J Mol Evol (1996) 42:73–78

JOURNAL OF MOLECULAR EVOLUTION

The Close Proximity of *Escherichia coli* Genes: Consequences for Stop Codon and Synonymous Codon Use

Adam Eyre-Walker*

Department of Biological Sciences, Rutgers University, Piscataway, NJ 08855-1059, USA

Received: 17 February 1995 / Accepted: 15 September 1995

Abstract. It is shown that synonymous codon usage is less biased in favor of those codons preferred by highly expressed genes at the end of *Escherichia coli* genes than in the middle. This appears to be due to the close proximity of many *E. coli* genes. It is shown that a substantial number of genes overlap either the Shine-Dalgarno sequence or the coding sequence of the next gene on the chromosome and that the codons that overlap have lower synonymous codon bias than those which do not. It is also shown that there is an increase in the frequency of A-ending codons, and a decrease in the frequency of G-ending codons at the end of *E. coli* genes that lie close to another gene. It is suggested that these trends in composition could be associated with selection against the formation of mRNA secondary structure near the start of the next gene on the chromosome. Stop codon use is also affected by the close proximity of genes; many genes are forced to use TGA and TAG stop codons because they terminate either within the Shine-Dalgarno or coding sequence of the next gene on the chromosome. The implications these results have for the evolution of synonymous codon use are discussed.

Key words: Synonymous codons — Stop codons — Shine-Dalgarno — *Escherichia coli*

Introduction

Many organisms preferentially use a subset of the synonymous codons available to them (Sharp 1989). In *Escherichia coli* the codons used are those which bind the commonest tRNAs with normal base pairing (Ikemura 1985; Bulmer 1988). Similarly the three stop, or termination, codons are used unequally—TAA is used in *E. coli* in preference to TGA, which itself is used in preference to TAG (Sharp and Bulmer 1988). It seems likely that in both cases the preferential use of certain synonymous sense or stop codons is a strategy to maximize the rate of translation or to minimize the costs of translational errors (Sharp and Bulmer 1988; Kurland 1991; Bulmer 1991). However, since selection on synonymous codon bias is thought to be quite weak (Bulmer 1991), any conflict with other selection pressures can lead to the use of suboptimal synonymous codons. For instance, synonymous codon use is less biased at the start of *E. coli* genes, probably because of constraints imposed by ribosome binding (Eyre-Walker and Bulmer 1993).

Many genes in bacteria are in the same orientation and very close together, if not actually overlapping; for example, the last base pair of *hisC* is the first base pair of *hisB* in *E. coli*. This could lead to a conflict of interest since the region upstream from the start codon is involved in binding the ribosome (Gold 1988). Although the requirements for ribosome binding are not yet fully characterized, two factors, of relevance here, are thought to be involved. First, there is a sequence, the Shine-Dalgarno (S-D) sequence, which binds the end of 16S rRNA component of the ribosome (Gold 1988). This sequence is generally required for gene expression. Second, it seems likely that there is an avoidance of mRNA

* *Present address:* Department of Plant Science, Foran Hall, Cook Campus, Rutgers University, New Brunswick, NJ 08903, USA

secondary structure around the start codon, since the introduction of secondary structure at the start of *E. coli* gene substantially reduces expression (de Smit and van Druin 1990; Vellanoweth and Rabinowitz 1992). As I will show below, many genes are sufficiently close together that the end of the one gene impinges on the region upstream from the next gene on the chromosome where the S-D sequence is located, and where selection is likely to be acting against the formation of mRNA secondary structure.

In this paper I will show that synonymous codon use is less biased in favor of those codons preferred by highly expressed genes at the end of *E. coli* genes. An investigation will be made into to what extent this can be attributed to the close proximity of genes. I will also briefly consider whether S-D sequences that are overlapped by a coding sequence are different from those which are not in terms of strength and position.

