Molecular Microbiology (1998) **29**(6), 1341–1355

Codon usages in different gene classes of the *Escherichia coli* genome

Samuel Karlin,[1]* Jan Mrázek[1] and Allan M. Campbell[2]
[1]*Department of Mathematics, and*
[2]*Department of Biological Sciences, Stanford University, Stanford, CA 94305-2125, USA.*

Summary

A new measure for assessing codon bias of one group of genes with respect to a second group of genes is introduced. In this formulation, codon bias correlations for *Escherichia coli* genes are evaluated for level of expression, for contrasts along genes, for genes in different 200 kb (or longer) contigs around the genome, for effects of gene size, for variation over different function classes, for codon bias in relation to possible lateral transfer and for dicodon bias for some gene classes. Among the function classes, codon biases of ribosomal proteins are the most deviant from the codon frequencies of the average *E. coli* gene. Other classes of 'highly expressed genes' (e.g. amino acyl tRNA synthetases, chaperonins, modification genes essential to translation activities) show less extreme codon biases. Consistently for genes with experimentally determined expression rates in the exponential growth phase, those of highest molar abundances are more deviant from the average gene codon frequencies and are more similar in codon frequencies to the average ribosomal protein gene. Independent of gene size, the codon biases in the 5′ third of genes deviate by more than a factor of two from those in the middle and 3′ thirds. In this context, there appear to be conflicting selection pressures imposed by the constraints of ribosomal binding, or more generally the early phase of protein synthesis (about the first 50 codons) may be more biased than the complete nascent polypeptide. In partitioning the *E. coli* genome into 10 equal lengths, pronounced differences in codon site 3 G+C frequencies accumulate. Genes near to oriC have 5% greater codon site 3 G+C frequencies than do genes from the ter region. This difference also is observed between small (100–300 codons) and large (>800 codons) genes. This result contrasts with that for eukaryotic genomes (including human, *Caenorhabditis elegans* and yeast) where long genes tend to have site 3 more AT rich than short genes. Many of the above results are special for *E. coli* genes and do not apply to genes of most bacterial genomes. A gene is defined as alien (possibly horizontally transferred) if its codon bias relative to the average gene exceeds a high threshold and the codon bias relative to ribosomal proteins is also appropriately high. These are identified, including four clusters (operons). The bulk of these genes have no known function.

Received 13 February, 1998; revised 8 April, 1998; accepted 8 June, 1998. *For correspondence. E-mail fd.zgg@forsythe.stanford.edu; Tel. (650) 723 2204; Fax (650) 725 2040.

Introduction

The nature of gene codon choices varies considerably from organism to organism; for recent reviews, see Andersson and Kurland (1990), Kurland (1993) and Sharp and Matassi (1994). Pronounced differences in codon usage raise questions about the mechanisms producing and maintaining these differences. Codon preferences have been considered from two main perspectives: (i) translational efficiency and fidelity; (ii) selective and non-selective substitutional biases operating during DNA transcription, replication and repair processes. Variations in tRNA availabilities is proffered by several authors as a key factor in producing codon bias of the 'highly expressed genes' during the rapid growth phase of *Escherichia coli* (e.g. Andersson and Kurland, 1990). Codon/anticodon interaction strength (e.g. Kurland, 1993), site-specific codon biases (e.g. Maynard-Smith and Smith, 1996), time of replication (e.g. Deschavanne and Filipski, 1995), codon context (e.g. Irwin *et al.*, 1995) and evolutionary age may also contribute to codon selectivity. Codon choices might usefully be divided by gene sequence location (amino end, carboxyl end, central parts) (Chen and Inouye, 1994) and by protein structural domains and protein secondary structure determinations (Thanaraj and Argos, 1996).

It has been suggested that recently imported genes generally show overwhelmingly deviant codon usage from the host gene repertoire (Médigue *et al.*, 1991; Lawrence and Ochman, 1997). Indeed, Médigue *et al.* (1991) applied clustering techniques based on codon usage to a set of 780 genes of *E. coli* supporting a partitioning of the genes into three classes. Two of the classes correlated with the level of expression of the genes. They suggested that the genes of the third class were introduced into the *E. coli* genome by horizontal gene transfer.

Codon preferences among different gene classes in E. coli

Gene codon usages can be dissected in many ways. In this paper we focus on the following comparisons:

1. Correlations of codon biases with level of gene expression.
2. A means for forming gene classes based on partitioning the genome into 100 kb, 200 kb or longer contigs and then assembling all genes of each contig to define a gene group. In prokaryotic genomes, biases related to location may reflect consequences of replication timing.
3. Codon usages for genes characterized by function and/or cellular localization of gene products. In this context, bacterial genes have been divided into 14 major function/cellular categories (adapted from Riley, 1993), each generally comprising several subclasses. For example, the Translation category includes the subclasses: (i) amino acyl tRNA synthetases (abbreviated tRN); (ii) proteins of degradation (D); (iii) genes of protein modifications (M); (iv) ribosomal proteins (RP). Group (i) generally consists of about 20 genes and group (iv) includes about 50–70 genes. Group (iii) includes elongation factors (tufB, fusA), DNA-processing chain A (dprA), ribosome-releasing factor (frr), initiation factor 3 (infC), adenylyl-transferase (glnE), peptide chain release factor 3 and translation factor (dsbB), among others. Group (ii) centres on proteinases (e.g. sohB, pepN, pepT, pepE, sppA, clpP, clpX).
4. Correlations related to gene size.
5. Genes (longer than 100 codons) might be divided into three parts, the amino third, the middle third, and the carboxyl third. We then form the amino codon collection, the middle codon collection, and the carboxyl codon collection and investigate the nature of codon usages for these three ‘gene’ groups.
6. Comparisons generated by forming k (e.g. $k = 2$, 3, 5, 10) clusters dividing all genes by similarity of codon usage (in 61-dimensional space corresponding to the 61 sense codons), or alternatively by similarity of amino acid usage or by similarity of a reduced set of amino acids or codons.
7. Codon usage related to genes encoded from the leading versus lagging strand (see Lobry, 1996; Francino and Ochman, 1997; Mrázek and Karlin, 1998).
8. Characterizations of ‘alien’ genes (possibly laterally transferred or part of pathogenicity island or unusual in other respects).

Methods

Measures of relative codon biases

Codon adaptation index. A quantitative measure proposed for assessment of codon bias is the codon adaptation index (CAI, Sharp and Li, 1987). This specifies a reference set of genes, almost invariably a set, H, chosen from among ‘highly expressed genes.’ Defining $w_{xyz} = f^{H}_{xyz}/\max_{xyzEa} f^{H}_{xyz}$ as the ratio of the frequency of the codon (xyz) to the maximal codon frequency in H for the same amino acid a, the CAI of a gene of length L is taken as $(\Pi^{L}_{i=1} w_i)^{1/L}$ (the log average) where i refers to the i^{th} codon of the gene and w is calculated as above. High values (near 1) of CAI correlate with high expression levels. Classification of genes according to their CAI values has been carried out in several publications (e.g. see Bulmer, 1988; Eyre-Walker, 1996). Genes that are known (experimentally) to be highly expressed, at least during cellular fast growth, include most ribosomal protein genes and genes coding for elongation factors (tuf and fus) and some membrane genes (Dong *et al*., 1996). However, not all ribosomal proteins have a high CAI value (for examples, see Parker, 1992; Eyre-Walker, 1996).

