The Rate of Adaptive Evolution in Enteric Bacteria

Jane Charlesworth* and Adam Eyre-Walker*†

*Centre for Study of Evolution, School of Life Sciences, University of Sussex, Brighton, United Kingdom; and
†National Evolutionary Synthesis Center, Durham, USA

Here we estimate the rate of adaptive substitution in a set of 410 genes that are present in 6 Escherichia coli and 6 Salmonella enterica genomes. We estimate that more than 50% of amino acid substitutions in this set of genes have been fixed by positive selection between the E. coli and S. enterica lineages. We also show that the proportion of adaptive substitutions is uncorrelated with the rate of amino acid substitution or gene function but that it may be correlated with levels of synonymous codon usage bias.

Introduction

The rate at which adaptive substitutions occur has long been an interesting evolutionary question (Nei 2005) and one which data emerging from genome sequences promise to shed some light on. Many recent studies have applied a diverse set of methods (several of which are reviewed by Nielsen [2001]) to data from a number of species, attempting to quantify the importance of adaptive substitution in evolution or to identify regions of genomes whose variation has been shaped by positive selection. The first large-scale analysis of adaptive evolution was carried out by Endo et al. (1996) who compared nonsynonymous and synonymous substitution rates in 3,595 genes from a range of taxa. They found evidence of adaptive evolution in only 17 of these genes. However, such an approach is most powerful when the taxa are closely related; a recent analysis of genes in humans and chimpanzees has suggested that as many as ~11% of genes show evidence of adaptive evolution (Clark et al. 2003). Genomic scans of DNA diversity have also suggested that a number of genes have undergone adaptive evolution in humans (Ronald and Akey 2005), Drosophila (Glinka et al. 2003; Orengo and Aguade 2004), and maize (Wright et al. 2005). The signature of adaptive evolution is most conspicuous in populations that have experienced a novel habitat—in both humans (Akey et al. 2004) and Drosophila (Glinka et al. 2003), adaptive evolution appears to have been more prevalent in European than in African populations. Similarly, it has been estimated that 2–4% of maize genes have undergone adaptive evolution during domestication (Wright et al. 2005).

Although the above studies suggest that adaptive evolution is reasonably common, they do not give us an estimate of the proportion of evolutionary change that is due to positive selection. However, this quantity can be estimated using an extension of the test of McDonald and Kreitman (1991) that was first suggested by Charlesworth (1994). By combining data from both between and within species, Fay et al. (2001) estimated that ~35% of amino acid substitutions were due to adaptive evolution in primates, but recent studies have found little evidence of positive selection in hominids (Chimpanzee Sequencing and Analysis Consortium 2005; Zhang and Li 2005). In contrast, estimates of the rate of adaptive substitution in Drosophila have been consistently high. However, the precise estimate depends upon the methods used. Using approaches similar to those employed in the analysis of primate data, the proportion of amino acid substitutions that are adaptive (α) is estimated to be between 25% and 45% in Drosophila (Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004). This rises to over 90% under the method used by Sawyer et al. (2003).

The fact that some genes show evidence of adaptive evolution whereas others do not suggests that adaptive evolution may be restricted or more prevalent in some genes. This is certainly what one might expect; for example, genes involved in the immune system would be expected to undergo more adaptive evolution than housekeeping genes. However, the only analysis to formally test whether the proportion of substitutions driven by positive selection varies between genes found no evidence of variation in a set of Drosophila simulans genes (Bierne and Eyre-Walker 2004) (though see Fay et al. 2002, for contrary evidence using a less formal approach). This was striking given that the rate of nonsynonymous substitution varied by several orders of magnitude in the genes studied. However, it is unclear whether this apparent constancy was real or due to a lack of power in the method.

Here we attempt to estimate the rate of adaptive amino acid substitution in enteric bacteria. We also investigate whether the proportion of adaptive substitutions varies between genes. We do this using the method of Bierne and Eyre-Walker (2004)—that is, by comparing models in which the proportion of adaptive substitutions is allowed to vary between genes, relative to a model in which all genes share a single proportion of adaptive substitutions. However, we also investigate whether the proportion of adaptive substitutions is correlated with the rate of amino acid substitution, gene function, or codon bias. This latter analysis was motivated by the fact that there is a strong negative correlation between the nonsynonymous substitution rate and codon bias in several taxa (Sharp 1991; Akashi 1994; Pal et al. 2001; Betancourt and Presgraves 2002; Rocha and Danchin 2004) and by the suggestion that one might be able to estimate rates of adaptive evolution using measures of codon usage bias (Plotkin et al. 2004; Stoeckli et al. 2005). We investigate these questions using a large set of genes sampled from several complete-genome sequences of both E. coli and S. enterica.
Materials and Methods