Materials and Methods

The analyses use two datasets. The first is a compilation of 1,286 complete *E. coli* genes from the Ecoseq6 database (Rudd 1992, pers comm). The second is a set of sequences from Ecoseq5 (Rudd 1992, pers comm) which cover the 100 bases upstream and the 100 bases downstream from the start codon of 1,037 genes.

The codon adaptation index (CAI) was calculated by the method of Sharp and Li (1987) with minor modifications suggested by Bulmer (1988): i.e., let $x_{ik} = \mathrm{Log}_e\, w_{ik}$where w_{ik} is the frequency of codon k of amino acid i in a reference set of highly expressed genes, divided by the frequency of the most commonly used codon of amino acid i. The CAI value for a set of codons is then computed as

$$\mathrm{CAI} = \mathrm{Exp}(\bar{x}) \qquad (1)$$

where $\bar{x}$ is the average value of x_{ik} for the codons considered; 95% confidence intervals for the CAI value were calculated by first computing the confidence interval for $\bar{x}$.

Following Schurr et al. (1993) the probable S-D sequence of a gene was located by searching the 30 bp upstream from the start codon for the sequence that formed the strongest interaction with the last 13 nucleotides of the *E. coli* 16S rRNA sequence (3′-AUUCCUCCACUAG-5′). The strength of the interaction between the mRNA and 16S rRNA was calculated using the rules of Freier et al. (1986) and Jaeger et al. (1989). The calculations were kindly performed by Dr. D.G. Higgins. Terminal mismatches and the 3.4 kcal/mol associated with duplex initiation were included in the calculation of the free energy, whereas dangling ends were not. If two sequences upstream from a gene gave the same free energy upon binding the 16S rRNA sequence one of them was chosen at random. No hairpin loops were allowed in either the mRNA or the 16S rRNA (i.e., the sequences were not allowed to bind to themselves). For each gene a sequence was located that gave a reduction in free energy on binding the 16S rRNA sequence. The distance between the ribosome binding site and the start codon was calculated as the number of nucleotides separating the start codon and the base that would bind the capitalized U in the 16S rRNA sequence 3′-auuccUccacuag-5′. This base appears to be involved in most interactions with the S-D sequence (Schurr et al. 1993).

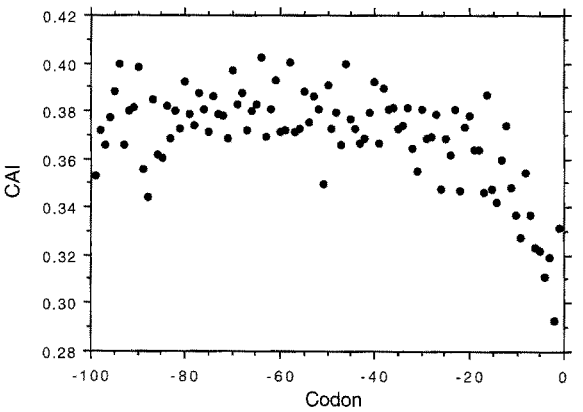

Fig. 1. The CAI value plotted against the distance from the stop codon for the 1,286 complete genes in Ecoseq6.

Results

Synonymous Codon Use

In general *E. coli* tends to use those synonymous codons that bind the commonest tRNAs, with normal base pairing (Ikemura 1985; Bulmer 1988). This is especially the case in highly expressed genes (Ikemura 1985; Bulmer 1988). Figure 1 shows the codon adaptation index (CAI) value over the last 100 codons of 1,286 *E. coli* genes. The CAI value is a measure of how frequently codons preferred by highly expressed genes are used (they are the codons that bind the commonest tRNAs, or that bind the tRNA with normal base pairing); it varies from 1.000 when only those codons that are most favored by highly expressed genes are used to 0.240 when synonymous codon use is random (Eyre-Walker and Bulmer 1993). It is evident that synonymous codon use is less biased in favor of the codons used by highly expressed genes in the last 20 codons of the gene than in the middle.

