J Mol Evol (2000) 50:264–275
DOI: 10.1007/s002399910031

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary Lability of Context-Dependent Codon Bias in Bacteria

Gilean A.T. McVean,[1] Gregory D.D. Hurst[2]

[1] Institute of Cell, Animal and Population Biology, King's Buildings, West Mains Road, University of Edinburgh, Edinburgh EH9 3JT, UK
[2] Department of Biology, University College London, 4 Stephenson Way, London NW1 2HE, UK

Received: 6 May 1999 / Accepted: 29 October 1999

Abstract. In bacteria, synonymous codon usage can be considerably affected by base composition at neighboring sites. Such context-dependent biases may be caused by either selection against specific nucleotide motifs or context-dependent mutation biases. Here we consider the evolutionary conservation of context-dependent codon bias across 11 completely sequenced bacterial genomes. In particular, we focus on two contextual biases previously identified in *Escherichia coli*; the avoidance of out-of-frame stop codons and AGG motifs. By identifying homologues of *E. coli* genes, we also investigate the effect of gene expression level in *Haemophilus influenzae* and *Mycoplasma genitalium.* We find that while context-dependent codon biases are widespread in bacteria, few are conserved across all species considered. Avoidance of out-of-frame stop codons does not apply to all stop codons or amino acids in *E. coli*, does not hold for different species, does not increase with gene expression level, and is not relaxed in *Mycoplasma* spp., in which the canonical stop codon, TGA, is recognized as tryptophan. Avoidance of AGG motifs shows some evolutionary conservation and increases with gene expression level in *E. coli*, suggestive of the action of selection, but the cause of the bias differs between species. These results demonstrate that strong context-dependent forces, both selective and mutational, operate on synonymous codon usage but that these differ considerably between genomes.

Correspondence to: Gilean A.T. McVean; *e-mail:* g.mcvean@ed.ac.uk

Key words: Codon bias — Bacteria — *Escherichia coli* — Out-of-frame stop codons — AGG motifs — *Haemophilus influenzae* — *Mycoplasma genitalium*

Introduction

For many bacteria, the use of synonymous codons is highly nonrandom. In *Escherichia coli*, highly expressed genes are biased toward those codons recognized by more abundant tRNAs (Grantham et al. 1981; Ikemura 1981), suggesting the influence of selection for translational efficiency. In addition, there exist contextual influences on synonymous codon usage, such that there is significant nonrandomness between base composition at sites in neighboring codons and the codon used (Yarus and Folley 1985; Shpaer 1986; Gouy 1987). For example, in *E. coli*, of the two triplets which code for lysine, AAG is preferred when the next three prime nucleotide is C or A, but AAA is preferred when G follows (Shpaer 1986; Berg and Silva 1997), leading to underrepresentation of sequences of the type AAG:G (a colon indicates the division between adjacent codons) in coding regions (Gutman and Hatfield 1989).

Context-dependent codon bias may be caused by either selection or mutational pressures. Bulmer (1990) demonstrated a correlation between contextual effects in the coding and complementary strands in weakly expressed genes of *E. coli*, suggesting the influence of processes acting at the DNA level, such as mutation bias and selection on DNA structure. Similar forces are suggested by the observations that some contextual biases are

maintained in noncoding regions (Hanai and Wada 1989) and others are independent of the level of gene expression (Berg and Silva 1997). Context-dependent mutational biases may be caused by either differential mutability of sequence motifs or intrinsic bias in DNA repair mechanisms. For example, in *E. coli*, Dcm DNA methylase targets cytosine residues in specific motifs (Boyer et al. 1973) which, because of the high mutability of 5-meC, tend to become depleted from the genome (Coulondre et al. 1978). This effect is counteracted by the very short patch (VSP) repair system, which corrects T:G mismatches found in the same contexts (Lieb 1991). However, VSP repair also "corrects" mismatches found in contexts unassociated with DNA methylation, leading to depletion of other oligonucleotide motifs (Bhagwat and McClelland 1992; Merkl et al. 1992).

Selection acting at the level of translation may also cause context-dependent codon bias. If the efficiency or accuracy with which a particular codon is translated is influenced by neighboring codons, the selective advantage of alternative codons will vary between occurrences of the same amino acid within a gene. For example, in *E. coli*, the efficiency of suppressor tRNAs in reading the TAG[1] amber codon varies considerably with context (Bossi 1983) and misreading of phenylalanine codons as leucine under phenylalanine starvation occurs in some positions of the *argI* gene but not others (Precup et al. 1989). Certain codon pairs are also known to induce frame-shift mutations during translation (Farabaugh 1996). Mechanistically, context dependency during translation can be interpreted in terms of either the interaction between the peptidyl and aminoacyl tRNAs at the A and P sites of a ribosome or codon–anticodon recognition (Yarus and Folley 1985; Shpaer 1986; Gutman and Hatfield 1989). Of course, both mutation bias and translation-mediated selection may be acting, but selection is sufficient to generate an observable effect only in highly expressed genes. In agreement, there appear to be significant differences between contextual biases in highly and weakly expressed genes (Yarus and Folley 1985; Shpaer 1986; Gouy 1987; Gutman and Hatfield 1989; Berg and Silva 1997).

Recently, Maynard Smith and Smith (1996) have suggested that selection acting at the level of translation is responsible for two particular context-dependent codon biases in *E. coli*: the avoidance of out-of-frame stop codons and AGG motifs. Codons resembling termination codons (TRR, where R is A or G) are severely constrained in highly expressed genes, suggesting that the avoidance of premature termination represents a significant selective force (Shpaer 1986). Similarly, highly expressed genes in *E. coli* preferentially use AGA over AGG at arginine codons (for the AGR subset) (Sharp and Li 1987). By considering the sensitivity of such biases to reading frame, Maynard Smith and Smith (1996) found that while both stop codons and the sequence AGG are avoided in the +1 frame [the latter also demonstrated by Gutman and Hatfield (1989)], there is no such underrepresentation in the +2 frame. Sensitivity to reading frame is not predicted by either mutational bias or selection on DNA structure. Hence, it was suggested that selection acting at the level of translation must be the dominant force.

Here we consider the evolutionary conservation of the specific context-dependent codon biases noted by Maynard Smith and Smith (1996). Selection on contextual codon usage due to translational accuracy and/or efficiency must reflect characteristics of the translational machinery, features that we expect to be highly conserved between species. In contrast, the mutational environment varies considerably between genomes, as evidenced by genome wide variation in base composition and differences in the number of DNA repair enzymes. We investigate the generality of the observations made in *E. coli* by analyzing patterns of codon usage in 11 largely or completely sequenced bacterial genomes. This comprises two archaea, *Methanococcus jannaschii* (Bult et al. 1996) and *Archaeglobus fulgidus* (Klenk et al. 1998); one thermatogale, *Thermotoga maritima* (Nelson et al. 1999); two spirochaetes, *Borrelia burgdorferi* (Fraser et al. 1997) and *Treponema pallidium* (Fraser et al. 1998); three firmicutes, *Mycoplasma genitalium* (Fraser et al. 1995), *M. pneumoniae* (Himmelreich et al. 1996), and *Mycobacterium tuberculosis* (www.tigr.org); and three proteobacteria, *Escherichia coli* (Blattner et al. 1997), *Haemophilus influenzae* (Fleischmann et al. 1995), and *Helicobacter pylori* (Tomb et al. 1997). These genomes represent a broad phylogenetic sample and a range in genomewide base composition of 28.5% G+C in *Borrelia burgdorferi* (Fraser et al. 1997) to 52.7% G+C in *Treponema pallidium* (Fraser et al. 1998). The number of specific DNA repair enzymes also differs widely between these species. For example, *E. coli* may have as many as 100 genes involved in DNA repair (Kornberg and Baker 1992), while 30 have been putatively identified in *H. influenzae* (Fleischmann et al. 1995) and considerably fewer in *M. genitalium* (Fraser et al. 1995).