Data

Genes were extracted from 6 complete E. coli genomes: E. coli K12, O157:H7, O157:H7 EDL933, CFT073 (GenBank accession numbers: U00096, BA000007, AE006174, AE014075, respectively), and E2348 (these sequence data were produced by the E. coli and Shigella Comparative Sequencing Group at the Sanger Institute and can be obtained from ftp://ftp.sanger.ac.uk/pub/pathogens/Escerichia_Shiella) and 6 complete S. enterica genomes: S. enterica LT2, Choleraesuis, Paratyphi A, Typhi CT18, and Typhi Ty2 (GenBank accession numbers: AE006468, AE017220, CP000026, AL513382, AE014613, respectively), PT4, Salmonella typhimurium DT104 and S. typhimurium SL1344 (these sequence data were produced by the Salmonella Comparative Sequencing Group at the Sanger Institute and can be obtained from ftp://ftp.sanger.ac.uk/pub/pathogens/Salmonella). Note that S. typhimurium is a strain of S. enterica. Sequences were downloaded from GenBank http://www.ncbi.nlm.nih.gov/ genomes/lproks.cgi and from the Sanger ftp sites as cited.

All protein-coding sequences longer than 100 bp were extracted from the E. coli K12 and S. typhimurium LT2 genomes, and Stand Alone Blast (Altschul et al. 1997) was used to query each gene against databases containing the remaining available genomes for each species plus the outgroup genome sequence (S. typhimurium LT2 for E. coli and E. coli K12 for S. enterica). An alignment for a gene was accepted if 85% of the coding sequence was present in all strains, and no premature stop codons were detected. Genes which were present in all strains but incomplete were aligned by hand using SeAl v. 2.0a11 Carbon (Rambaut 1996). In a large number of cases, variants were found in a single sequence that appeared to be single base pair insertions or deletions. It was not possible to tell whether these were real indels or sequencing errors. In order to be certain that no genuine pseudogenes were accidentally included in the analysis, any genes containing such single base pair indels were excluded from the final data set; in addition, our analysis excluded singleton polymorphisms (i.e., polymorphisms present in a single sequence) to reduce the effect of sequencing errors. Our final data set contained 410 genes present in all 6 E. coli and all 6 S. enterica genomes.

Analysis

The relationship between strains was examined by concatenating 200 randomly chosen genes and constructing a Neighbor-Joining tree using synonymous divergence between strains as estimated by the Nei–Gojobori method I (Nei and Gojobori 1986). This analysis was performed in Mega3 (Kumar et al. 2004). We also estimated overall synonymous diversity between the strains using Watterson’s θ (Watterson 1975) (θs). θs was calculated for all 410 genes and averaged across genes.

Polymorphisms were counted using software written for this purpose. Counts were identical to those produced by DnaSP (Rozas J and Rozas R 1999), except where we resolved codons containing multiple mutations, which DnaSP omits. We adopted the following algorithm. For all pairs of codons present at a site, we estimated the number of synonymous and nonsynonymous differences between them using the method of Nei and Gojobori (1986)—that is, we took an unweighted average of the paths. We identified for each codon the codon it was most closely related to, and then we summed the numbers of nonsynonymous and synonymous polymorphisms across these pairs of codons. This seems a reasonable approach for species, such as E. coli and S. enterica, in which recombination is sufficiently rare that it will not often introduce a recombination breakpoint within a codon (Maynard Smith et al. 1993). As we had no information about which sequences were ancestral, a polymorphism segregating at a frequency of i out of n sequences was indistinguishable from a polymorphism segregating in n–i sequences. Numbers of nonsynonymous and synonymous polymorphisms were denoted by Pn and Ps, respectively. Polymorphism frequencies were defined as singletons (a single polymorphic sequence), doubletons (2 strains have a variant), and tripletons (3 strains have a variant).

The numbers of substitutions between E. coli and S. enterica were estimated using the sequences from the 2 strains E. coli K12 and S. typhimurium LT2; substitutions were counted using the “Codeml” program in the PAML package (Yang 2002), with the F3x4 model of codon frequencies. We follow Dunn et al. (2001) and denote the number of nonsynonymous and synonymous substitutions by Dn and Ds, respectively, with the number of substitutions per site denoted by dN and dS. We restricted the estimate of divergence to those sites for which we had polymorphism data. Since this varies slightly between E. coli and S. enterica (i.e., a gene may be present in both species, but the exact length of alignable sequence might differ between species), comparisons in the 2 directions between the species yielded slightly different estimates for the numbers of synonymous and nonsynonymous substitutions (Dn and Ds) for some genes.