It seems possible that this decline in synonymous codon bias at the end of genes could be associated with the close proximity of genes oriented in the same direction on the chromosome; many genes may be sufficiently close together so that the end of one gene overlaps either the Shine-Dalgarno sequence or the coding sequence of the next gene on the chromosome. To test this idea we need to know where a gene terminates relative to the next gene on the chromosome and where the S-D sequence of the downstream gene is located. In order to find the position of one gene relative to another on the chromosome the 100 bp upstream and the 100 bp downstream from 1,037 *E. coli* genes were searched for the last 20 bp of 1,286 complete *E. coli* genes. Out of the 1,286 complete gene sequences 389 were found to terminate within 80 bp of the start codon of another gene. The distribution of the intergenic distances is shown in Fig. 2. There is a clear tendency for termination to occur within 20 bp of

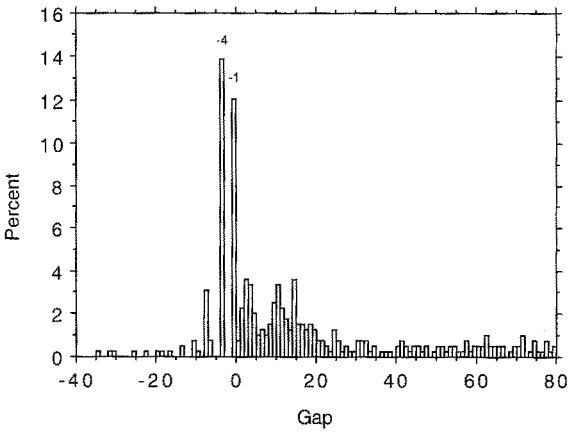

Fig. 2. The distribution of intergenic distances.

the start codon of the next gene, with an overlap between the stop codon of one gene and the start codon of the next gene (the −1 and −4 columns in Fig. 2) being particularly common. It is clear that many genes overlap the S-D or coding sequence of the next gene on the chromosome. The probable position of the S-D sequence upstream from each of the 389 downstream genes was located by searching the 30 bp upstream from each gene for the sequence that formed the most stable interaction with the last 13 bp of the 16S rRNA sequence (see Methods).

Table 1 shows the CAI value of those codons at the end of the upstream gene that overlap either the coding or S-D sequence of the downstream gene, and those codons that lie upstream and downstream from the S-D sequence; only those genes that overlap the S-D sequence are considered. It is evident that codon bias is substantially reduced in the part of the gene that overlaps the S-D or coding sequence; in fact the CAI value is significantly below the value one would expect if codon usage were random (0.24 in *E. coli;* Eyre-Walker and Bulmer 1993). This is probably because the perfect complement to the 16S rRNA sequence (5′-TAAGGAGGTGATC-3′) only has two optimal codons in the three potential reading frames, GGT and ATC, but three very rare codons, GGA and AGG (twice). Note that the GGA is complementary to the most commonly bound part of the 16S rRNA sequence, the central 3′-CCUC-5′ (Schurr et al. 1993). Any gene that overlaps the S-D sequence will therefore tend to show low codon bias at its 3′ end.

However, overlaps of the upstream gene with the S-D and coding sequence of the downstream gene do not appear to be the sole cause of the low codon bias at the end of the gene; there is still a significant trend in codon bias even when the overlapping codons are removed from the analysis (Fig. 3, Spearman's rank correlation $r_s = 0.305$, $P < 0.0028$). Note that the sample sizes are different in Figs. 1 and 3, in part explaining why any trend in codon bias is less obvious in the latter. Interestingly, this low codon bias appears to be

Table 1. The CAI value of codons whose third position overlaps an S-D sequence or another coding sequence

	CAI	95% confidence interval
4 codons 3′ to S-D	0.363	0.316–0.417
Overlapping S-D or distal gene	0.196	0.171–0.225
4 codons 5′ to S-D	0.336	0.306–0.370

