797

44 Wright, S.I. *et al.* (2005) The effects of artificial selection on the maize genome. *Science* 308, 1310–1314

45 Voight, B.F. *et al.* (2006) A map of recent positive selection in the human genome. *PLoS Biol.* 4, e72

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

47 Begun, D.J. and Aquadro, C.F. (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356, 519–520

48 Smith, J.M. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35

49 Charlesworth, B. *et al.* (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134, 1289–1303

50 Stephan, W. (1995) An improved method for estimating the rate of fixation of favourable mutations based on DNA polymorphism data. *Mol. Biol. Evol.* 12, 959–962

51 Wiehe, T. and Stephan, W. (1993) Analysis of a genetic hitch-hiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol. Biol. Evol.* 10, 842–854

52 Hoekstra, H.E. *et al.* (2004) Ecological genetics of adaptive color polymorphism in pocket mice: geographic variation in selected and neutral genes. *Evolution* 58, 1329–1341

53 Nachman, M.W. *et al.* (2003) The genetic basis of adaptive melanism in pocket mice. *Proc. Natl. Acad. Sci. U. S. A.* 100, 5268–5273

54 Eyre-Walker, A. *et al.* (2002) Quantifying the slightly deleterious model of molecular evolution. *Mol. Biol. Evol.* 19, 2142–2149

55 Xue, Y. *et al.* (2006) Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am. J. Hum. Genet.* 78, 659–670

56 Welch, J.J. Estimating the genome-wide rate of adaptive protein evolution in *Drosophila*. *Genetics* (in press)