In addition, by using data from *E. coli* on the relative expression level of the genes, and assuming that conservation of gene sequence implies conservation of gene function (hence relative selection on codon usage), we can investigate how the intensity of context-dependent codon bias varies with expression level in each species. This is achieved by using homologues of *E. coli* genes of different inferred expression levels and requires no a priori knowledge of preferred codons in different species. Combined, these analyses provide a means of as-

[1] For convenience, both U and T are referred to as T.

sessing the contribution of mutation bias and translation-mediated selection to context-dependent codon bias. For the case of avoidance of out-of-frame stop codons we can test an additional hypothesis, namely, that where the canonical genetic code has been altered such that the codon TGA is recognized as tryptophan, such as in *Mycoplasma* (Yamao et al. 1985), selection against out-of-frame stop codons should be relaxed.

Methods

Sequence Data

Files containing all known and predicted ORFs in simple FASTA format from each complete or largely complete genome sequence were downloaded from the web. Sequences for *Methanococcus jannaschii* (Mjann), *Archaeglobus fulgidus* (Afulg), *Thermotoga maritima* (Tmar), *Borrelia burgdorferi* (Bburg), *Treponema pallidium* (Tpall), *Mycoplasma genitalium* (Mgen), *Mycobacterium tuberculosis* (Mtub; nearly complete), *Haemophilus influenzae* (Hinf), and *Helicobacter pylori* (Hpyl) are available from www.tigr.org, sequences for *M. pneumoniae* (Mpneu) are available from www.zmbh.uni-heidelberg.de/M_pneumoniae, and *E. coli* (Ecoli) sequences from www.genetics.wisc.edu/html/k12.html. Genes found to contain ambiguous nucleotides and in-frame stop codons were excluded from the analysis.

Sequence Analysis

For each amino acid in each gene we tabulate context-dependent codon usage (CDCU) relative to base composition at the first site in the next 3′ codon. We then calculate the expected number of codons used in each context assuming independence (as for a contingency table). Applying this method to each amino acid separately ensures that our results are not biased by nonrandom amino acid doublets. Observed and expected values are summed over genes (divided into expression level groups where appropriate), and the ratio

$$B = \frac{\sum_{Obs} - \sum_{Exp}}{\sqrt{Var(\sum_{Exp})}} \qquad (1)$$

is used as a summary statistic for the relative bias shown by each codon, where the variance of the sum of expected values is that expected under independence with constant row and column totals. This variance is equivalent to that obtained by randomly shuffling the position of codons in genes while maintaining the amino acid sequence (Appendix). We have found that $\sum_{Exp}$ is well approximated by the normal distribution, with mean and variance given by Eqs. (A4) and (A5), respectively, hence the B value can be directly converted into a P value from tables of critical values. As an example, Fig. 1 shows the simulated distribution of $\sum_{Exp}$ for CAG:G codon contexts in *H. influenzae* obtained by shuffling codon location and the normal distribution approximation. The relative bias shown by each amino acid can be considered from the mean square value of B across codons.

To assess the avoidance of specific motifs, such as out-of-frame stop codons and AGG motifs, we can sum the observed and expected values for each codon context that contains the motif. For stop codons of the type TAA this comprises the codon contexts TTA:A and CTA:A for leucine, ATA:A for isoleucine, and GTA:A for valine. For TAG stop codons we use TTA:G, CTA:G, ATA:G, and GTA:G and for TGA stop codons we consider TTG:A, CTG:A, and GTG:A. For AGG motifs we consider the codon contexts CAG:G for glutamine, GAG:G for glutamic acid, and AAG:G for lysine. For further analysis we also analyze each amino acid separately.

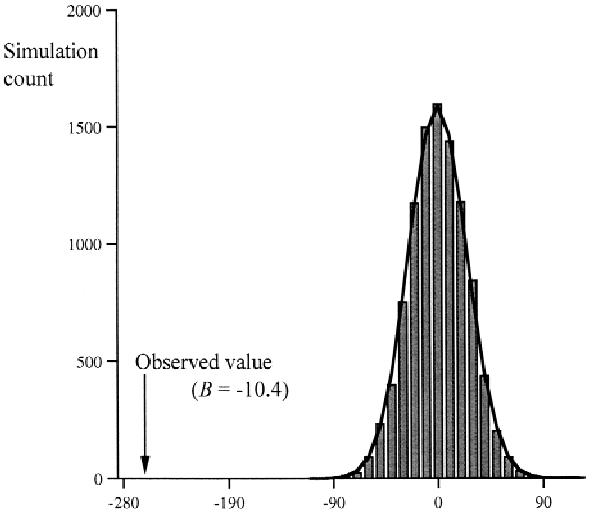

Fig. 1. The distribution of the excess of CAG:G motifs in the *H. influenzae* genome obtained by shuffling codon locations 10,000 times (*columns*), the normal distribution approximation calculated from Eqs. (A4) and (A5) (*solid line*), and the observed value.

To assess the effect of gene expression level on CDCU, the codon adaptation index (CAI) was calculated for genes from *E. coli* (Sharp and Li 1987) by the program CODONS (Lloyd and Sharp 1992). The CAI value can be used as an indirect measure of the level of gene expression (in *E. coli*), given that low-expressed genes tend to have low codon bias (Sharp and Li 1986). From this, we identified a subset of 100 genes for which putative orthologues of *E. coli* genes could be identified in both *H. influenzae* and *M. genitalium.* These genes represent highly conserved functions such as elongation factors, tRNA synthetases, metabolic enzymes, and ribosomal proteins. Hence if we assume that conservation of function implies conservation of relative gene expression level, we can assess the effects of expression level in species where we have no direct estimate of expression levels. The genes were divided into four classes of equal size, assigned by expression level (as measured by CAI in *E. coli*): very highly expressed genes (25 sequences, 8775 codons in *E. coli*; CAI in *E. coli* values in the range 0.66–0.83; *aceE, ackA, dnaK, efp, eno, fus, gap, pgk, oyk, rpL1, rpL11, rpL13, rpL15, rpL2, rpL3, rpL31, rpL34, rpL4, rpL7, rpoC, rpS17, rpS3, rpS9, tkt, tsf, tufA*); highly expressed genes (25 sequences, 8502 codons; CAI range, 0.57–0.65; *aceF, apt, aspS, atpE, gltX, leuS, pdhD, pta, ptsH, recA, rpL10, rpL18, rpL19, rpL21, rpL32, rpL5, rpL6, rpoB, rpoD, rpS10, rpS11, rpS18, rpS5, rpS6, valS*); moderately expressed genes (25 sequences, 10,163 codons; CAI range, 0.46–0.56; *alaS, argS, atpF, fruA, ftsH, ftsZ, glySa, glySb, ileS, metG, pheS, pheT, proS, rpL22, rpL24, rpL29, rpS19, rpS4, rpS7, rpS8, serS, secA, trpS, trxA, tyrS*); and low-expressed genes (25 sequences, 11,278 codons; CAI range, 0.32–0.45; *clpB, dnaB, dnaN, fmt, ftsY, galU, hisS, lig, msbA, potA, potB, potC, rbsC, rimK, rpL14, rpoA, rpS13, secY, thrS, topA, trxB, uvrA, uvrB, uvrC, uvrD*).