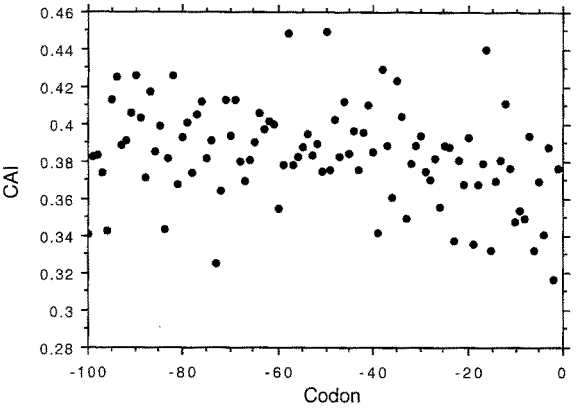

Fig. 3. The CAI value plotted against the distance from the stop codon for the 389 genes that terminate within 80 bp of another gene excluding codons that overlap the S-D or coding sequence of the downstream gene.

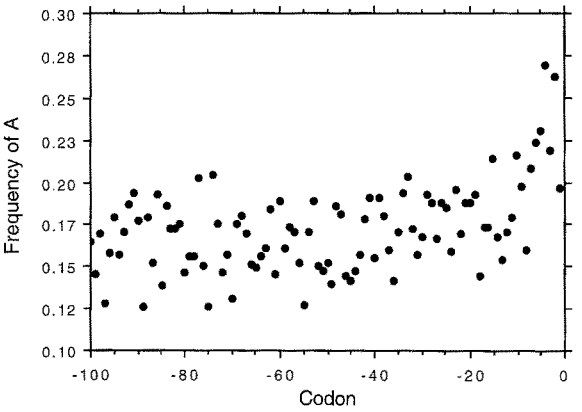

Fig. 4. The frequency of A at the third codon position, at the end of the 389 *E. coli* genes, excluding those codons that overlap either the S-D or coding sequence of the next gene on the chromosome. The stop codon is omitted.

associated with some trends in composition. Figure 4 shows the frequency of A in the third codon position over the last 100 codons, excluding those codons that overlap either the S-D or coding sequence of the next gene on the chromosome. It is clear that the frequency of A increases toward the end of the gene particularly over the last five codons ($r_s = 0.374$, $P = < 0.0005$). There is also a less obvious decrease in the frequency of G ($r_s = 0.328$, $P < 0.002$), but no discernible changes in the frequency of C ($r_s = 0.174$, $P < 0.09$) and T ($r_s = 0.072$, $P < 0.5$). Interestingly, the trends in A and G are also

Table 2. The frequency of A relative to the stop codon, and to the start codon of the next gene[a]

	Within 15 bp of start codon of next gene	>15 bp away from start codon of next gene
Within 5 codons of stop codon	0.274 (0.020)	0.219 (0.013)
Between 6 and 30 codons from stop codon	0.301 (0.045)	0.181 (0.004)

[a] Figures in parentheses are standard errors. The stop codon and codons overlapping the S-D or coding sequence of the distal gene were excluded

detectable at the first and second codon positions (results not shown).

There are two general explanations for the increase in the frequency of A: either it could be associated with the end of the gene itself—for instance, there might be signals at the end of the gene important for efficient termination (Brown et al. 1990; Arkov et al. 1993)—or it might be a consequence of forces acting at the start of the next gene: for instance, the avoidance of mRNA secondary structure around the S-D sequence (Gold 1988; Eyre-Walker and Bulmer 1993). To differentiate between these two possibilities a 2×2 table was constructed for those codons that did not overlap either the S-D or coding sequence of the next gene on the chromosome. Codons were classified as being either within 5 codons, or between 6 and 30 codons of the stop codon; and within 15 bp, or more than 15 bp from the start codon of the next gene. The dataset was also divided into two groups according to CAI value (less than 0.37 and greater than 0.37) to ensure that gene expression level was not a confounding factor. However, there was no significant difference between the contingency tables for the two groups of genes ($\chi^2 = 7.565$, $df = 4$, $P > 0.10$), so the tables were combined to give Table 2. Table 2 shows that it is the close proximity of genes that is largely responsible for the high frequency of A at the end of the genes; i.e., codons that are near the start codon of the next gene have a high frequency of A whether or not they are near the end of the gene. However, there might also be an increase in A associated with the proximity of the stop codon, since in those codons that are further than 15 bp from the start codon of the next gene, the codons that are within five codons of the stop codon have a higher frequency of A in the third codon position than those codons that are further away ($z = 2.79$, $P < 0.006$). This is also true for codons that are further than 30 bp or 45 bp from the start codon of the next gene (results not shown). It is important to appreciate that the changes in composition are likely to be causing the reduction in codon bias rather than vice versa, since a reduction in codon bias need not lead to a change in composition (see Sharp 1989).