We tested for recombination within the E. coli and S. enterica strains separately using Maynard Smith’s “maxchi” test (Maynard Smith 1992). We ran the test with the adjustment suggested by Piganeau and Eyre-Walker (2004), in which chi-square values were ignored if they were generated by contingency tables with any expected value less than 2. This prevents very high chi-square values being produced when there is little data. Maynard Smith’s test was originally designed to test for recombination in bacteria and has proved to be one of the most powerful tests of recombination in general (Posada and Crandall 2001). The maxchi test will detect both recombinations between strains within the sample being considered and between strains from different species. We differentiated these 2 possibilities by testing whether the level of polymorphism was different either side of the breakpoint; we tested this using a chi-square test of independence where the cell entries in the 2 × 2 table are number of polymorphic and nonpolymorphic sites before the breakpoint and number of polymorphic and nonpolymorphic sites after the breakpoint. The breakpoint was considered to be midway between the single-nucleotide polymorphisms (SNPs) flanking the breakpoint. We considered it likely that the
recombinant was from outside the set of strains if the chi square was significant at the 5% level.

The proportion of amino acid substitutions due to positive selection, \( \alpha \), was calculated using the maximum likelihood method of Bierne and Eyre-Walker (2004). We investigated whether there was evidence for variation in \( \alpha \) using 2 approaches. First, we tested for correlations between \( \alpha \) and the rate of amino acid evolution (\( dN \)), codon bias, or functional category. Second, we applied the method of Bierne and Eyre-Walker (2004), using software written by Welch (2006), to investigate whether models which allowed \( \alpha \) to vary fit the data better than models that assumed a constant value of \( \alpha \) for all genes. To investigate the relationship between \( dN \) and \( \alpha \), we split each gene into 2 equal-sized sets of mutually exclusive codons by randomly selecting codons without replacement. One set was used to estimate \( dN \), whereas the other set was used to estimate \( Dn \) and \( \alpha \). By using different sets of codons to measure \( dN \) and \( \alpha \), we ensured that \( \alpha \) and \( dN \) were statistically independent and that any correlation between the 2 was not due to using the same sites to measure \( dN \) and the \( Dn \) count used to calculate \( \alpha \). To reduce sampling error, we divided the data set into 9 groups in order of increasing \( dN \), each containing an approximately equal number of genes. Within each group, we summed \( Dn \), \( Ds \), \( Pn \), and \( Ps \) and used these summed counts to calculate \( \alpha \) using the second set of codons. \( \alpha \) was calculated by the simple equation:

\[
\alpha = 1 - \frac{\sum (Pn) \sum (Ds)}{\left( \sum (Ps) \right) \left( \sum (Dn) \right)},
\]

as used by (Fay et al. 2001). An average \( dN \) for each group was calculated using the first set of \( dN \) values. We also divided \( dN \) into 2 parts, that which appeared to be due to adaptive evolution (\( dN_{\text{adaptive}} \)) and that which appeared to be neutral (\( dN_{\text{neutral}} \)). Simple rearrangements of equation (1) and division by the number of nonsynonymous sites (\( L_n \)) reveal that these are calculated by:

\[
dN_{\text{adaptive}} = \alpha \frac{\sum (Dn)}{\sum (L_n)}, \tag{2a}
\]

\[
dN_{\text{neutral}} = (1 - \alpha) \frac{\sum (Dn)}{\sum (L_n)}. \tag{2b}
\]

We performed a similar analysis to test for a correlation between codon usage bias and rates of adaptive evolution. Codon usage bias was measured using the codon adaptation index (CAI) (Sharp and Li 1987). Optimal codons were assumed to be the same in both species, and we used the codon fitness values defined for a set of \( E. coli \) genes by Sharp and Li (1987). Genes were divided into 17 groups of approximately equal size by ascending CAI values; \( \alpha \) was calculated for each group using the summation method given above.

To investigate the relationship between \( \alpha \) and gene function, we divided our genes into functional categories by searching the database of information on \( E. coli \) genes available at http://coli.berkeley.edu/cgi-bin/ecoli coli_entry.pl. We used the functional categories defined by Riley and Labedan (1996). Genes were sorted by functional category, and categories containing 20 or more genes were investigated further. In order to control for the correlation between \( \alpha \) and codon bias, we first looked at the mean CAI of each group; analysis of variance (ANOVA) was used to test whether any group had a significantly different CAI. To test for variation between \( \alpha \) among genes belonging to different functional groups, we compared the variance in \( \alpha \) between different groups of genes with the variance obtained by randomly permuting genes between groups, preserving the same number of genes in each group as in the real functional categories.