An alternative method of analyzing gene expression levels effects would be to use relative synonymous codon usage in orthologues of very highly expressed genes to calculate the codon adaptation index for each species. Ranking genes by CAI within genomes could then be used as an indirect method of assessing gene expression level. However, we have found that this approach gives artifactual results in genomes with strongly biased base composition. In such circumstances, ranking genes by inferred CAI reflects variation between genes in

Table 1. Highly conserved context-dependent codon biases in bacteria

Amino acid	Underrepresented	Overrepresented
Asp	GAT:A	GAC:A
Tyr	TAT:A	TAC:A
Ala		GCC:A[a]
Gly		GGT:T[a]
		GGG:C[a]
Val	GTT:A[a]	
Arg		AGA:A
Leu	CTT:A	CTT:T
Ser		TCC:A[a]

[a] Direction of bias conserved in all species with $|B| > 3.29$ (minimum nine species).

codon usage due to stochasticity, or variation in mutational biases, rather than gene expression level.

Results

Conservation of Context-Dependent Codon Bias

For each amino acid in each gene analyzed we have calculated both synonymous codon usage dependent on the context of the first base of the next 3′ nucleotide (N1) and that expected if codon usage were independent of neighboring base composition. By summing these values over genes we can demonstrate genome wide departures from independence. In addition, we can consider the extent to which particular contextual biases are conserved across species, whether some amino acids consistently show greater context-dependent biases than others, and the degree to which genomes vary in the extent of context-dependent bias.

Of the 236 codon × N1 combinations (excluding methionine, tryptophan, and stop codons), 7 show contextual bias in the same direction in every species analyzed (Table 1), and another 5 show contextual bias in every species for which the absolute bias is greater than 3.29 (with a minimum of nine species), corresponding to a P value of 0.001. Of these codon contexts, six represent codon pairs in which an overrepresented context differs from an underrepresented codon context by a single base pair, a pattern which has been suggested to arise from DNA repair associated mutation biases (Karlin et al. 1997).

The relative degree of context-dependent bias for different amino acids can be calculated by ranking amino acids according to their mean square B value in each species, dividing them into those showing low, moderate, and high levels of context-dependent bias, and plotting the frequency with amino acids fall into particular classes across species (Fig. 2). The amino acids phenylalanine, glutamic acid, and leucine consistently show high levels of context-dependent codon bias, while serine, arginine, and cysteine consistently show low levels of context-dependent bias. Across species, there is considerable variation in the degree of context-dependent codon bias, but there are no obvious trends; *Helicobacter pylori* is the most biased, while *Treponema pallidium* is the least.

The Avoidance of Out-of-Frame Stop Codons and AGG Motifs

We can assess the avoidance of specific motifs by summing the effects over all codon contexts which contain the motif. Figure 3 shows the avoidance of out-of-frame stop codons and AGG motifs in the 11 species considered, mapped onto a phylogeny representing the current understanding of phylogenetic relationships of the bacterial groups (although the branching order of the spirochaetes, proteobacteria, and firmicutes is not resolved). Motifs including out-of-frame stop codons show strong context-dependent bias. However, none are consistently avoided across all species [or all species with significant ($P < 0.001$) context-dependent bias]. In *E. coli*, NTA:G motifs show the strongest avoidance, while NTA:A motifs, as in 9 of 11 species, are considerably overrepresented. Furthermore, while NTG:A motifs are underrepresented in *E. coli* (and several other species), there is no evidence for relaxation of this pressure in *Mycoplasma* spp. Indeed, *M. pneumoniae* shows the strongest avoidance of this motif among all species considered. Analysis of individual codon contexts contributing to out-of-frame stop codons also indicates a lack of conservation of the observations made in *E. coli*; no single codon context containing an out-of-frame stop codon is consistently underrepresented across species.

Context-dependent biases involving AGG motifs in the +1 frame show a similar pattern. Biases are strongest in *E. coli*, in which motifs of the form RAG:G show considerable underrepresentation (R is A or G), while the codon context CAG:G is overrepresented. In those species showing significant ($P < 0.001$) context-dependent biases involving AGG motifs, there is no general over- or underrepresentation of either CAG:G or AAG:G motifs, however, GAG:G motifs are underrepresented in 8 of 10 species. These results suggest no strong conservation of any context-dependent forces, except perhaps for GAG:G motifs. Given the extreme nature of some of the contextual biases (for example NTG:A is overrepresented by 80 standard deviations in *T. maritima*), such a lack of conservation seems remarkable.

The Effect of Expression Level

To what extent does conservation (or lability) of context-dependent codon bias between amino acids and species argue for a mutational or selective explanation? While we can conclude that avoidance of out-of-frame stop

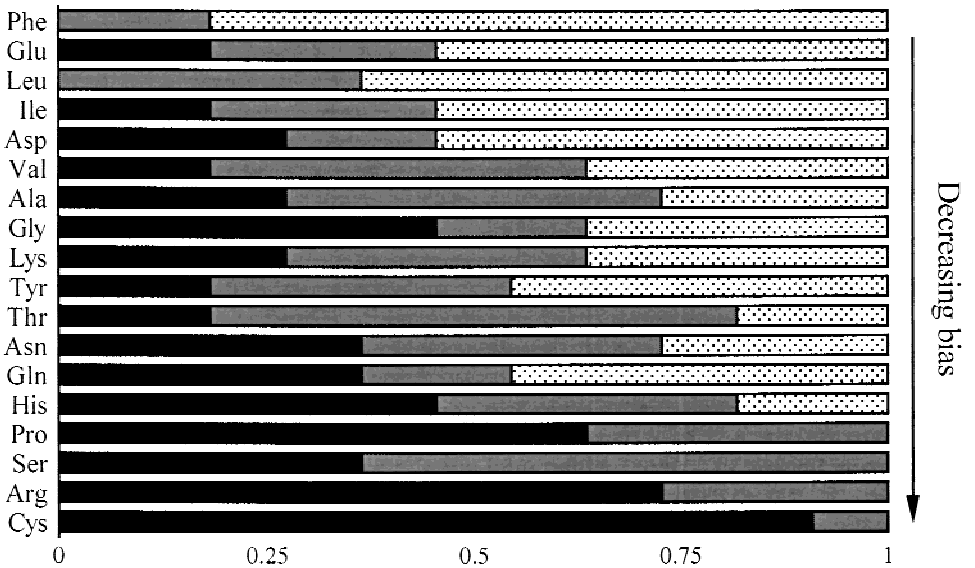

Fig. 2. Context-dependent codon bias for different amino acids. Each *bar* represents the proportion of species in which the amino acid is ranked as showing high (*dotted*), moderate (*gray*), or weak (*black*) context-dependent bias, by mean square bias.

codons is not a general force in the evolution of contextual codon bias (given that it does not apply to all species, amino acids, or even stop codons), we cannot rule out the possibility that avoidance of TAR stop codons in the +1 frame is an important factor in determining codon usage in *E. coli.* Conversely, conservation of context-dependent codon bias across species of different mutational environments (such as suggested for the avoidance of GAG:G motifs) could well be explained by highly conserved selective forces, such as those originating from the process of translation. But they might equally well be caused by highly conserved features of mutation, either factors inherent to the mutation process or biases caused by highly conserved DNA repair enzymes.