Stop Codon Use

E. coli uses the stop codon TAA in preference to TGA, especially in highly expressed genes, and TAG is used only very rarely (Sharp and Bulmer 1988). Sharp and Bulmer suggested that this might be due to selection upon the termination rate, or against translational read-through. In this section I explore another factor: Stop codon use reflects the use of stop codons that have other simultaneous coding functions. For instance, 54 genes in the sample overlap the next gene by 4 bp; this constrains them to terminate with TGA. Likewise those genes that overlap by 1 bp must generally terminate either with a TGA or a TAA since most genes start ATG. The 16S rRNA sequence that binds the S-D sequence is 3′AUUCCUCCACUAG-5′. This can bind either TAA or TGA at the 5′end of the S-D sequence (note that U binds both A and G) or a TGA 2 bp from the 3′ end (Shine and Dalgarno 1974). Table 3 lists the stop codon use at each of these sites. Interestingly, at least 55% of TGA stop codons appear to be constrained to be TGA because they overlap the coding sequence (excluding the 1-bp overlaps) or the S-D sequence (excluding those that bind the 3′ end of 16S rRNA). Similarly, 4 out of 16 TAG stop codons are constrained by overlapping the coding sequence.

With so many stop codons having other functions, it is tempting to suggest that much of stop codon use can be understood in terms of conflicting pressures and the distance between genes; i.e., TAA is the preferred codon (because of faster termination or lower levels of translational read-through), and TGA and TAG are only used when the sequence has other coding functions. This causes TGA to be used more often than TAG because of the tendency for genes to terminate either in the S-D sequence or the start codon of the distal gene. TAA increases in frequency with increasing gene expression level because highly expressed genes do not tend to overlap the next gene on the chromosome (Eyre-Walker 1995). To test the latter parts of this hypothesis (i.e., [1] that TGA is used in preference to TAG because of the high frequency of 4-bp overlaps and [2] that TAA use increases with CAI value because the distance between genes is related to the CAI value [Eyre-Walker 1995]), the dataset of upstream genes was divided into three groups according to their CAI value: less than 0.35, greater than 0.35 but less than 0.50, and greater than 0.50. Table 4 shows the stop codon use for those genes that terminate upstream from the S-D sequence. It is evident that TGA is used in preference to TAG even in those genes that do not overlap either the S-D or coding sequence of the distal gene. Furthermore, the use of TAA is still correlated to the CAI value of the gene. Furthermore, the use of TAA is still correlated to the CAI value of the gene. Overlaps between stop codons and the S-D or coding sequence of the next gene on the chromosome

Table 3. The number of stop codons at various positions of known function

Stop codon	Overall	5′ end of S-D	3′ end of S-D	1-bp overlap	4-bp overlap	Overlap >4 bp (231 frame)	Overlap >4 bp (312 frame)
TAA	232	25	0	28	0	3	10
TGA	141	5	6	18	54	0	12
TAG	16	0	0	1	0	3	1

Table 4. Percentage stop codon use in genes that terminate upstream from the S-D sequence of the distal gene, expressed in percents[a]

	High CAI	Medium CAI	Low CAI
TAA	84.6 (33)	79.0 (49)	61.1 (33)
TGA	12.8 (5)	16.1 (10)	31.5 (17)
TAG	2.6 (1)	4.8 (3)	7.4 (4)

[a] Figures in parentheses are the actual stop codon use

are therefore not solely responsible for the biases in *E. coli* stop codon use.