**Results**

**Data**

We extracted all the protein-coding sequences which were readily identifiable as being common to 6 \( E. coli \) and 6 \( S. enterica \) genomes (i.e., present in all 12 genomes under consideration). This yielded 410 genes, about one-tenth of the genes present in \( E. coli \) K12. The number of genes in our data set is much smaller than the total number of genes in any individual microbial genome, for several reasons. First, many genes are species specific: the fraction of \( S. enterica \) LT2 genes having a homologue in 8 other enterobacterial genomes (including several \( E. coli \) and \( S. enterica \) strains) is estimated to be only \( \sim 55\% \) (McClelland et al. 2001). Second, many genes are not present in all strains of a species—a comparison of 3 \( E. coli \) genomes showed that they only shared 39\% of their gene content (Welch et al. 2002). Third, we excluded genes containing either an indel which changed the frame or a premature stop codon. Some of these indels and stop codons are probably sequencing errors, but they could also indicate null alleles or pseudogenes. And finally, our data set may be biased toward genes which are relatively conserved due to the automated way in which we compiled the data.

**Species**

The McDonald–Kreitman framework relies on the distinction between inter- and intraspecies processes because neutral and advantageous mutations behave differently in these 2 settings—advantageous mutations contribute relatively more to interspecies differences (i.e., substitutions) than they do to intraspecies differences (i.e., polymorphism) when compared to neutral mutations. However, species boundaries are notoriously difficult to infer in prokaryotes, although species, in a population genetic sense, do exist—that is, there are collections of strains which undergo genetic drift together (Hey 2001). To proceed, we took a pragmatic approach to the analysis and determined levels of synonymous divergence between our strains by concatenating 200 randomly selected genes and constructing a phylogenetic tree. The strains are closely related to each other within both \( E. coli \) and \( S. enterica \); the maximum synonymous divergence between any 2 strains is less than 10\% within both species (fig. 1). Furthermore, the average values of Watterson’s \( \theta \) across all genes are 0.05 and 0.02 for \( S. enterica \) and \( E. coli \), respectively, whereas the average synonymous divergence as measured by \( dS \) is 0.76. The values of Watterson’s \( \theta \) are very similar to those estimated in several \( Drosophila \) species (Moriyama and Powell 1997; Andolfatto 2001). It is therefore plausible, though by no
means proved, that the *E. coli* and *S. enterica* strains used here act as distinct species in the population genetic sense.

To further investigate whether the strains were part of a single species, we tested for recombination between them. In both *E. coli* and *S. enterica*, we observed substantial numbers of genes that had apparently undergone recombination (104/334 = 31% of genes in *E. coli* and 64/276 = 23% of genes in *S. enterica* were significant at the 5% level using the maxchi test) (table 1). In 85% of *E. coli* genes and 82% of *S. enterica* genes, in which we had evidence of recombination, there was no evidence of an increase in the diversity in the recombinant region, suggesting that recombination is occurring between strains within a species and not between species. Almost all strains in both *E. coli* and *S. enterica* appear to have exchanged genetic material; however, a couple of pairs appear to have undergone more recombination than others—for example, *E. coli* strains K12 and 0157:H7 (table 1). The presence of recombination does not prove that the strains form species at all loci, but it does suggest that the strains are acting as a single unit. It should also be appreciated that we will underestimate the level of adaptive evolution if some of the strains are actually separate species; this is because we will count some substitutions as polymorphisms.

**Slightly Deleterious Mutations**

Many nonsynonymous SNPs appear to be slightly deleterious in bacterial species because they segregate at lower frequencies than synonymous SNPs (Hughes 2005). Such slightly deleterious SNPs can lead to an underestimate of the rate of adaptive evolution because they contribute to polymorphism, but rarely become fixed (Fay et al. 2001). In our data, nonsynonymous polymorphisms segregate at significantly lower frequencies than synonymous polymorphisms, in both *E. coli* and *S. enterica* (table 2), which suggests that some nonsynonymous mutations are slightly deleterious. This difference in frequency is greater in *S. enterica*, suggesting that slightly deleterious mutations are more prevalent in this species than in *E. coli*. A similar pattern might be caused by sequencing errors, but the average allelic frequency of nonsynonymous mutations is significantly less than that for synonymous mutations even if we remove singletons.

**Estimating \( \alpha \)**

It is possible to partially remove the effects of these slightly deleterious mutations by removing mutations segregating at low frequencies (Fay et al. 2001). If we calculate \( \alpha \) excluding low-frequency variants, we find that \( \alpha \) increases as the frequency of the polymorphisms being considered increases. Unfortunately, there is no tendency for the \( \alpha \) value to approach an asymptote as the frequency of the mutations being considered increases (table 3). This is not unexpected when the number of alleles sampled is small. Nevertheless, it makes it difficult to estimate of the rate of adaptive evolution. The best we can do is to estimate a lower bound for the rate of adaptive evolution.