Codon usage differences between gene classes. We introduce a versatile way to assess bias of one group of genes (or a single gene) relative to a second group of genes. Let C be a family of genes with codon frequencies $c(x, y, z)$ for codon xyz normalized such that for each amino acid codon family $\Sigma_{(x,y,z)=a}\, c(x, y, z) = 1$ where the sum extends over all codons (x, y, z) translated to amino acid a. Let $F = \{f(x, y, z)\}$ indicate the corresponding codon frequencies for the gene family F, again normalized to 1 in each amino acid codon family. We assess the codon usage difference of the gene family F relative to the gene family C by the formula

$$B(F|C) = \sum_{a} p_a(F)\left(\sum_{(x,y,z)=a} |f(x,y,z) - c(x,y,z)|\right) \qquad (1)$$

where $\{p_a(F)\}$ is the set of amino acid frequencies of the genes of F. The maximum possible value for $B(F|C)$ is 2.00, but rarely does $B(F|C)$ exceed 0.5. Codon usage differences between two gene families generally range from 0.05 to 0.300.

Notice the asymmetry of $B(F|G)$ in that only the amino acid frequencies of F appear as weights. We refer to the gene collection C as the standard to which different gene groups $F^{(1)}, F^{(2)}, \ldots, F^{(r)}$ are compared. C could be a specific gene class, an average gene meaning that C is the set of all genes of the genome, or an absolute set of frequencies (e.g. $c(x, y, z) = 1/d(x, y, z)$, the reciprocal of the degeneracy count of the codon $(x, y, z) =$ random). F could be a single gene. A symmetric version of (1) for the two sets of genes F and G is $B^*(F, G) = [B(F|G) + \mathrm{B}(G|F)]/2$. In general, $B(F|G)$ and $B(G|F)$ differ little (see Tables 3, 5, 7 and 8), indicating that differences in amino acid usages between F and G have little influence on the calculated relative codon biases.

When C is the set of all genes, then $B(F|C) = B(F|\text{all})$ measures the codon usage difference of the class of genes F from the average gene and we refer to $\underline{B(F|\text{all})}$ as the codon bias (CB) of F. When no ambiguity is likely, we refer to $B(F|G)$ as the codon bias of F with respect to G. The assessments implied by (1) can be made for any two gene groups from the same genome or from different genomes. The formula (1) can also be applied to a subset of amino acids (e.g. a restricted to hydrophobic, charge or aromatic types).

For $C = H$ 'the set of highly expressed genes' and with F a single gene, formula (1) can be construed as a species-specific measure of the 'fitness' of a gene in terms of its codon usage. Figure 1 plots the CAI index versus CB determinations [formula (1)] for all individual genes relative to the standard of ribosomal protein (RP) genes and to the average gene. Clearly CAI and CB values are inversely correlated with respect to RP genes. In assessing CAI and CB values relative to the average gene, we obtain a central mass flanked by two horns corresponding to 'alien' and 'highly expressed' genes; compare with Fig. 2.

Codon biases in terms of gene expression level in *E. coli* are detailed below.

Results

Codon usage differences among translation genes

Among these gene classes (subgroups of the translation functional category), RP by a factor of at least two comprise the most biased genes relative to the average *E. coli* gene (Table 1). Codon usages for each of the five classes are manifestly non-random with RP genes featuring the most non-random group (last column). Codon usage of the average *E. coli* gene deviates most from the RP class compared with the Ch (chaperonins), D, M and the tRN gene groups. On the other hand, codon choices are reasonably similar for tRN and modification (M) genes, codon bias $B(\text{tRN}|\text{M}) \approx 0.088$.

Comparing codon usages of ribosomal proteins (RP) and amino acyl tRNA synthetase (tRN) genes (Table 2)

Most ribosomal proteins are highly expressed during the *E. coli* exponential growth phase. Similarly, during fast growth, most genes facilitating translation are expected to be 'highly expressed' and essential. These include genes for aminoacyl tRNA synthetases (tRN), and other gene classes [e.g. genes for protein modification (M) and degradation functions (D), and chaperonins (Ch), although the last group is primarily inducible under appropriate cellular conditions].

It is useful to compare the overall RPs and tRN codon frequencies per amino acid and indicate by an asterisk reversals to the average codon frequency tendencies.

The foregoing tabulation indicates that codon preferences can differ significantly among highly expressed genes relative to the average gene for certain amino acid types. Note especially the disparities for alanine, aspartate, histidine, isoleucine, phenylalanine, threonine and valine amino acids.

Codon bias and gene expression level in E. coli

It is accepted dogma, at least for *E. coli* genes, that there is a high correlation between 'optimal codons' and level of gene expression. Ninety-seven protein abundances of *E. coli* genes during the exponential growing phase have been experimentally determined (VanBogelen *et al*., 1996). These genes include five ribosomal proteins, 12 aminoacyl tRNA synthetase genes, RecA, elongation factor Tuf, the heat shock protein 70 homologue DnaK, several OMP genes and others. All these 97 proteins have high CAI values (≥0.4). A means to standardize the protein abundances is to divide the abundance value by protein molecular weight giving molar abundances and create 10 molar abundance groups of about 10 genes each. We label these bin groups a, b, c, . . . , j, where a consists of those genes among the 97 of lowest molar abundance level up to j, the genes of highest molar abundance level. Codon biases were evaluated for each molar abundance group relative to an average *E. coli* gene (All) and codon usage differences to subclasses of genes of the translation category, the RP genes and the group of aminoacyl tRNA synthetases (tRN) (Table 3). The following conclusions stand out. (i) Relative to the totality of *E. coli* genes, codon biases (read down the column marked *All*

Table 1. Relative codon bias with respect to five subgroups of genes and with respect to the average of (All) *E. coli* genes (values multiplied by 1000).

F\C	All *E. coli*	Ch.	D	M	tRN	RP	Random
All	•	237	190	259	297	530	377
Chaperone (abbreviated Ch)	233	•	201	145	197	328	444
Degradation (D)	184	197	•	112	131	380	526
Modification (M)	249	137	115	•	88	281	518
tRNA-synthetase (tRN)	284	195	129	86	•	271	557
Ribosomal proteins (RP)	520	326	399	289	289	•	709
Random (R)	395	460	562	560	626	768	•