The Strength of the Ribosome Binding Site

Since many genes overlap the S-D sequence of the next gene it is of interest whether S-D sequences that also code for part of a protein are compromised in any way in terms of strength and position relative to the start codon. The strength in this case is the estimated free energy of the structure that forms between the 16S rRNA and the S-D sequence, and the distance is the number of bases between the start codon and one of the central bases in the 16S rRNA sequence (see Materials and Methods).

The strength of the ribosome binding site is slightly, but significantly, greater for those S-D sequences that overlap the upstream gene (−6.74 ± 0.15 kcal/mol) than for those that do not (−5.97 ± 0.16 kcal/mol) ($t = 3.41, P < 0.0007$), whereas the distance between the start codon and the S-D sequence is slightly greater (11.24 ± 0.41 against 9.48 ± 0.39, t = 2.98, $P < 0.003$). However, in both cases the differences are small relative to the overall spread of values (they amount to less than half a standard deviation, which is ~2 kcal/mol for the energy, and ~5 bp for the distance) so the S-D sequences that code for part of a protein are not very different from those that do not.

Discussion

It has been shown that synonymous codon use is less biased in favor of codons preferred by highly expressed genes at the end of *E. coli* genes and that this is, at least in part, due to the fact that many *E. coli* genes terminate very close to the next gene on the chromosome; in doing so they overlap either the coding sequence or sequences just upstream from the start codon that are known to be involved in gene expression. It seems likely that Lawrence and Hartl (1991) did not observe this trend in codon bias because they used tenfold-fewer genes. It has also been shown that many genes use TGA or TAG stop codons because they overlap a sequence with another function. In particular many genes overlap the next gene by 4 bp, which constrains them to terminate with TGA.

Since a proportion of genes terminate in the opposite orientation to the next gene on the chromosome it is possible that some of the low synonymous codon bias at the end of *E. coli* genes is caused in part by the end of one gene overlapping the end of another gene or sequences downstream from another gene important for termination (of transcription or translation). To investigate this possibility the 1,286 complete gene sequences were searched for perfect matches to the complement of the last 15 bp of each complete gene. Only four gene pairs were found, so overlaps, between the ends of coding sequences at least, seem an unlikely explanation of the low codon bias at the end of *E. coli* genes.

The changes in composition at the end of genes are consistent with selection against the formation of mRNA secondary structure around the start codon of the next gene on the chromosome. As Eyre-Walker and Bulmer (1993) point out, A is likely to be the most favored base in such regions since it only binds U weakly, whereas G is likely to be the least favored base since it binds U weakly and C strongly. The origins of the increase in the frequency of A associated with the proximity to the stop codon (rather than to the start codon of the next gene) are less evident. It might be associated in some way with termination (Brown et al. 1990; Arkov et al. 1993).

The preferential use of certain synonymous and stop codons in *E. coli* appears to be due to selection, either in favor of rapid elongation or termination rate or against translational errors. However, this leaves a problem—why "suboptimal" synonymous and stop codons are ever used. There are four main possibilities. First, the use of suboptimal stop codons could be a strategy to regulate gene expression (Grosjean and Fiers 1982; Konigsberg and Godson 1983). This seems unlikely since initiation appears to be rate limiting in translation (Bulmer 1991). Second, there could be a balance between selection and

mutation, with both forces sufficiently strong to make genetic drift inconsequential. This hypothesis predicts very high levels of synonymous site polymorphism (Shields 1990) that are not observed in *E. coli* (e.g., see Hall and Sharp 1991; Guttman and Dykhuizen 1994). Third, selection in favor of the optimal codons could be sufficiently weak so that genetic drift counteracts the effect of selection (Sharp 1989). And fourth, there could be other conflicting selection pressures acting upon synonymous codon bias; for instance, Eyre-Walker and Bulmer (1993) have recently shown that suboptimal synonymous codons are used more frequently at the start of the genes because of conflicting selection pressures, probably associated with ribosome binding. The findings in this paper support a role for the fourth of these alternatives, particularly in the case of stop codon use, where over 50% of the TGA codons of genes that terminate within 80 bp of another gene in the same orientation have another easily identifiable coding function.