---

**Table 1**

The Number of Genes Showing Significant Evidence of Recombination \( (P > 0.05) \) for Each Pair of Strains in (a) *Escherichia coli* and (b) *Salmonella enterica*

<table>
<thead>
<tr>
<th>(a) <em>Escherichia coli</em></th>
<th>K12</th>
<th>0157:H7</th>
<th>0157:H7 EDL933</th>
<th>E2348</th>
<th>042</th>
<th>CFT073</th>
</tr>
</thead>
<tbody>
<tr>
<td>K12</td>
<td>N/A</td>
<td>19</td>
<td>1</td>
<td>9</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>0157:H7</td>
<td>N/A</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0157:H7 EDL933</td>
<td>N/A</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E2348</td>
<td>N/A</td>
<td>6</td>
<td></td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>042</td>
<td>N/A</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFT073</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b) <em>Salmonella enterica</em></th>
<th>LT2</th>
<th>Paratyphi A</th>
<th>PT4</th>
<th>Ty2</th>
<th>CT18</th>
<th>Choleraesuis</th>
</tr>
</thead>
<tbody>
<tr>
<td>LT2</td>
<td>N/A</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Paratyphi A</td>
<td>N/A</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>PT4</td>
<td>N/A</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ty2</td>
<td>N/A</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT18</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choleraesuis</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.**—N/A: not applicable. We only tabulate the most significant recombination event for each gene and only recombination events that are likely to have been within the group of strains being considered.
using SNPs from the highest frequency class, which in this case is tripletons. Using polymorphism data from E. coli, this is 56%, with a lower confidence interval (CI) of 45%. Using polymorphism data from S. enterica, it is 34%, with a lower CI of 14%. Because slightly deleterious nonsynonymous mutations seem to be more prevalent in S. enterica than in E. coli, we take the estimate of $\alpha$ using the E. coli polymorphism data as our best minimum estimate.

### Variation in $\alpha$ Across the Genome

It has been suggested that adaptive evolution may be more prevalent in fast evolving genes (Fay et al. 2002). However, the method of Biener and Eyre-Walker (2004) and further investigations (Welch 2006) find strong support for models in which a single estimate of $\alpha$ is shared across genes, rather than models where $\alpha$ is allowed to vary according to a specified distribution. Here, we investigated variation in $\alpha$ by testing whether $\alpha$ was correlated with a number of parameters and by model comparisons. To test whether $\alpha$ correlates with the rate of evolution, as suggested by Fay et al. (2002), we divided each gene into 2 sets of mutually exclusive randomly chosen codons and measured the nonsynonymous divergence ($dN$) for each set. The 2 estimates of $dN$ are highly correlated (Spearman’s $\rho = 0.618$). One set of codons was used to estimate $\alpha$, $dN_{\text{adaptive}}$ and $dN_{\text{neutral}}$, and the other set was used to estimate $dN$; in this way we made our measures of $dN$ and our measures of $\alpha$, $dN_{\text{adaptive}}$, and $dN_{\text{neutral}}$, statistically independent. To reduce sampling errors, we grouped genes into 9 classes of roughly equal size and used both doubleton and triplet polymorphisms from E. coli to estimate the level of adaptive evolution; the results were qualitatively similar if we just used tripletons. Not surprisingly, $dN_{\text{adaptive}}$ and $dN_{\text{neutral}}$ were correlated with $dN$ (Spearman’s $\rho = 0.972$ for $dN_{\text{adaptive}}$ and 0.968 for $dN_{\text{neutral}}$); however, there was no significant correlation between $dN$ and $\alpha$ (Spearman’s $\rho = 0.167$, $P = 0.668$; fig. 2).

Previous studies have found that the rate of nonsynonymous substitution is correlated with codon bias in a variety of organisms (Sharp 1991; Rocha and Danchin 2004; Stoletzki et al. 2005) This pattern is also evident in our data ($\rho = -0.674$, $P = 0.001$ and $\rho = -0.783$, $P = 0.000$ for E. coli and S. enterica, respectively). To investigate the relationship further, we divided the rate of nonsynonymous substitution into 2 components, a part apparently due to positive selection ($dN_{\text{adaptive}}$) and a part apparently due to neutral evolution ($dN_{\text{neutral}}$); it should be appreciated that $dN_{\text{adaptive}}$ is probably underestimated, and $dN_{\text{neutral}}$ overestimated because of the segregation of slightly deleterious nonsynonymous mutations. As described above, we divided the data into 17 groups of genes, this time by their CAI value, using doubleton and triplet polymorphisms from E. coli to estimate the level of adaptive evolution. We found that $dN_{\text{neutral}}$ was significantly negatively correlated with codon bias ($\rho = -0.594$, $P = 0.006$), whereas $dN_{\text{adaptive}}$ and $\alpha$ were not significantly correlated ($\rho = 0.402$, $P = 0.079$). The negative correlation between $dN_{\text{neutral}}$ and CAI results in a positive correlation between $\alpha$ and CAI ($\rho = 0.596$, $P = 0.012$, fig. 3). This correlation could potentially arise through selection on synonymous codon use—as selection on synonymous codon use gets stronger, codon bias increases, the ratio $dS/pS$ decreases, and $\alpha$ consequently increases. To investigate this further, we divided our data set of genes into 3 groups according to their CAI values and calculated the average allelic frequency of synonymous and nonsynonymous SNPs in genes with high and low codon bias (table 2). As expected, $\alpha$ varies between these groups (low = 0.39, med = 0.59, high = 0.71). However, although synonymous polymorphisms segregate at slightly lower allelic frequencies in higher than low CAI genes, as expected, the frequency of synonymous mutations is not significantly lower than that of nonsynonymous mutations in the groups with high or low CAI in either species (table 2). This suggests that while