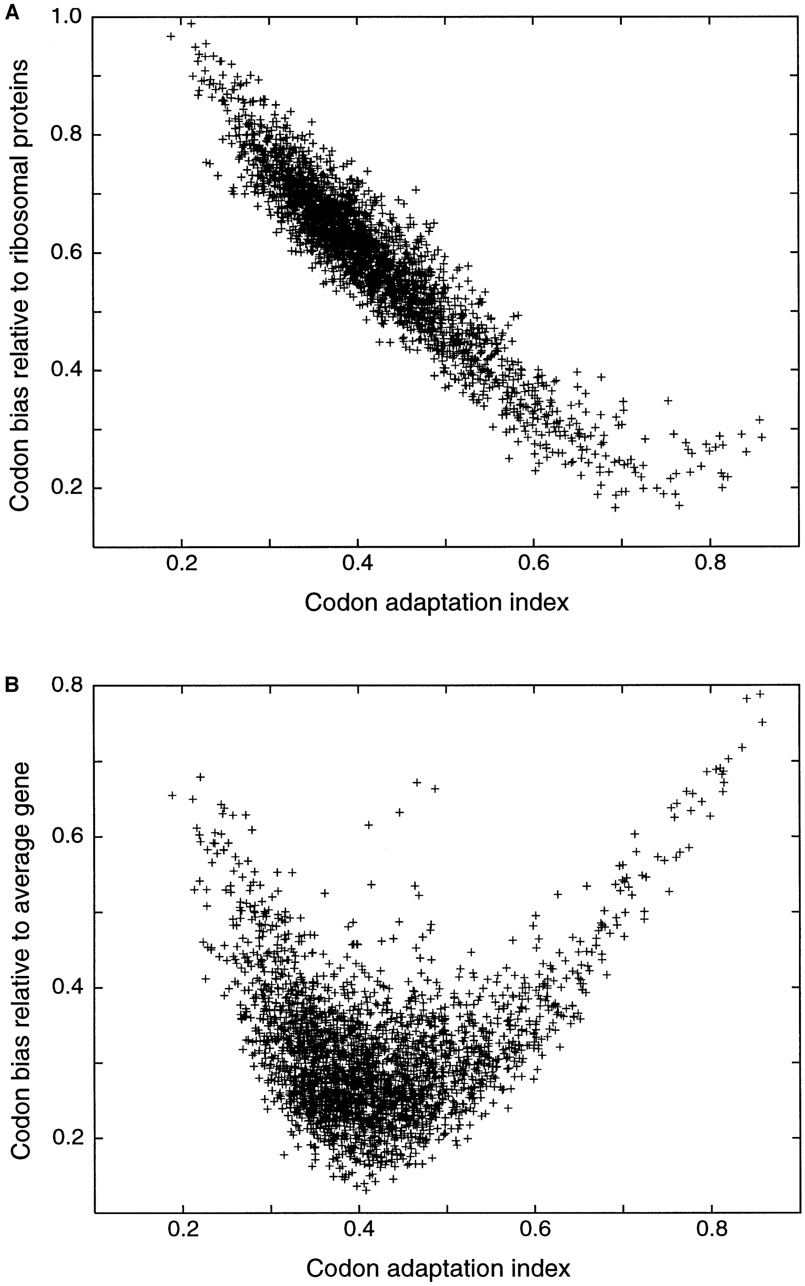

Fig. 1. Each *E. coli* gene of ≥ 200 codons is represented by a point with co-ordinates corresponding to its codon adaptation index and to the codon usage difference (codon bias) relative to ribosomal proteins (A) or codon bias relative to the average gene (B). Codon adaptation index is calculated using ribosomal protein genes as a reference set of highly expressed genes.

of Table 3) generally increase with increasing protein molar abundance. Thus, the greatest deviation from the general codon frequencies of the totality of *E. coli* proteins occurs for the highest expression level genes of bin(j) (B(j|All) = 545 compared with 100 for genes of bin(a). When comparing with the codon frequencies of ribosomal proteins, the codon usage differences generally decrease with respect to protein molar abundance. However, a negative correlation in the degree of protein molar abundance (bins a–j) relative to the tRN genes or other subclasses of the translation functional category is not evident. Thus, the molar abundance and codon usage differences unequivocally correlate negatively with the RP gene family and positively with the average *E. coli* gene respectively. Contrary to the case of the RP gene group, the highest and second highest molar abundance gene group (i and j) are most deviant in codon frequencies compared with the gene group tRN.

The 10 gene molar abundance classes were also evaluated for codon bias with respect to various subsets of amino acids. Codon frequency patterns observed for amino acids of codon degeneracy 2 ending with a pyrimidine (NNY) or

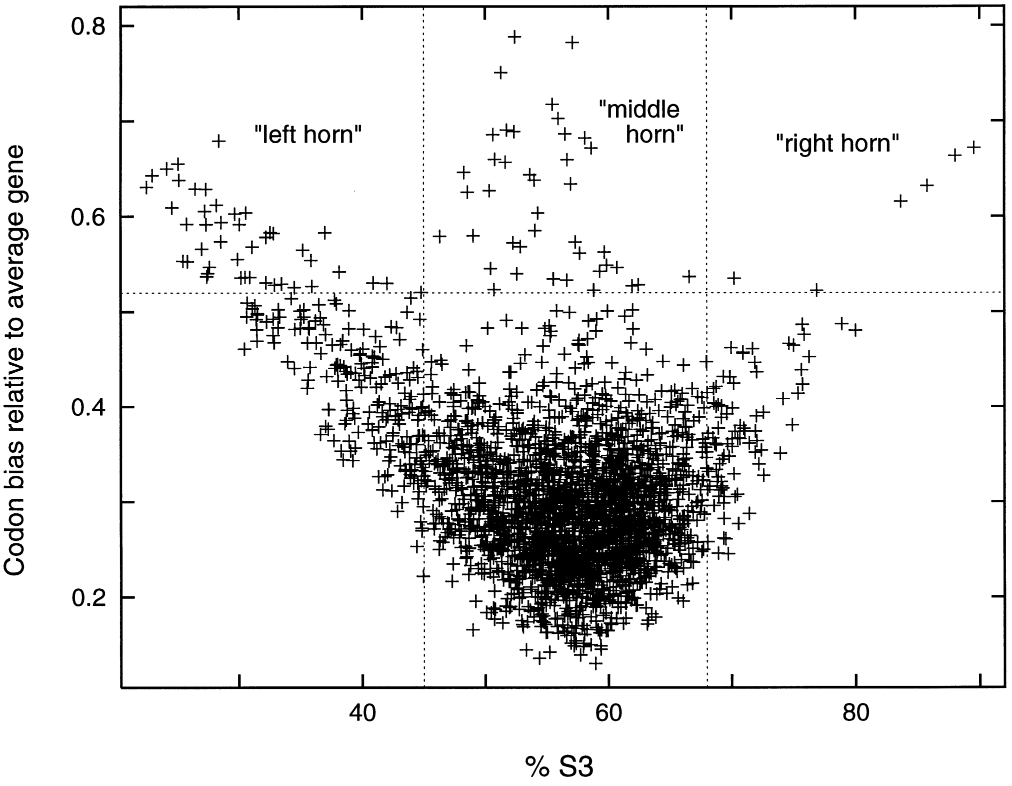

Fig. 2. Each *E. coli* gene of ≥ 200 codons is represented by a point with co-ordinates corresponding to its codon bias relative to the average gene and S3 percentage (G+C content of codon site 3).

amino acids of codon degeneracy 4, or amino acids of codon degeneracy 6, or restricted to hydrophobic amino acids, charged or amide hydrophilic amino acids parallel the calculations relative to all amino acids (data not shown).

Variation in codon biases around the genome (Table 4)

Comparisons are made to the average *E. coli* gene (All), to the RP gene collection and to the set of tRN genes. For most windows the codon bias is small <0.100. The largest bias among genes of the leading strand is shown in window 3L, which may be attributed to a cluster of 26 ribosomal protein genes. This is consistent with the result that this contig shows the smallest codon usage difference (0.430) relative to RP genes. A singular anomalous window, 2R, shows codon bias 0.116 for the lagging strand.

Table 2. Highest codon fractions. An asterisk indicates opposite codon usage compared with the average gene (all *E. coli*).