Acknowledgments. I am very grateful to Kenn Rudd, who provided the datasets used in this study; to Des Higgins, who performed the calculations of Shine-Dalgarno strength; to Michael Bulmer and Ken Wolfe for helpful discussions; and to Holly Hilton for her comments on this manuscript.

References

Arkov AL, Korolev SV, Kisselev LL (1993) Termination of translation in bacteria may be modulated via specific interaction between peptide chain release factor 2 and the last peptidyl-tRNA$^{Ser/Phe}$. Nucleic Acids Res 21:2891–2897

Brown CM, Stockwell PA, Trotman CNA, Tate WP (1990) The signal for the termination of protein synthesis in procaryotes. Nucleic Acids Res 18:2079–2086

Bulmer M (1988) Are codon usage patterns in unicellular organisms determined by a mutation-selection balance? J Evol Biol 1:15–26

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

de Smit MH, van Druin J (1990) Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. Proc Natl Acad Sci USA 87:7668–7672

Eyre-Walker A, Bulmer M (1993) Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res 21: 4599–4603

Eyre-Walker A (1995) The distance between *Escherichia coli* genes is related to gene expression levels. J Bacteriol 177:5368–5369

Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T, Turner DH (1986) Improved free-energy parameters for predictions of RNA duplex stability. Proc Natl Acad Sci USA 83:9373–9377

Guttman DS, Dykhuizen DE (1994) Detecting selective sweeps in naturally occurring *Escherichia coli.* Genetics 138:993–1003

Gold L (1988) Posttranscriptional regulatory mechanisms in *Escherichia coli.* Ann Rev Biochem 57:199–233

Grosjean H, Fiers W (1982) Preferential codon usage in prokaryotic genes: the optimal codon-anti-codon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18:199–209

Hall BG, Sharp PM (1991) Molecular population genetics of *Escherichia coli:* DNA sequence diversity at the *celC, crr,* and *gutB* loci of natural isolates. Mol Biol Evol 9:654–665

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Jaeger JA, Turner DH, Zuker M (1989) Improved prediction of secondary structures for RNA. Proc Natl Acad Sci USA 86:7706–7710

Konigsberg WJN, Godson GN (1983) Evidence for the use of rare codons in the *dnaG* gene and other regulatory genes of *Escherichia coli.* Proc Natl Acad Sci USA 80:687–691

Kurland CG (1991) Codon bias and gene expression. FEBS Lett 285: 165–169

Lawrence JG, Hartl DL (1991) Unusual codon bias occurring within insertion sequences in *Escherichia coli.* Genetica 84:23–29

Rudd KE (1992) Alignment of *E. coli* DNA sequences to a revised integrated genomic restriction map. In: Miller J (ed) A short course in bacterial genetics: a laboratory manual and handbook for *Escherichia coli* and related bacteria. Cold Spring Harbor Press, Cold Spring Harbor

Schurr T, Nadir E, Margalit H (1993) Identification and characterization of *E. coli* ribosomal binding sites by free energy computation. Nucleic Acids Res 21:4019–4023

Sharp PM (1989) Evolution at silent-sites in DNA. In: Hill WG, McKay TFC (eds) Evolution and animal breeding: reviews of molecular and quantitative approaches in honour of Alan Robertson. CAB International, Wallingford, p 23

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

Sharp PM, Bulmer M (1988) Selective differences among translation termination codons. Gene 63:141–145

Shields DC (1990) Switches in species-specific codon preferences: the influence of mutation biases. J Mol Evol 31:71–80

Shine J, Dalgarno L (1974) The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc Natl Acad Sci USA 71:1342–1346

Vallanoweth RL, Rabinowitz JC (1992) The influence of ribosome-binding-site elements on the translational efficiency in *Bacillus subtilis* and *Escherichia coli in vivo.* Mol Microbiol 6:1105–1114