### Table 2

<table>
<thead>
<tr>
<th></th>
<th>Escherichia coli</th>
<th>Salmonella enterica</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Synonymous</td>
<td>Nonsynonymous</td>
</tr>
<tr>
<td>Low CAI</td>
<td>0.28 (0.006)</td>
<td>0.26 (0.009)</td>
</tr>
<tr>
<td>High CAI</td>
<td>0.27 (0.007)</td>
<td>0.25 (0.01)</td>
</tr>
<tr>
<td>All genes</td>
<td>0.27 (0.002)</td>
<td>0.26 (0.004)</td>
</tr>
</tbody>
</table>

**Note.**—All comparisons were significant at $P < 0.0001$.

### Table 3

<table>
<thead>
<tr>
<th></th>
<th>E. coli</th>
<th>S. enterica</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nonsynonymous</td>
<td>Synonymous</td>
</tr>
<tr>
<td>Divergence</td>
<td>6,413</td>
<td>86,576</td>
</tr>
<tr>
<td>Polymorphism frequencies included</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>839</td>
<td>13,302</td>
</tr>
<tr>
<td>Doubletons and tripletons</td>
<td>385</td>
<td>7,203</td>
</tr>
<tr>
<td>Tripletons only</td>
<td>61</td>
<td>1,902</td>
</tr>
</tbody>
</table>

**Note.**—$\alpha$ is calculated independently for all polymorphisms, doubleton and triplet polymorphisms, and triplet polymorphisms only (CIs are given in brackets).
selection against synonymous polymorphisms is stronger in high bias genes, there is also stronger selection on nonsynonymous polymorphisms.

Finally, we tested whether $\alpha$ varied between genes with different functions by grouping the *E. coli* genes according to functional category. All our genes belonged to the general category of housekeeping genes, genes of unknown function excepted (Riley and Labedan 1996). In order to control for the correlation we observed between $\alpha$ and CAI, we first investigated the variation in CAI between functional groups. We find a significant difference in CAI between genes belonging to different functional groups (table 4; ANOVA, $F = 27.24, P = 0$). Tukey’s test showed ribosomal components to be the significantly different group, so we excluded this group from our investigation of variation in $\alpha$. One might intuitively expect this group of genes to have a high CAI; ribosomal components are possibly the most essential of genes and have long been known to be highly expressed and have high codon bias in bacteria (Gouy and Gautier 1982). We then estimated $\alpha$ for each of the remaining functional categories represented by our data and found no variation between categories (figure 4).

**Discussion**

We estimate that at least 50% of amino acid substitutions have been driven by positive selection between the entric bacteria *E. coli* and *S. enterica* in our data set of 410 genes. Unfortunately, we can only estimate a lower bound because some of the nonsynonymous polymorphisms appear to be slightly deleterious, and it is difficult to remove their effects when so few alleles have been sampled. If the sequences had been sampled randomly, then we might have been able to estimate the strength of selection acting upon these polymorphisms and, hence estimate $\alpha$ more accurately (see Fay et al. [2001] and Sawyer et al. [2003] for examples of this approach in other species). However, the sequences are far from randomly sampled, and limited recombination makes the population genetics difficult to model.

**Artifacts**

Although highly statistically significant, the evidence of adaptive evolution could be artifactual for 2 reasons. First, it could be caused by a combination of slightly deleterious nonsynonymous mutations and expanding population size. It is difficult to rule out this possibility completely.

**Table 4**

<table>
<thead>
<tr>
<th>Functional Category</th>
<th>No. of Genes</th>
<th>CAI</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid biosynthesis</td>
<td>28</td>
<td>0.39</td>
<td>0.5</td>
</tr>
<tr>
<td>Central intermediary metabolism</td>
<td>34</td>
<td>0.45</td>
<td>0.54</td>
</tr>
<tr>
<td>Energy metabolism (carbon)</td>
<td>38</td>
<td>0.47</td>
<td>0.63</td>
</tr>
<tr>
<td>Macromolecule synthesis and modification</td>
<td>45</td>
<td>0.43</td>
<td>0.82</td>
</tr>
<tr>
<td>Ribosome constituents</td>
<td>35</td>
<td>0.64</td>
<td>1</td>
</tr>
<tr>
<td>Transport/binding proteins</td>
<td>49</td>
<td>0.41</td>
<td>0.52</td>
</tr>
<tr>
<td>Unknown</td>
<td>79</td>
<td>0.39</td>
<td>0.63</td>
</tr>
</tbody>
</table>

**Figure 2**—Mean $\alpha$ and 95% CIs plotted against dN. Mean $\alpha$ was calculated using *Escherichia coli* polymorphism data (a similar result was observed for *Salmonella enterica*). Spearman’s $\rho = 0.167, P = 0.668$.