	All *E. coli*	tRN (tRNA-synthetase)	RP (ribosomal proteins)
Ala	GCG 35%, GCC 26%	GCG 39%	*GCT 47%, GCA 27%
Arg	CGC 40, CGT 38	*CGT 59, CGC 37	*CGT 67, CGC 31
Asn	AAC 55	AAC 78	AAC 86
Asp	GAT 63	*GAC 50.5	*GAC 62
Cys	TGC 56	TGC 68	TGC 75
Gln	CAG 65	CAG 80	CAG 76
Glu	GAA 69	GAA 71	GAA 75
Gly	GGC 40, GGT 34	GGC 46, GGT 44	*GGT 59, GGC 38
His	CAT 57	*CAC 63	*CAC 72
Ile	ATT 50, ATC 42	*ATC 58, ATT 41	*ATC 73, ATT 25
Leu	CTG 49.5	CTG 73	CTG 83
Lys	AAA 76.5	AAA 77	AAA 70
Phe	TTT 57.4	*TTC 66	*TTC 75
Pro	CCG 52.6	CCG 73	CCG 69
Ser	AGC 28 uniform otherwise	AGC 29, TCT 26, TCC 21	*TCT 38, TCC 26, AGC 26
Thr	ACC 54	ACC 54	*ACT 47, *ACC 43
Tyr	TAT 57	*TAC 58	*TAC 75
Val	GTG 37, GTT 26	GTT 34, GTG 33	*GTT 51, GTA 27

Table 3. Relative codon usage bias (multiplied by 1000) among E. coli gene bins by expression level.[a]

{*F*}\{*G*}	a	b	c	d	e	f	g	h	i	j	All	tRN	RP	rnd
a	*	226	272	210	205	247	268	334	435	495	100	231	462	445
b	221	*	92	75	99	61	121	141	247	283	276	78	294	547
c	271	95	*	99	111	83	105	124	196	253	337	86	251	602
d	204	77	97	*	71	93	122	165	266	305	266	53	321	560
e	205	101	109	74	*	104	111	146	249	307	266	72	299	553
f	244	64	83	94	103	*	98	112	213	260	302	69	261	580
g	266	124	105	125	113	95	*	88	192	243	316	100	225	547
h	333	143	122	166	147	109	90	*	166	194	383	133	204	592
i	432	255	200	271	253	214	192	165	*	114	488	238	138	689
j	493	295	262	311	319	264	248	197	118	*	545	286	149	726

a. A total of 97 genes with experimentally determined expression levels under exponential growth conditions (VanBogelen *et al*., 1996) were arranged into 10 bins labelled a (lowest expression level) to j (highest expression level) according to their molar abundance. Each gene ensemble {*F*} was compared with a standard {*G*} using formula (1). In addition, the bins were compared with the ensemble of all *E. coli* genes (labelled All), amino acyl tRNA synthetases (22 genes, labelled tRN), and ribosomal proteins (55 genes, labelled RP). The rightmost column (rnd) indicates codon bias relative to a random standard (equal frequencies of all codons).

This region contains many hypothetical open reading frames (ORFs). However, in general, the *E. coli* genome is homogeneous with relatively weak codon biases among the genes distributed over the genome. Bidirectional replication proceeds along with distance from oriC, and it appears that codon bias hardly depends on timing in the replication cycle except near the ter region (Table 5).

Table 4. Codon biases of gene ensembles in 200 kb windows traversing the *E. coli* genome from oriC to the ter region and demarcating coding sequences of the leading and lagging strand.

Leading strand genes[a]				Lagging strand genes[a]			
Region[b]	All	RP	tRN	Region[b]	All	RP	tRN
1R	63	513	276	1R	51	522	286
1L	45	526	285	1L	52	528	299
2R	91	464	215	2R	116	528	277
2L	87	525	278	2L	61	537	296
3R	52	532	303	3R	74	495	256
3L	106	430	221	3L	57	525	300
4R	42	517	283	4R	79	527	283
4L	78	485	247	4L	45	542	307
5R	75	481	243	5R	63	521	310
5L	42	520	283	5L	71	513	301
6R	66	530	282	6R	54	554	320
6L	53	518	293	6L	61	574	344
7R	67	571	334	7R	38	546	310
7L	79	499	247	7L	54	552	332
8R	59	495	256	8R	71	530	282
8L	36	530	293	8L	79	564	348
9R	46	544	312	9R	55	516	284
9L	57	564	316	9L	64	581	345
10R	56	554	328	10R	85	559	348
10L	80	579	352	10L	79	499	247
11R	88	583	365	11R	80	531	326
11L	67	565	338	11L	80	550	335
12R	86	574	350	12R	107	592	385
12L	103	578	360	12L	92	569	358

a. Genes from the leading strand (i.e. transcribed away from *oriC*) separated from those transcribed from the lagging strand.
b. The regions are labeled by the position relative to the origin of replication (i.e. 1L refers to the first 200 kb left of the origin). Regions 12R and 12L are next to each other separated by the ter region near position 1588 kb. These two windows are 104 kb and 135 kb long.

An alternative partition of the *E. coli* genome is generated from 10 non-overlapping clockwise-oriented gene contigs of about equal length (about 460 kb each) commencing at oriC, labelled 0, 1, . . . , 9. The contig 5 is diametrically opposite to the origin, and contig 9 is adjacent counter-clockwise to contig 0. Codon biases are described in Table 5.

Codon biases with respect to the average gene are small with the largest bias, 0.081, in the ter region. This is the region with the smallest site 3 G+C frequency, about 5% lower than for the genes about oriC. The deviation in codon usage from RP and tRN genes is most emphatic at the same ter region.

Gene size and site 3 G+C frequency

Genes of *E. coli* were divided into size classes with approximately equal numbers of codons (Table 6).

The S3 variation in Table 6 is striking. In *E. coli* there is a 5% increase in the S3 frequency traversing genes of length 100–300 codons to genes of length >800 codons, whereas in eukaryotic site 3 G+C frequency tends to decrease with increasing gene length.

Relative codon usage bias among 5′, middle, and 3′ parts of genes in E. coli

Six classes of genes were defined by gene length. Within each class, all genes were divided into three equal parts (5′, middle, 3′). Codon biases were compared among the three parts (Table 7).

The calculations for genes ≥100 codons in *E. coli* show that the middle and the last third of genes are more similar in codon usage than either is to the initial third of the gene. This data supplements earlier findings of Eyre-Walker and Bulmer (1993), Chen and Inouye (1994) and Goldman *et al.* (1995).

Table 5. Relative codon bias (multiplied by 1000) among genes of 10 contigs of equal length around the *E. coli* genome.[a]

$\{F\}\backslash\{G\}$	0	1	2	3	4	5	6	7	8	9	All	RP	tRN	S3%
0	*	29	26	22	67	112	70	36	34	28	32	519	277	57.1
1	29	*	28	32	73	113	71	43	38	28	36	513	274	57.2
2	27	28	*	31	80	121	78	50	47	32	43	518	278	57.9
3	22	32	31	*	58	97	55	31	31	34	24	530	288	57.0
4	67	73	81	58	*	51	29	34	39	69	38	540	314	54.2
5	113	113	121	97	51	*	47	78	80	115	81	576	358	52.5
6	70	72	79	55	29	47	*	43	45	78	41	567	334	54.7
7	36	43	50	31	34	78	43	*	14	42	13	528	296	55.7
8	34	38	47	31	39	80	44	14	*	42	12	529	295	55.7
9	28	28	32	34	68	114	78	41	42	*	38	499	263	57.0

a. The contigs are labelled 0–9. Contig 0 is centred at oriC. The numbering continues clockwise with contig 5 being opposite to oriC (containing the ter region) and contig 9 being next to 0. Codon biases of genes in each contig relative to all other contigs are shown in the left part. Also shown are biases relative to all *E. coli* genes (all), ribosomal proteins (RP), and amino acyl tRNA synthetases (tRN). The rightmost column shows average codon site 3 G+C frequencies (S3%).