We therefore require an additional means of assessing the contribution of mutation bias and selection to context-dependent codon biases. One approach is to investigate the effect of expression level on the magnitude of the bias (Berg and Silva 1997). Explicit population genetic models of codon bias (Bulmer 1991) can be manipulated to predict how the degree of context-dependent codon bias is affected by the strength of selection resulting from gene expression level, when the cause of the contextual bias is either mutation bias or selection. In general, contextual biases resulting from selection tend to become exaggerated with increasing expression level, while mutationally caused biases are little affected or become weaker.

To assess the effect of gene expression level on context-dependent codon bias, we require data on the relative expression level of the genes considered. In *E. coli*, the codon adaptation index (CAI) has been used as an indirect measure of gene expression level (Bulmer 1990), as the CAI is a measure of codon bias relative to codon usage in highly expressed genes (Sharp and Li 1987), and codon bias increases with gene expression level (Grantham et al. 1981; Ikemura 1981; Sharp and Li 1986). For other species, no equivalent data are available. However, if we make the assumption that conservation of gene function implies conservation of relative gene expression level, then we can use orthologues of genes of known expression level in *E. coli* to investigate the effects of expression level in other species.

To achieve this, we identified 100 putative orthologous genes in *E. coli*, *H. influenzae*, and *M. genitalium* and used CAI values in *E. coli* to group genes by expression level. For each class of gene we can then calculate context-dependent codon frequencies. If the cause of a contextual bias is mutational, we expect context-dependent bias to decrease with increasing expression level, and lines representing codon usage dependent on the nucleotide at the first position of the following codon (N1) either to be parallel or to converge with increasing expression level. In contrast, if the cause is selective, bias should increase with expression level and lines should diverge.

Figures 4–6 show the effects of inferred expression level on context-dependent codon usage for amino acids which can generate stop codons in the +1 frame (isoleucine, valine, and leucine), for *E. coli*, *H. influenzae*, and *M. genitalium*. Figure 7 shows the effects of inferred expression level on context-dependent codon usage for amino acids which can generate AGG motifs in the +1 frame (glutamine, glutamic acid, and lysine). Each graph shows the relative frequency of a particular codon dependent on the nucleotide at the first position of the next 3′ codon; congruent lines indicate an absence of context-dependent bias.

Three points should be noted. First, as is well established in *E. coli* and *H. influenzae*, several codons con-

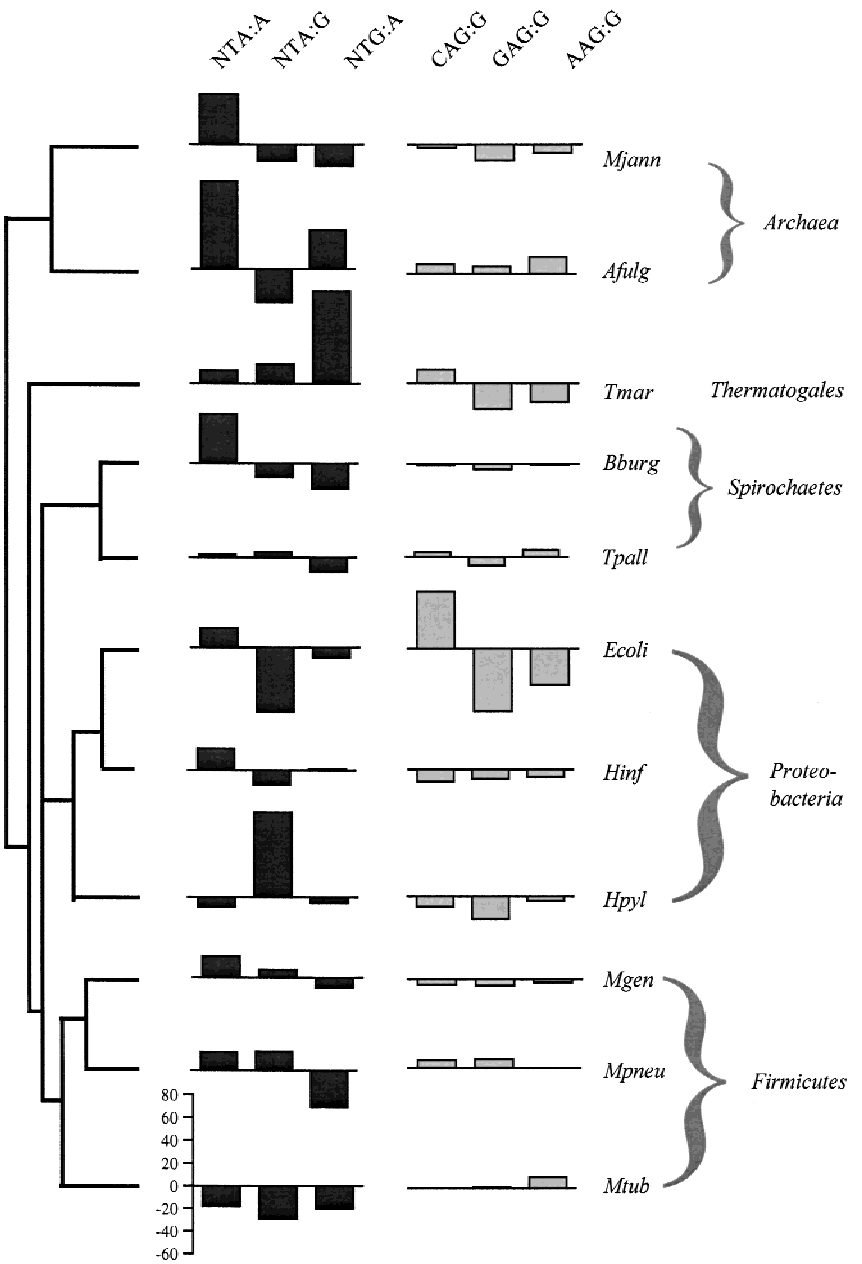

Fig. 3. Phylogenetic patterns of context-dependent codon bias for motifs which contain stop codons and AGG triplets in the +1 frame. *Columns* represent the *B* value [Eq. (1)] and all graphs are on the same scale.

sistently increase or decrease in frequency over inferred expression level, indicating selective effects on synonymous codon usage. This is not generally the case in *M. genitalium*, where variation in codon usage is thought largely to reflect genomewide variation in base composition rather than selection relating to gene expression level (Kerr et al. 1997). However, some codons increase in frequency in certain contexts, but not others. For example, ATC increases when N1 is T for isoleucine and CTT usage also increases when N1 is T for leucine, suggesting that although selection on codon usage may be weak in *M. genitalium*, it is not completely absent. Second, almost all codons in all three species show strong context-dependent effects. Codons such as AAG, which shows little context-dependent bias in either *H. influenzae* or *M. genitalium*, are the exception. For some codons in *E. coli*, changes in codon frequency across expression level are almost entirely the result of changes in contextual bias. For example, the decrease in usage of GAG for glutamic acid is simply a result of increased avoidance of GAG:G motifs (see also Berg and Silva 1997). Finally, the majority of context-dependent biases demonstrate patterns more consistent with mutational bias than selection. That is, context-dependent biases are strong in weakly expressed genes and are little affected by, or decrease with, increasing expression level.