**Figure 3**—Mean $\alpha$ and 95% CIs plotted against CAI. Mean $\alpha$ was calculated using *Escherichia coli* polymorphism data (a similar result was observed for *Salmonella enterica*). Spearman’s $\rho = 0.596, P = 0.012$.

**Figure 4**.—Variation in (a) CAI and (b) $\alpha$ between functional categories. Functional groups are: aa = amino acid biosynthesis, cm = central intermediary metabolism, em = energy metabolism (carbon), mm = macromolecule biosynthesis and modification, t/b = transport/binding, rib = ribosomal constituents, u = unknown function.
Although *E. coli* and *S. enterica* are found in a broad variety of mammals and other animals (Selander et al. 1996), many of the complete genome sequences appear to be of human-specific pathogens (e.g., *E. coli* 0157:H7; CFT0373; *S. enterica* typhii; *Choleraesuis*). Their population sizes may therefore have expanded with human populations. However, although the human census population size has increased, estimates of the effective population size in humans are either very similar to or smaller than estimates in other hominids, including chimpanzees (Eyre-Walker et al. 2002), gorillas (Yu et al. 2004), and the ancestor of humans and chimpanzees (Rannala and Yang 2003). This suggests that most recent demographic changes have not increased the effective population size of humans greatly. So unless the diversity in *E. coli* and *S. enterica* is much younger than the diversity in humans, we would also not expect the diversity in them to reflect recent increases in human population sizes. The nucleotide mutation rate has been estimated to be $\sim 5 \times 10^{-10}$ in *E. coli* (Drake 1991). If we accept that the *E. coli* strains we have analyzed belong to the same species, then the effective population size is estimated to be $50 \times 10^9$ using the estimate of synonymous diversity given above. A similar calculation yields an estimate of $20 \times 10^9$ for *S. enterica*. If we accept that *E. coli* goes through a generation every 2 days (Savageau 1983; Selander et al. 1987), then the diversity in *E. coli* is on average 275,000 years old. Thus, the diversity in *E. coli* is older than the diversity in human mitochondria but a little younger than the age of diversity in nuclear genes (Jorde et al. 1998). We currently do not know the average generation time for *S. enterica* and so cannot estimate the age of the diversity.

Second, the evidence of adaptive evolution could be due to the systematic underestimation of $d\ell$ either because synonymous sites are approaching saturation between *E. coli* and *S. enterica* or because there is substantial variation in the strength of selection acting upon synonymous sites. It is difficult to rule out this possibility completely; however, if we divide our data set into 2 halves according to $d\ell$ ($d\ell < 0.75$, 191 genes and $d\ell > 0.75$, 219 genes), we find the estimates of $\alpha$ to be almost identical in the 2 data sets: 0.51 (0.31, 0.78) and 0.58 (0.45, 0.72), respectively. And yet the mean and median values of $d\ell$ are substantially different between the 2 data sets: 0.438 and 0.516 for $d\ell < 0.75$ and 1.06 and 1.03 for $d\ell > 0.75$. If $\alpha$ is an overestimate because $d\ell$ has been underestimated, it would seem remarkable that this bias is the same in genes which have different values of $d\ell$.

**Comparisons with Other Species**

Although we can only set a lower limit on the proportion of substitutions driven by adaptive evolution, it does appear to be higher than that found in either primates (0% to $\sim 35$%; Fay et al. 2001; Chimpanzee Sequencing and Analysis Consortium 2005; Zhang and Li 2005) or *Drosophila* ($\sim 25$%; Bierne and Eyre-Walker 2004)—the very high level of adaptive substitution ($\sim 94$%) in *Drosophila* inferred by Sawyer et al. (2003) may have been a consequence of the method they used; their method assumed that the amino acid mutations contributing to polymorphism and divergence are drawn from the same normal distribution. This may not be the case. The high proportion of adaptive substitution we have found in *E. coli* and *S. enterica* is possibly not surprising given the very large effective population sizes of these bacteria—estimates (see above) of their effective population sizes are at least an order of magnitude greater than *Drosophila* (Begun and Aquadro 1993; Andolfatto 2001) and 3 orders of magnitude greater than primates (Eyre-Walker et al. 2002; Rannala and Yang 2003; Yu et al. 2004). Such large effective population sizes will decrease the probability that slightly deleterious mutations are fixed and increase the probability that advantageous mutations are fixed.