The variation S3% over each third part increases steadily about 5% concomitant with gene size. The middle third codons pervasively show about one-half percentage higher S3% than the carboxyl third and about 2% higher than the amino third. This may suggest more accurate translation for the central parts of genes.

How does codon usage vary with function?

The standard protein function classification distinguishes 14 functional categories (Riley, 1993, see also legend to Table 8). The homogeneity of several of these categories is problematic. Proteins involved in subclasses of translation are described in the *Introduction*. The category Cellular Processes includes proteins of cell division, cell killing, chaperones, detoxification, protein and peptide secretion and transformation; the Transcription category focuses on DNA-dependent RNA polymerase and proteins contributing to degradation of RNA.

Mutual codon biases for the Translation, Energy metabolism and Transcription (Tl, EM, Tc) categories are all relatively small <0.100, Table 8. The Replication, Regulatory (R, RF) and the set of categories (CM, T, X, CE) register mutual codon bias <0.060. But these two sets of categories entail relative codon biases greater by a factor of two (>0.130) (Tl, EM, Tc). Notably Tl, EM and Tc gene classes are the most distant from the average gene and somewhat closer to RP genes. This result applies for several enteric bacterial gene collections including *E. coli* and *H. influenzae*.

Codon bias and 'alien' genes.

Different bacterial genomes display wide variation in their overall G+C content that is commonly attributed to varying mutational mechanisms and processes. Genes within a species tend to be rather homogeneous in base composition (similar in residue and codon usages), although the 'highly expressed genes' are often significantly different from the average gene. There are other genes, *g*, in the horns (see Fig. 2) that may possibly be considered to be DNA imported through recent horizontal transfer or deviant

Table 6. G+C% at site 3 (S3) and number of genes (below).

Gene size (codons)	Bacterial genomes			Eukaryotes		
	E. coli	*H. influenzae*		Human	Yeast	*C. elegans*
100–300	53.7 (1970)	28.4 (812)	100–300	59.9 (488)	40.7 (2179)	39.1 (565)
200–400	55.2 (1795)	28.6 (685)	200–400	60.4 (494)	40.2 (1789)	36.4 (513)
300–500	56.2 (1363)	29.0 (473)	300–500	60.0 (497)	39.4 (1615)	36.4 (439)
400–700	57.3 (916)	29.4 (327)	400–700	60.0 (526)	38.6 (1779)	36.6 (390)
500–1000	57.7 (539)	29.4 (201)	500–900	58.5 (423)	37.8 (1556)	36.0 (314)
>800	58.3 (139)	29.1 (51)	700–1200	56.0 (287)	36.4 (864)	35.4 (207)
			>1000	53.1 (199)	35.4 (477)	34.5 (118)

Table 7. Relative codon biases among 5′, middle and 3′ parts of *E. coli* genes.[a]

$\{F\}\backslash\{G\}$	Size 100–300 codons			Size 200–400 codons				Size 300–500 codons				Size 400–700 codons				Size 500–1000 codons				Size ≥800 codons			
	5′	m	3′	$\{F\}\backslash\{G\}$	5′	m	3′	$\{F\}\backslash\{G\}$	5′	m	3′	$\{F\}\backslash\{G\}$	5′	m	3′	$\{F\}\backslash\{G\}$	5′	m	3′	$\{F\}\backslash\{G\}$	5′	m	3′
5′	*	66	56	5′	*	66	58	5′	*	73	67	5′	*	71	71	5′	*	70	75	5′	*	71	75
m	65	*	34	m	65	*	35	m	72	*	36	m	70	*	38	m	69	*	41	m	71	*	52
3′	54	34	*	3′	57	35	*	3′	66	36	*	3′	70	38	*	3′	74	41	*	3′	74	52	*
S3%[b]	52.4	54.7	54.0	S3%	53.9	56.1	55.4	S3%	55.0	57.2	56.5	S3%	56.2	58.0	57.5	S3%	56.6	58.5	57.8	S3%	57.4	59.4	58.0

a. Genes of similar sizes were divided into three equal parts (5′, middle, 3′) and codon biases among the codon aggregates of the three parts were calculated using formula (1).
b. Average G+C per cent at codon site 3 (S3%) for the three parts.

due to other disrupting influences. In terms of codon bias assessments, we characterize genes as alien if they fulfil the following criteria: (i) codon bias [formula (1)] of *g* compared with the average gene of the species exceeds an appropriately high threshold; (ii) codon bias of *g* relative to ribosomal proteins $B(g|RP)$ is also appropriately high.

Requirement (ii) excludes most ‘highly expressed genes’ as alien genes. At the time of introgression, horizontally transferred genes reflect the genome composition of the donor genome that over time acquire the DNA compositional ‘biases and asymmetries’ of the new genome (Lawrence and Ochman, 1997). In particular, we expect that the transferred gene yields to mutational and selection tendencies adapting it to the ‘genome signature’ and ‘codon signature’ of the new host (Karlin and Mrázek, 1996).

Figure 2 plots for each gene (of length at least 200 codons) its codon bias (CB) and S3 frequency. The thresholds of (i) and (ii) are generally chosen to define natural clusters in the Fig. 2 plot. In the *E. coli* genome an appropriate threshold in (i) is $B = (g|\text{all}) \geq 0.52$ and in (ii) is $B = (g|RP) \geq 0.45$. These distinguish about 3% of genes of Fig. 2. Three horns are conspicuous corresponding to $G_1 = (CB \geq 0.52,\ \text{S3} < 45\%)$, $G_2 = (CB \geq 0.52,\ 45\% < \text{S3} < 68\%)$ and $G_3 = (CB \geq 0.52,\ \text{S3} > 68\%)$. The plot is reminiscent of that of Médigue *et al.* (1991), which projected two horns. About one-third of group G_1 qualify as laterally transferred genes according to the compilation of Lawrence and Ochman (1997) (see Table 9a). These authors only considered genes from a 1.43 Mb contig *M* of *E. coli* centred at ori-C but including genes of all sizes. Group G_2 predominantly consists of ‘highly expressed genes’ generally violating condition (ii). Group G_3 genes apparently are also ‘laterally transferred’ genes.

Table 9a–c lists all the genes of codon bias ≥0.520 relative to the average gene. The left horn (Table 9a, see also Fig. 2) contains 44 genes most of ORF status or homologous to a hypothetical protein. This collection features three clusters of contiguous genes with high codon biases relative to the average *E. coli* gene and to RPs. Their indicated starting positions and orientation are as follows: three genes $C_1 = \{1196755(-), 1197460(-), 1197918(+)\}$ (total length about 5 kb); four genes $C_2 = \{2102531(-), 2104079(-), 2105248(-), 2107606(-)\}$, extending about 7–8 kb; five genes, $C_3 = \{3794575(+), 3796939(-), 3797823(-), 3799626(-), 3802743(-)\}$, extending more than 10 kb.

The distribution of the genes of the left horn around the *E. coli* genome is highly non-random including three clusters. Except for cluster C_2, the other cluster genes are of unknown function. The genes of the right horn are listed in Table 9b. This features an alien cluster of length about 8 kb. This cluster again represents genes of unknown function.