However, some context-dependent biases do appear to increase with expression level, indicative of selection. Examples include the avoidance of AAG:G motifs in *E. coli* and GTT:C motifs in *M. genitalium.* Furthermore,

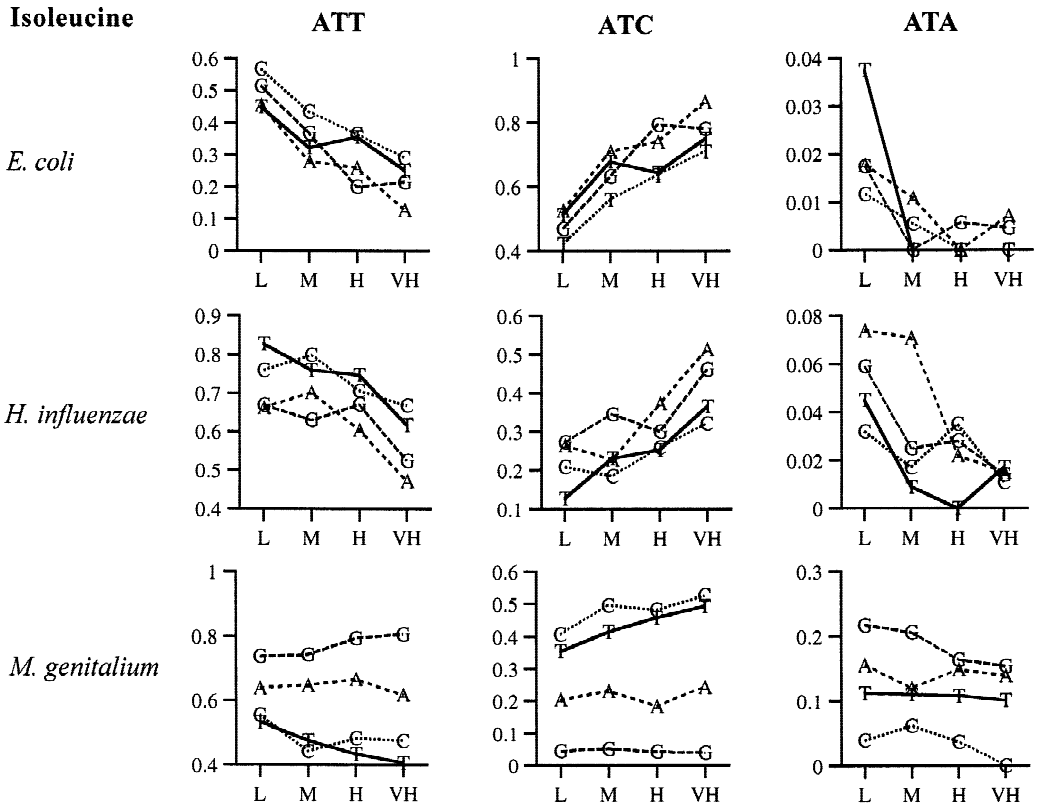

Fig. 4. The effect of inferred expression level on context-dependent codon usage for isoleucine (relative to N1) for 100 putatively orthologous genes in *E. coli*, *H. influenzae*, and *M. genitalium.*

when all genes in *E. coli* are used for the analysis (ranked by CAI), it is clear that other biases also show the hallmark of selection, for example, the avoidance of GAG:G and GTC:C motifs (data not shown). The discrepancy between the results based on the entire genome and those based on the subset of 100 conserved genes is due to a tendency for the conserved genes to be more biased than average. For example, in very low-bias *E. coli* genes (CAI <0.2) there is no significant context-dependent bias for the valine codon GTC, while in the 100 conserved genes, the lowest-bias group (CAI <0.45) shows very strong avoidance of GTC:C motifs. Similarly, the codon GAG for glutamic acid is used with equal frequency when N1 is A or G in genes with CAI <0.2, but usage with N1 equal to G decreases with increasing bias, such that there is considerable avoidance of GAG:G motifs in the lowest-bias group in our subsample.

Underrepresented motifs which result in a significant genomewide dearth of out-of-frame stop codons comprise ATA:G, GTA:A, GTA:G, TTA:G, CTA:G, and CTG:A in *E. coli* and GTA:G and CTA:G in *H. influenzae* (there are none in *M. genitalium*). With the possible exception of CTA:G avoidance in *E. coli*, none of these biases show features suggestive of the action of selection. Context-dependent biases which result in a significant genomewide dearth of AGG motifs are GAG:G and AAG:G in *E. coli*, CAG:G, GAG:G, and AAG:G in *H. influenzae*, and CAG:G and GAG:G in *M. genitalium.* However, while there is evidence that RAG:G motif avoidance has a selective basis in *E. coli*, plots of codon frequencies (Fig. 7) indicate that the dearth of AGG motifs in the other species is simply a consequence of the overrepresentation of other motifs, notably overrepresentation of G-ending codons for glutamine, glutamic acid, and lysine when N1 is C in *H. influenzae* and overrepresentation of G-ending codons for the same amino acids when N1 is T in *M. genitalium.* That is, there is no specific avoidance of AGG motifs in either of these species.

In sum, while context-dependent codon biases are widespread in bacteria and can be considerable in magnitude, there is very little evolutionary conservation of the causative forces and most demonstrate patterns suggestive of mutational bias. In only a few cases can we identify biases which increase with inferred expression level in *E. coli*, as may be expected from a selective cause. However, it is also possible that such patterns may be caused by expression-associated mutational biases (Francino and Ochman 1997). The avoidance of out-of-frame stop codons in *E. coli* appears to be a fortuitous consequence of unknown mutational bias, and while RAG:G motifs appear to be selectively avoided in *E. coli*, this does not seem to be a general pattern in bacteria.

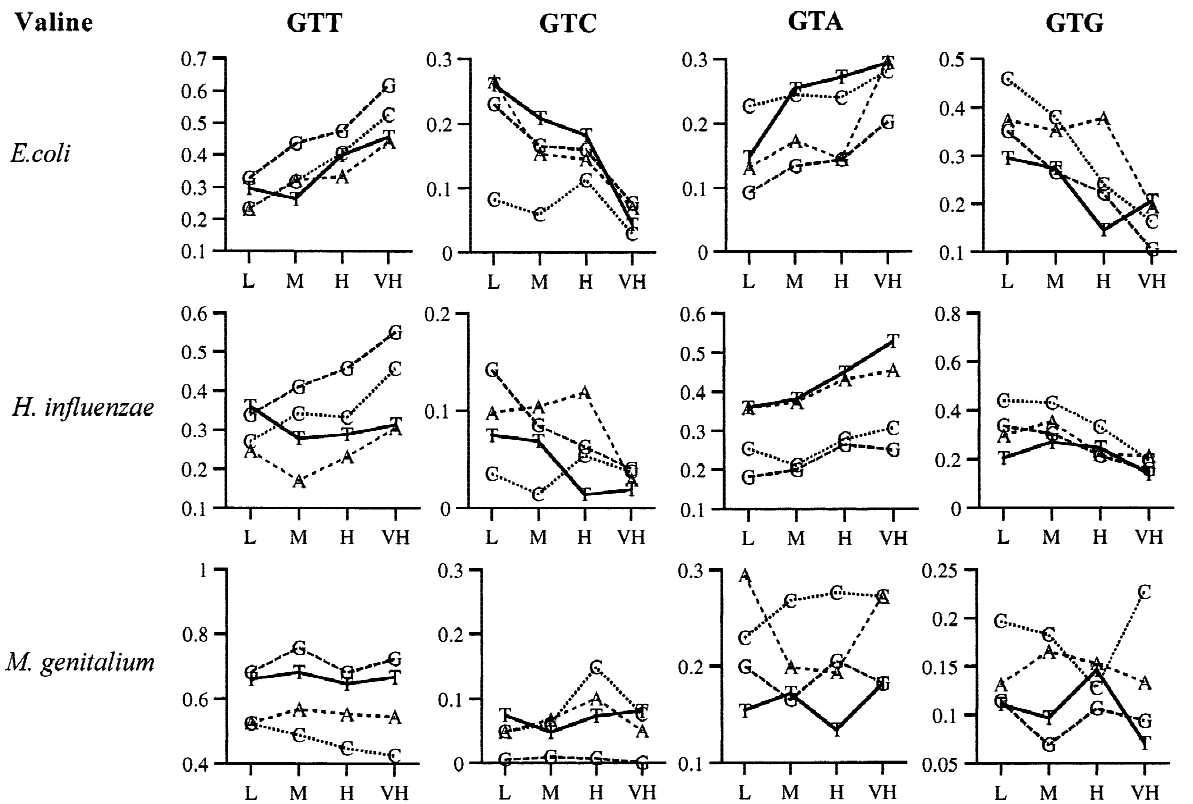

Fig. 5. The effect of inferred expression level on context-dependent codon usage for valine for 100 putatively orthologous genes in *E. coli*, *H. influenzae*, and *M. genitalium.*