**Genomic Rates**

Bacteria can potentially adapt to their environment in 2 distinct ways; evolution can occur via mutation in the genes they already have, or they may acquire novel genes through horizontal transfer events (Ochman et al. 2000). The full sequences of bacterial genomes have made it evident that many genes are acquired and lost during bacterial evolution—many genes are not shared between strains belonging to a single species (e.g., $\sim 500$ genes found in *E. coli* K12 are not present in *E. coli* 0157:H7) (Perna et al. 2001). Much of this difference in gene content may be adaptive. We have estimated the level of adaptive evolution for a set of genes which are present in all the *E. coli* and *S. enterica* genomes we surveyed; one might think of these as the core genes. On average, these core genes are $\sim 310$ amino acids long (data used here and Blattner et al. [1997]) and differ between *E. coli* and *S. enterica* by $\sim 0.05$ amino acid substitutions per codon. So given that there are 4,288 genes in *E. coli* K12 (Blattner et al. 1997) and a minimum of 55% of these are present in *S. enterica*, we estimate that these 2 species differ by at least 37,000 amino acid differences. Because we estimate that at least 50% of these helped either *E. coli* or *S. enterica* adapt to their environment, this means they differ by a minimum of 18,500 adaptive amino acid substitutions in these core genes. It has been estimated that these 2 species diverged $\sim 100$ MYA, which means they have gone through at least of one adaptive substitution every 11,000 years. This rate of adaptive substitution is far lower than *Drosophila* (1 every 45 years [Smith and Eyre-Walker 2002]); this is possibly due to the smaller genome size of these bacteria.

**Variation in $\alpha$**

We have also investigated whether $\alpha$ varies between genes. Surprisingly, the proportion of nonsynonymous substitutions which are adaptive does not seem to be correlated to the overall rate of nonsynonymous substitution. This is contrary to the results of Fay et al. (2002), who found that $\alpha$ was higher in faster evolving *Drosophila* genes, but agrees with the results of Bierne and Eyre-Walker (2004), who found no evidence of variation in $\alpha$ between genes, also in *Drosophila*. The lack of a correlation between $\alpha$ and $d\ell$ might be due to the fact that our data set is biased toward conserved “core” genes; however, there is more than an order of magnitude variation in $d\ell$ within our data.
Although we found no correlation between \( \alpha \) and \( dN \), we did find a highly significant correlation between \( \alpha \) and the level of codon bias. This seems largely to be a consequence of a correlation between \( dN_{\text{neutral}} \) and codon bias rather than \( dN_{\text{adaptive}} \). This is as expected if the correlation between \( dN \) and codon bias is a result of selection on translational accuracy because \( dN_{\text{neutral}} \) is probably a better indicator of how many of the amino acid sites are critical for function and therefore how strong selection for translational accuracy is likely to be. Adaptive substitutions might occur at both critical and noncritical sites. If this interpretation is correct, then this means that although codon bias might give us some information about the rate of amino acid substitution, it is not likely to be useful in identifying genes undergoing a high rate of adaptive evolution as suggested by Plotkin et al. (2004). We also find no evidence that \( \alpha \) varies between genes of different function, with the caveat that the majority of genes represented by our data set are core genes.

Although Bierne and Eyre-Walker (2004) found no evidence of variation in \( \alpha \) in Drosophila, it remained unclear whether this was a genuine result or a failure of power. Our bacterial data allowed us to investigate the matter further because we found a significant correlation between \( \alpha \) and codon bias—that is, evidence that there is significant variation in \( \alpha \) between genes within these bacteria. When we ran the method of Bierne and Eyre-Walker (2004) on our data set, we found no evidence of variation in \( \alpha \). Models in which \( \alpha \) was beta distributed converged upon a single spike with very little variance and were not favored over models in which alpha was assumed to be constant across genes. Likewise, a constant \( \alpha \) model was favored over a model in which 2 categories of genes were assumed to have different \( \alpha \) values; this model again converged on a single spike where \( \alpha \) was extremely close to our maximum likelihood estimate. This unfortunately suggests that Bierne and Eyre-Walker’s result in Drosophila must be viewed with caution. Furthermore, it implies that it will generally be difficult to test for heterogeneity in \( \alpha \), unless correlates of \( \alpha \) can be found. One option may be to constrain the model further as Bustamante et al. (2005) have done.

**Acknowledgments**

The authors are grateful for the comments of several referees, John Welch, Brian Charlesworth, and Deborah Charlesworth. The authors are supported by the Biotechnology and Biological Sciences Research Council and the National Evolutionary Synthesis Center.

**Literature Cited**


William Martin, Associate Editor

Accepted April 10, 2006