The genes of high codon bias but with G+C site 3

Table 8. Relative codon usage bias (multiplied by 1000) among *E. coli* gene major functional classes.[a]

{*F*}\{*G*}	Tl	EM	Tc	CP	NN	FA	AA	CM	T	X	CE	R	RF	BC	RP	All
Tl	*	60	89	75	88	134	168	159	173	173	182	202	206	245	309	238
EM	62	*	93	83	77	119	154	152	165	168	181	199	204	231	329	224
Tc	91	93	*	135	147	180	229	213	230	225	233	260	265	298	250	286
CP	76	80	137	*	87	107	138	109	120	131	134	150	153	203	354	189
NN	90	79	149	88	*	74	90	94	111	108	125	143	143	171	392	168
FA	139	120	181	108	76	*	86	91	96	88	94	126	121	144	414	134
AA	173	156	231	139	91	86	*	69	65	66	92	90	82	98	482	103
CM	164	153	215	109	95	89	69	*	49	59	70	71	72	107	452	97
T	183	171	237	123	115	97	64	53	*	52	58	73	60	105	489	92
X	179	168	229	134	107	88	65	58	51	*	54	71	60	90	477	84
CE	189	181	238	137	127	95	92	73	60	55	*	87	71	86	482	79
R	207	200	261	151	144	121	89	73	68	71	85	*	56	82	497	88
RF	207	202	264	151	142	118	82	73	58	60	69	57	*	89	493	84
BC	252	235	298	205	173	144	98	109	104	90	86	84	90	*	538	57
RP	326	337	259	361	395	411	482	455	472	470	467	504	502	539	*	520

a. The functional classes (adapted from Riley, 1993) include translation (excluding ribosomal proteins; labelled Tl); energy metabolism (labelled EM); transcription (Tc); cellular processes (CP); purines, pyrimidines, nucleosides and nucleotides (NN); fatty acid and phospholipid metabolism (FA); amino acid biosynthesis (AA); central intermediary metabolism (CM); transport and binding proteins (T); cell envelope (CE); replication (R); regulatory functions (RF); biosynthesis of cofactors, prosthetic groups, and carriers (BC); ribosomal proteins (RP); and other categories (X). In addition, all functional classes were compared to a standard of all *E. coli* genes (labelled All).

frequencies in the normal range (45% < S3 < 68%), the middle horn, are reported in Table 9c. These genes have mostly been recognized by BLAST searches and correspond in *E. coli* to highly expressed genes particularly active during exponential growth. These include eight ribosomal proteins (of ≥ 200 residue length), three membrane OMP (porin) genes, the heat shock proteins DnaK, GroEL, several elongation factors and other essential transcription or translation processing genes. Except for a contiguous array of four ribosomal proteins, the genes of Table 9c (see Fig. 3), unlike the alien genes, do not cluster.

Genes in the left horn located within the 1.43 Mb contig *M* studied by Lawrence and Ochman (1997) are considered horizontally transferred by their analysis. Much less conservative, they list 228 such genes (139 in size greater than 200 codons, many less than 100 codons) compared with only 16 genes of Fig. 3 in the region *M*. Thus, most of their genes do not satisfy $B = (g|\text{all}) \geq 0.52$ and, in fact, 88 of their genes have $B = (g|\text{all}) \leq 0.40$ and 21 of these genes have $B = (g|\text{all}) \leq 0.31$, which is the individual average bias among all genes of *E. coli*. We do not imply that the 16 genes we list in region *M* are the only ones that have been horizontally transferred or that Lawrence and Ochman's figure of 228 is a less accurate estimate of the actual amount of horizontal transfer – only that the 16 we identified are more significantly alien than many of their 228. Parenthetically, codon bias is negatively correlated with gene size.

Codon biases of coli phage

Codon usage of all the genes of the bacteriophages lambda and T7 are compared with average codon usage in *E. coli*. The relative biases are displayed in Table 10.

Clearly, the genes of lambda are quite similar in codon frequencies with *E. coli* (B(Lambda|Eco) = 0.194), whereas

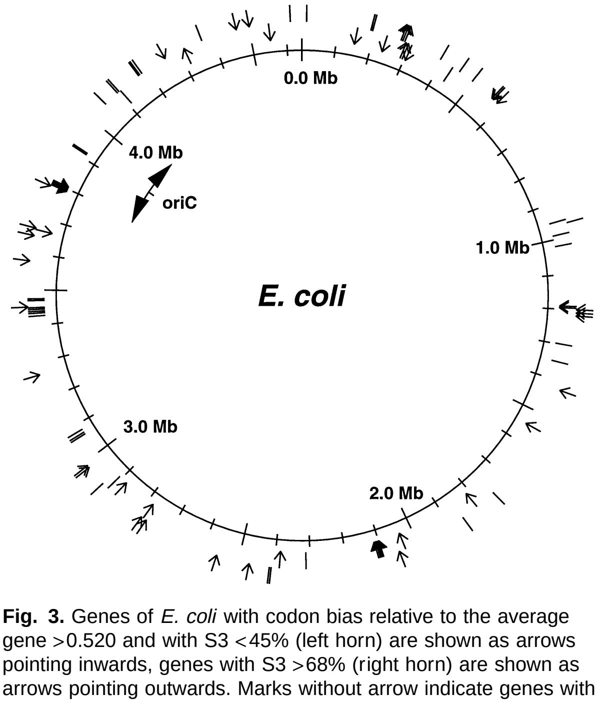

Fig. 3. Genes of *E. coli* with codon bias relative to the average gene >0.520 and with S3 <45% (left horn) are shown as arrows pointing inwards, genes with S3 >68% (right horn) are shown as arrows pointing outwards. Marks without arrow indicate genes with S3 between 45% and 68% (middle horn). The genes transcribed clockwise are indicated in the outer circle, the genes transcribed counterclockwise are indicated in the middle circle. The inner circle indicates position in the genome. Genes with arrows are considered alien genes, see text. The genes without arrows are predominantly highly expressed genes.

Table 9a. Genes of length ≥200 codons with codon usage bias ≥0.520 relative to all *E. coli* genes and codon site 3 G+C content <45%.