Discussion

We have claimed that while the avoidance of out-of-frame stop codons appears to be the fortuitous consequence of context-dependent mutation bias in *E. coli*, the avoidance of RAG:G motifs is the result of selective forces, but forces which are not conserved across bacteria. But exactly what factors may be responsible in each case?

Out-of-Frame Stop Codons

Underrepresentation of the motifs ATA:G, GTA:A, GTA:G, TTA:G, CTA:G, and CTG:A in the *E. coli* genome can be explained only partially by biases known to exist within *E. coli*. For example, the underrepresentation of the motif CTAG in the *E. coli* genome is well documented (Phillips et al. 1987; Burge et al. 1992; Blattner et al. 1997) and may, to some extent, represent intrinsic biases in the very short patch (VSP) DNA repair system (Bhagwat and McClelland 1992; Merkl et al. 1992). VSP repair corrects T:G mismatches in certain contexts to C:G pairs, some of which (e.g., in 5′CTWGG/3′-GGW′CC; W is A or T) result from methylcytosine deamination, but others of which do not (e.g., TWGG/GW′CC and CTWG/GGW′C), leading to the depletion of sequences of the type TWGG and CTWG and their complements. CTAG motifs are most underrepresented in protein-coding regions (Blattner et al. 1997), which could also suggest a selective explanation based on interference with secondary structure or *trpR* binding (Médigue et al. 1991; Burge et al. 1992). However, CTAG avoidance shows little increase with expression level (Fig. 3) (Eyre-Walker 1995; but see Gutiérrez et al. 1994), suggesting a mutational explanation.

Although the VSP repair system can act on several different sequence substrates, and appears to lead to the depletion of certain oligonucleotide motifs largely in proportion to substrate affinity (Gläsner et al. 1995), it alone cannot explain the observed set of context-dependent codon biases. For example, neither the motif GTA-R nor its complement is a substrate for VSP action (unless there is a particularly high proportion of GTA:GG, but even this does not explain GTA:A avoidance). In addition, CTAG motifs are also underrepresented in the *H. influenzae* genome (Karlin et al. 1997), which has no homologue of the *vsr* gene involved in VSP repair (Fleischmann et al. 1995). The avoidance of palindromic motifs (Karlin et al. 1997), potentially associated with restriction enzyme systems, is similarly unable to explain the observed contextual biases. We are therefore left to postulate the presence of a mutational bias with an unknown cause. Two pieces of evidence suggests that such biases may be widespread. First, the analysis of Berg and Silva (1997) has identified many contextual biases in *E. coli* suggestive of mutational effects (i.e.,

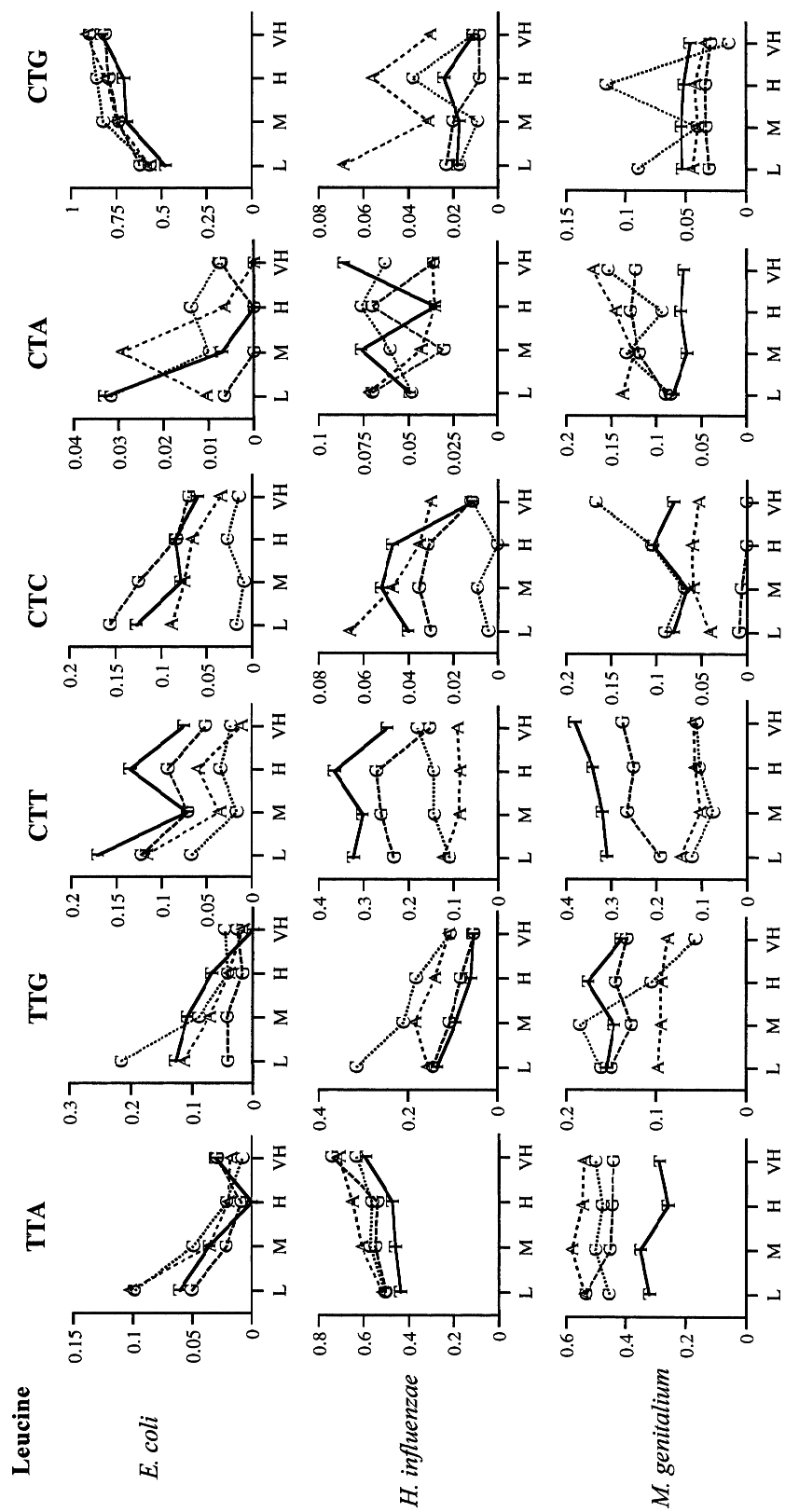

Fig. 6. The effect of inferred expression level on context-dependent codon usage for leucine for 100 putatively orthologous genes in *E. coli*, *H. influenzae*, and *M. genitalium*.