Position in the genome[a]	Gene	Bias[b] All	Bias[c] RP	S3%	Function/pathway/subcellular location
152231 −	yadL	525	733	34.5	h.p.
234782 −	f251	564	899	35.2	ORF
293142 −	yagL	527	823	35.9	ORF
568125 +	o508	593	875	28.6	ORF
570677 +	o265	679	937	28.4	ORF
582904 +	appY	611	950	28.2	M5 polypeptide, transcriptional regulator
1196755 −	f221	605	934	27.3	ORF
1197460 −	f234	566	894	27.3	ORF
1197918 +	lit	628	920	27.4	Bacteriophage T4 late gene expression blocking protein
1209569 +	mcrA	591	869	30.1	restriction enzyme A
1218824 +	o506	540	832	27.5	ORF
1421806 +	trkG	578	877	32.2	Potassium uptake protein, i.m.p.
1543738 −	f318	603	886	30.6	ORF
1811168 −	f271	530	909	32.2	ORF
2036893 −	f239	530	857	42.0	ORF
2054880 +	o238	534	662	27.4	ORF
2102531 −	yefI	552	839	25.9	l.b., possibly glycosyltransferase
2104079 −	yefG	573	877	28.6	l.b.
2105248 −	rfc	583	857	32.6	O-antigen polymerase, l.b., i.m.p.
2107606 −	rfbX	602	925	29.7	put. O-antigen transporter, l.b., i.m.p.
2386446 −	f238	554	856	35.9	ORF
2467151 +	o443	650	989	24.2	ORF
2558277 +	o213	521	664	44.8	ORF
2771339 +	o567	591	887	27.4	ORF
2782449 −	f263	536	893	30.5	ORF
2785627 +	o210	568	888	31.1	ORF
2879077 −	f226	530	899	40.9	ORF
2985498 +	o230	529	816	33.5	ORF
2990116 +	o458	536	798	30.2	ORF
3266056 +	yhaC[d]	547	886	27.7	h.p.
3453215 +	yheE[d]	542	867	38.1	put. general secretion, prob. inner membrane
3579494 +	yhhZ[d]	592	864	25.8	h.p.
3648921 +	yhiS[d]	582	879	32.8	h.p.
3663440 −	yhiX[d]	583	955	37.0	transcriptional regulator
3763964 +	yibA[d]	628	854	26.5	h.p.
3794575 +	rfaL[d]	655	967	25.1	l.c.b., i.m.p.
3796939 −	rfaK[d]	642	924	23.0	l.c.b., peripheral membrane
3797823 −	rfaZ[d]	553	851	25.5	l.c.b.
3799626 −	rfaJ[d]	609	901	24.6	Lipopolysaccharide glucosyltransferase, l.c.b.
3802743 −	rfaS[d]	630	925	22.6	l.c.b.
4258178 +	yjbM[d]	528	822	32.9	h.p.
4472975 +	yjgL[d]	555	877	29.9	h.p.
4501566 +	yjhB[d]	638	901	25.2	h.p., possibly i.m.p.
4553889 −	yjiC[d]	536	821	30.9	h.p.

a. Position of the 5′ end and DNA strand (+ or −).
b. Codon bias multiplied by 1000 relative to all *E. coli* genes.
c. Codon bias multiplied by 1000 relative to *E. coli* ribosomal proteins.
d. Putative horizontally acquired gene according to Lawrence and Ochman (1997).
h.p., hypothetical protein, put., putative, prob., probably, l.b., liposaccharide biosynthesis, l.c.b., liposaccharide core biosynthesis, i.m.p., integral membrane protein.

T7 genes deviate substantially. The codon bias of T7 relative to *E. coli* is not as extreme as that of the alien genes characterized in the preceding section.

Codon pair biases

It is widely acknowledged that codon context as well as other factors, some described in the *Introduction*, can influence codon bias. Paralleling formula (1) we introduce a measure of codon pair bias (dicodon bias) for a class of genes relative to a standard C in terms of dicodon frequencies, namely

$$\sum_{a,b} p_f(a,b)\left[\sum_{\substack{(xyz)=a\\(uvw)=b}} |f(xyz,uvw)-c(xyz,uvw)|\right] \tag{2}$$

where $(xyz)=a$, $(uvw)=b$ are successive codons for the diresidue $\{a,b\}$, $p_f(a,b)$, indicates diresidue frequencies of $\{a,b\}$ for F and the dicodon frequencies $f(xyz,\ uvw)$ and $c(xyz,\ uvw)$ are normalized for each diresidue to sum to 1. The gene sets F and C should be sufficiently large to avoid effects of statistical fluctuations.

Table 9b. Genes of length ≥200 codons with codon usage bias ≥0.520 relative to all *E. coli* genes and codon site 3 G+C content >68%.

Position in the genome[a]	Gene	Bias[b] All	Bias[c] RP	S3%	Function/pathway/subcellular location
282425 +	yagF	633	689	85.8	ORF
284619 +	yagG	672	706	89.5	ORF
286013 +	o536	664	650	88.0	ORF
288386 −	yagI	616	738	83.7	ORF
289529 −	argF	523	622	76.9	Ornithine carbamoyltransferase chain F
4315473 −	phnL[d]	535	648	70.2	Part of transport system for alkylphosphonates

a. Position of the 5′ end and DNA strand (+ or −)

b. Codon bias multiplied by 1000 relative to all *E. coli* genes.

c. Codon bias multiplied by 1000 relative to *E. coli* ribosomal proteins.

d. Putative horizontally acquired gene according to Lawrence and Ochman (1997).

Table 9c. Genes of length ≥200 codons with codon usage bias ≥0.520 relative to all *E. coli* genes and codon site 3 G+C content between 45% and 68%.

Position in the genome[a]	Gene	Bias[b] All	Bias[c] RP	S3%	AB[d]	Function/pathway/subcellular location
12163 +	dnaK	573	189	52.3	9.33	Heat shock protein 70
189874 +	rpsB	660	200	56.7	NA	30S r.p.S2
190857 +	tsf	627	262	50.4	4.7	e.f. Ts, cyt.
392239 +	o467	635	226	57.0	NA	ORF
431237 −	tsx	529	246	62.5	NA	Nucleoside-specific channel-forming
454357 +	tig	586	265	54.1	6.1	Trigger factor, involved in protein export
496399 +	adk	540	332	52.6	NA	Adenylate kinase, essential for cell growth
952777 −	pflB	672	219	58.6	3.9	Formate acetyltransferase, g.m., cyt.
961218 +	rpsA	689	269	52.3	15.10	30S r.p. S1, binds mRNA
986205 −	ompF	580	226	49.0	NA	o.m.p. F, porin, i.m.p., T2 phage rec.
1019276 −	ompA	708	218	55.9	20.	o.m.p. A, porin, i.m.p., T-even phage rec.
1297344 −	adhE	546	193	50.5	NA	Alcohol/acetaldehyde dehydrogenase
1349063 −	fabI	535	329	55.6	NA	Enoyl-[acyl-carrier-protein] reductase
1753722 +	pykF	569	318	52.9	NA	Pyruvate kinase, glycolysis,
1860795 +	gapA	788	315	52.4	NA	g3pd. A, glycolysis, cyt.
2310769 −	ompC	783	261	57.1	NA	o.m.p. C, porin, i.m.p.
2411490 +	ackA	534	240	56.6	1.0	acetate kinase, cyt.
2412767 +	pta	549	218	59.9	1.5	Phosphate acetyltransferase
2905963 −	eno	751	285	51.3	10.	Enolase, glycolysis, cyt.
2926251 +	sdaC	537	243	66.6	NA	put. serine transporter, i.m.p.
3069264 −	fba	683	224	58.1	NA	f.b.a., glycolysis
3070642 −	pgk	644	223	53.6	NA	Phosphoglycerate kinase, glycolysis, cyt
3079654 −	tktA	547	283	60.7	NA	Transketolase
3439312 −	rpsD	524	312	50.7	NA	30S r.p. S4, binds to 16S RNA, also t.r.
3447520 −	rpsC	646	236	48.3	NA	30S r.p S3, binding of initiator Met-tRNA
3449001 −	rplB	580	170	46.3	NA	50S r.p. L2, binds to 23S RNA
3449923 −	rplD	638	215	54.0	NA	50S r.p. L4, binds to 23S RNA, also t.r.
3450563 −	rplC	626	291	48.6	NA	50S r.p. L3, binds to 23S RNA
3468966 −	tufA	718	291	55.5	55.	e.f. EF-Tu, cyt.
3471151 −	fusA	657	258	51.6	16.	e.f. EF-G, cyt.
3915003 −	atpD	563	307	59.7	5.6	ATP synthase F1 beta sub.
3917485 −	atpA	604	235	54.3	6.9	ATP synthase F1 alpha sub.
4055987 +	yihK	543	193	59.3	NA	binds GTP, prob. interacts with ribosomes
4098391 +	sodA	528	348	61.9	NA	Manganese superoxide dismutase
4105132 +	pfkA	562	306	57.7	0.5	6-phosphofructokinase, glycolysis
4109087 −	tpiA	660	276	50.8	NA	Triosephosphate isomerase
4173523 +	tufB	687	272	56.5	NA	e.f. EF-Tu, duplicate gene
4176457 +	rplA	686	274	50.6	NA	50S r.p. L1, binds to 23S RNA, also t.r.
4182928 +	rpoC	574	199	57.3	NA	DNA-dep. RNA polymerase, beta′-sub.
4368603 +	mopA	691	288	51.7	13.50	GroEL, 60 kDa chaperonin
4618452 +	deoD	523	248	58.8	NA	Purine-nucleoside phosphorylase