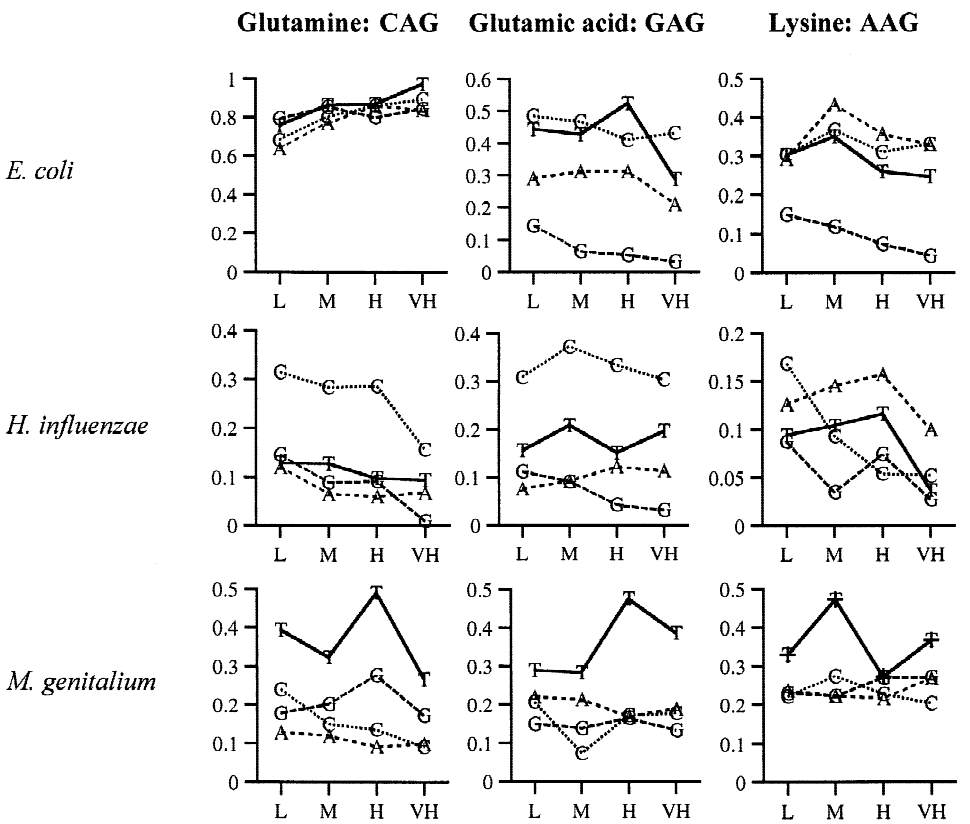

Fig. 7. The effect of inferred expression level on context-dependent codon usage for glutamine, glutamic acid, and lysine for 100 putatively orthologous genes in *E. coli*, *H. influenzae*, and *M. genitalium*.

where the magnitude of the effect decreases with expression level). Second, in *H. influenzae* there exists a wide range of under- and overrepresented tetranucleotide motifs which differ by a single base change (Karlin et al. 1997), which suggests the action of biased DNA mismatch repair enzymes.

Avoidance of RAG:G Motifs

We favor a selective explanation for the avoidance of RAR:G motifs in *E. coli* because the bias increases with expression level. Through analyzing the effect of expression level on the same biases, Berg and Silva (1997) also came to the conclusion that a selective force was responsible. One possibility is that the avoidance of sequences similar to the Shine–Dalgarno motif (TAAGGAGGT), important for the identification of the translation initiation site on the mRNA by the ribosome, is responsible for the death of AGG in the +1 frame (Berg and Silva 1997). Berg and Silva (1997) discount this suggestion by noting that increased resemblance to the Shine–Dalgarno motif has no apparent selective disadvantage. Avoidance of RAG:G motifs due to Shine–Dalgarno similarity also seems unlikely given that the bias is not conserved across species, while Shine–Dalgarno initiation is. In addition, in *E. coli*, Maynard Smith and Smith (1996) found avoidance of the AGG motif only in the +1 and not in the +2 frame. However, as they were unaware that the context-dependent bias was restricted to RAG:G motifs, the simple tests they employed may not have been sufficient to identify RA:GG avoidance. We have therefore repeated the analysis of avoidance of AGG in the +2 frame in *E. coli* by considering the relative use of A- and G-ending codons for glutamine, glutamic acid, and lysine when the following codon begins with either GG or GH (where H is A, C, or T). We find that neither glutamic acid nor lysine shows a reduced preference for A-ending codons in front of glycine (GGN), as would be expected if avoidance of the Shine–Dalgarno motif were important. For glutamic acid, we find that, contrary to predictions, A-ending codons are actually less avoided when the following codon begins with GG (data not shown). In sum, there is no evidence that RA:GG motifs are avoided, further suggesting that the resemblance to the Shine–Dalgarno motif is purely coincidental.

Frame sensitivity of RAGG motif avoidance suggests that factors acting during translation are responsible. Possible factors are the propensity for tRNAs coding for inappropriate amino acids to bind or for frameshift mutations to occur. We can, however, see no obvious correlation between contextual bias and features of the translation machinery. For the twofold degenerate amino acids, the preferred codon (that which increases in frequency with gene expression level) always corresponds

to one for which the appropriate transfer RNA is present in the genome (Sprinzl et al. 1998). And those amino acids which show the strongest selective increase in contextual codon bias with expression level (Berg and Silva 1997) are also those where only one of the two tRNAs are present in the genome, suggesting that contextual effects may be due to destabilization of noncognate base pairings. However, not all amino acids where only a single tRNA is present show strong selective contextual effects, and while *E. coli*, *H. influenzae*, and *M. genitalium* each have only the tRNA corresponding to GAA for glutamic acid, each species has a different context-dependent bias. Finally, it is worth noting that it is unlikely that any simple set of rules applies to the interaction between codon context and translation. For example, misreading of phenylalanine by leucine under phenylalanine starvation in the *argI* gene of *E. coli* appears to be influenced by context, but not the particular codon used (Precup et al. 1989), while asparagine misreading is strongly affected by codon and less so by context (Precup and Parker 1987).

In short, the high levels of evolutionary lability of context-dependent biases and the lack of any obvious relationship between the set of tRNAs in a genome and the level of bias suggest that the causes of bias are genome-specific and unlikely to reflect fundamental features of the translation machinery. Many biases appear to result from context-dependent mutational biases, but again, the majority of these mutational biases appear to differ between species, suggesting considerable differences in the set, or at least specificities, of DNA repair enzymes or other factors which can influence mutation.

Appendix: The Variance in Contingency Table Cell Frequencies Under Independence with Fixed Row and Column Totals

For a given gene we wish to calculate the distribution of codon contexts generated by randomly shuffling codon positions, while maintaining the amino acid order. In particular, we wish to know the expected value and the variance of that value. This problem is equivalent to calculating the distribution of cell frequencies in a contingency table with fixed row and column totals. If we consider each cell in turn, any R×C contingency table can be reduced to a 2×2 table. For example, if we are interested in the codon context GGG:G, we can consider the codon classes GGG and GG~G and the N1 classes G and ~G (where ~G means not G). We then have a table such as

Codon	N1 = G	N1 = ~G	Total
GGG	a	b	R_1
GG~G	c	d	R_2
Total	C_1	C_2	T

Under the assumption of independence of rows and columns, cell frequencies are multinomially distributed:

$$\Pr\{a,b,c,d\} = \frac{T!}{a!b!c!d!}\left(\frac{R_1C_1}{T^2}\right)^a\left(\frac{R_1C_2}{T^2}\right)^b\left(\frac{R_2C_1}{T^2}\right)^c\left(\frac{R_2C_2}{T^2}\right)^d \tag{A1}$$

With fixed row and column totals we can rewrite the unconditional probability A1 in terms of a alone:

$$\begin{aligned} b &= R_1 - a \\ c &= C_1 - a \\ d &= T - R_1 - C_1 - a \end{aligned} \tag{A2}$$

If we consider just those outcomes with positive values of a which give the observed row and column totals, the probability of observing cell count a is just the number of ways of obtaining cell counts $\{a,b,c,d\}$ divided by the total number of ways of obtaining the row and column totals. This gives

$$\Pr\{a\} = \frac{R_1!R_2!C_1!C_2!}{T!a!(R_1-a)!(C_1-a)!(T-R_1-C_1+a)!} \tag{A3}$$

with $a_{\min} = \max(0, R_1 - C_2)$ and $a_{\max} = \min(R_1, C_1)$. This probability is that calculated in Fisher's exact test (Sokal and Rohlf 1995, p. 733). The expected value of a is simply the product

$$E[a] = \frac{R_1C_1}{T} \tag{A4}$$

and the variance of a is

$$\sigma_a^2 = \frac{R_1C_1R_2C_2}{T^2(T-1)} \tag{A5}$$

Summing (A5) across genes gives the variance in the sum of a across genes. For the large sample sizes employed here, the sum of a across genes is well approximated by a normal distribution, with mean equal to the sum of (A4) across genes and variance equal to the sum of (A5) across genes (See Fig. 1).

Acknowledgments. We wish to thank John Maynard Smith and Noel Smith for comments and access to sequence data, Laurence Hurst for discussion, and three anonymous reviewers. G.M. is funded by the NERC; G.H. is funded by a BBSRC D. Phillips fellowship.

References

Berg OG, Silva PJN (1997) Codon bias in *Escherichia coli*: The influence of codon context on mutation and selection. Nucleic Acids Res 25:1397–1404

Bhagwat AS, McClelland M (1992) DNA mismatch correction by Very Short Patch repair may have altered the abundance of oligonucleotides in the *E. coli* genome. Nucleic Acids Res 20:1663–1668

Blattner FR, Plunket GI, Bloch CA, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462

Bossi L (1983) Context effects: Translation of UAG codon by suppressor tRNA is affected by the sequence following UAG in the message. J Mol Biol 164:73–87

Boyer HW, Chow LT, Dugaiczyk A, Hedgpeth J, Goodman HM (1973) Nature New Biol 244:40–43

Bulmer M (1990) The effect of context on synonymous codon usage in genes with low codon usage bias. Nucleic Acids Res 18:2869–2873

Bulmer MG (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907

Bult CJ, White O, Olsen GJ, et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii.* Science 273:1058–1073

Burge C, Campbell AM, Karlin S (1992) Over- and underrepresentation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci USA 89:1368–1362

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in *Escherichia coli.* Nature 274: 775–780

Eyre-Walker A (1995) Does Very Short Patch (VSP) repair efficiency vary in relation to gene expression levels? J Mol Evol 40:705–706

Farabaugh PJ (1996) Programmed translational frameshifting. Annu Rev Genet 30:507–528

Fleischmann RD, Adams MD, White O, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Francino MP, Ochman H (1997) Strand asymmetries in DNA evolution. Trends Genet 13:240–245

Fraser CM, Gocayne JD, White O, et al. (1995) The minimal gene complement of *Mycoplasma genitalium.* Science 270:397–403

Fraser CM, Casjens S, Huang WM, et al. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi.* Nature 390:580–586

Fraser CM, Norris SJ, Weinstock CM, et al. (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281:375–388

Gläsner W, Merkl R, Schellenberger V, Fritz H (1995) Substrate preferences of Vsr DNA mismatch endonuclease and their consequences for the evolution of the *Escherichia coli* K-12 genome. J Mol Biol 245:1–7

Gouy M (1987) Codon contexts in enterobacterial and coliphage genes. Mol Biol Evol 4:426–444

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 8:r49–r62

Gutiérrez G, Casadesús J, Oliver J, Marín A (1994) Compositional heterogeneity of the *Escherichia coli* genome: A role for VSP repair? J Mol Evol 39:340–346

Gutman GA, Hatfield GW (1989) Nonrandom utilization of codon pairs in *Escherichia coli.* Proc Natl Acad Sci USA 86:3699–3703

Hanai R, Wada A (1989) Novel third-letter bias in *Escherichia coli* codons revealed by rigorous treatment of coding constraints. J Mol Biol 207:655–660

Himmelreich R, Hilbert H, Plagens H, et al. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae.* Nucleic Acids Res 24:4420–4449

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151:389–409

Karlin S, Mrázek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179: 3899–3913

Kerr ARW, Peden JF, Sharp PM (1997) Systematic base composition variation around the genome of *Mycoplasma genitalium*, but not *Mycoplasma pneumoniae*. Mol Microbiol 25:1177–1179

Klenk HP, Clayton RA, Tomb JF, et al. (1998) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus.* Nature 390:364–379

Kornberg A, Baker TA (1992) DNA replication. Freeman, New York

Lieb M (1991) Spontaneous mutation at a 5-methylcytosine is prevented by Very Short Patch (VSP) mismatch repair. Genetics 128: 23–27

Lloyd AT, Sharp PM (1992) CODONS: A microcomputer program for codon usage analysis. J Hered 83:239–240

Maynard Smith J, Smith NH (1996) Site-specific codon bias in bacteria. Genetics 142:1037–1043

Médigue C, Viari A, Hénaut A, Danchi A (1991) *Escherichia coli* molecular genetic map (1500 kbp): Update II. Mol Microbiol 5: 2629–2640

Merkl R, Kröger M, Rice P, Fritz H (1992) Statistical evaluation and biological interpretation of non-random abundance in the *E. coli* K-12 genome of tetra- and pentanucleotide sequences related to VSP DNA mismatch repair. Nucleic Acids Res 20:1657–1662

Nelson KE, Clayton RA, Gill SR, et al. (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequences of *Thermotoga maritima.* Nature 399:323–329

Phillips GJ, Arnold J, Ivarie R (1987) Mono- through hexanucleotide composition of the Escherichia coli genome: A Markov chain analysis. Nucleic Acids Res 15:2611–2626

Precup J, Parker J (1987) Missense misreading of asparagine codons as a function of codon identity and context. J Biol Chem 262:11351–11355

Precup J, Ulrich AK, Roopnarine O, Parker J (1989) Context specific misreading of phenylalanine codons. Mol Gen Genet 218:397–401

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

Sharp PM, Li W-H (1987) The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential application. Nucleic Acids Res 15:1281–1295

Shpaer EG (1986) Constraints on codon context in *Escherichia coli* genes—Their possible role in modulating the efficiency of translation. J Mol Biol 188:555–564

Sokal RR, Rohlf FJ (1995) Biometry. W.H. Freeman, New York

Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S (1998) Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res 26:148–153

Tomb JF, White O, Kerlavage AR, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori.* Nature 388: 539–547

Yamao F, Muto A, Kawauchi Y, et al. (1985) UGA is read as tryptophan in *Mycoplasma capricolum.* Proc Natl Acad Sci USA 82: 2306–2309

Yarus M, Folley LS (1985) Sense codons are found in specific contexts. J Mol Biol 182:529–540