a. Position of the 5′ end and DNA strand (+ or −).

b. Codon bias multiplied by 1000 relative to all *E. coli* genes.

c. Codon bias multiplied by 1000 relative to *E. coli* ribosomal proteins.

d. Protein product abundance in exponential growth phase (from VanBogelen *et al*., 1996.

g3pd., glyceraldehyde 3-phosphate dehydrogenase; cyt., cytoplasmic; o.m.p., outer membrane protein; i.m.p., integral membrane protein; e.f., elongation factor; r.p., ribosomal protein; t.r., translational repressor; f.b.a., fructose 1,6-bisphosphate aldolase; g.m., glucose metabolism; rec., receptor; sub., subunit; prob., probably.

Table 10. Relative biases.

	Lambda	T7	Eco
Lambda	*	420	194
T7	417	*	457
Eco	189	448	*

Size classes. Let C = all genes and F be genes determined by the size classes {100–300, 200–400, 300–500, 400–700, 500–1000, >800} as in Table 6. Table 11 gives the results of formula (2) in this example. Clearly, dicodon bias increases with gene size.

Gene contigs around the genome (Table 12, see also Table 5).

The genome is divided equally into 10 contigs generating 10 gene classes as in Table 5. Dicodon bias is most pronounced between the gene classes of the region about oriC versus the ter region. However, the contrasts in bias are not as strong as for straight codon bias although the level of dicodon bias is twice as high. Compared with the average gene dicodon biases are approximately constant (about 0.095) throughout the genome.

Dicodon biases along genes. Dividing genes into thirds as in Table 7, the dicodon biases of the 5′ third, middle third, 3′ third parts of genes are much more similar, although the level of bias is about twice that of straight codon bias (data not shown). These three gene classes seem to indicate that codon pair utilization is more highly biased and species specific than straight codon bias.

Discussion

Establishing the rules of codon usage is of interest with respect to fundamental evolutionary questions, in gene prediction, in classifying gene families, and in the design of optimal expression vectors. Codon pair usage program (e.g. GRAIL, GENIE, GENSCAN) (hexanucleotide) evaluations have become essential information for gene finding in prokaryotic and eukaryotic genomes (Claverie, 1997; Burge and Karlin, 1997). The analysis of dicodon biases would be germane in this context.

A new broadly applicable measure for assessing codon bias of one group of genes with respect to a second group of genes is introduced [formulae (1) and (2) of text]. In this formulation, codon and dicodon bias correlations for *E. coli* genes were evaluated for level of expression, for contrasts along genes, for genes in different 200 kb or longer length contigs around the genome, for gene size classes, and for variation over different function classes. It is often stated that *E. coli* uses synonymous codons that bind the commonest tRNAs, minimize proofreading costs and maximize rate and accuracy of translation (e.g. Dong *et al*., 1996). Explanations for codon bias have generally involved combinations of selection and mutational pressures; optimizations of translation rate and accuracy in relation to tRNA abundances are considered among primary selection forces (Bulmer, 1988; Andersson and Kurland, 1990). Bulmer (1991) proposes that for the high expression level of a gene, high translation initiation rate is important, whereas high elongation rate is less important. An additional selective pressure (putatively active on mRNA) consists of preventing formation of secondary structures which competes with constraints optimizing elongation speed.

Contrasts between ribosomal protein and tRNA synthetase genes

RPs and aminoacyl tRNA synthetases (tRN) are generally ‘highly expressed’ gene classes during exponential growth of the *E. coli* cell. However, relative to the average *E. coli* gene, codon biases for RP genes are much more extreme than for tRN genes. Corresponding results apply to the experimentally determined proteins of abundance levels delineated by VanBogelen *et al.* (1996). Thus, in general, the greater molar abundance a gene’s product, the more its codon usage resembles that of RPs, whereas this is not the case for comparisons to tRN genes or for the class of modification (M) protein genes essential to translation activities. Similar contrasts hold for codon biases in

Table 11. Relative dicodon bias (multiplied by 1000) among genes of six size classes.[a]

$\{F\}\backslash\{G\}$	100–300	200–400	300–500	400–700	500–1000	>800	All
100–300	*	54	83	103	115	168	59
200–400	54	*	51	93	102	155	49
300–500	82	51	*	60	86	133	35
400–700	104	96	61	*	50	113	54
500–1000	116	104	87	49	*	104	68
>800	164	155	131	111	102	*	119

a. Dicodon biases among six partly overlapping aggregates of genes by the gene size (100–300, 200–400, 300–500, 400–700, 500–1000, and >800 codons; indicated at the top and left) were calculated using formula (2). Also shown are dicodon biases relative to all *E. coli* genes (all).

Table 12. Relative dicodon bias (multiplied by 1000) among genes of 10 contigs of equal length around the *E. coli* genome.[a]

{*F*}\{*G*}	0	1	2	3	4	5	6	7	8	9	All
0	*	132	146	121	134	155	143	138	145	129	92
1	135	*	125	122	141	138	144	107	159	142	93
2	145	124	*	129	130	144	142	143	150	131	96
3	121	126	135	*	124	145	134	121	127	139	88
4	139	139	134	125	*	116	143	101	135	136	81
5	159	141	149	144	115	*	103	121	180	161	98
6	147	146	143	132	140	104	*	120	163	160	95
7	143	106	145	120	101	123	121	*	125	139	71
8	148	159	145	128	131	176	161	124	*	154	113
9	134	143	133	141	134	162	159	141	153	*	93

a. Dicodon biases calculated using equation (2) are indicated for the same gene aggregates as in Table 5. The biases relative to all *E. coli* genes are shown in the rightmost column.

Bacillus subtilis but not for corresponding codon biases in the *Methanococcus jannaschii*, *Mycoplasma genitalium* and *Synechosystis* genomes.

Although RP and tRN genes tend to be highly expressed during exponential growth, they show a high relative mutual codon usage difference B*(RP, tRN) ≈ 0.460. Apparently, for these gene classes the source of the codon bias differ significantly. For unknown reasons, RP genes are among the most deviant from the average *E. coli* gene. Is it related to the fact that the *E. coli* organism is especially fast growing?

It has been suggested that codon usage and tRNA abundances are correlated for highly expressed genes simply to match substrate levels with cellular demands, and that frequently used codons are not required for rapid translation rates (Irwin *et al.*, 1995). From these considerations, perhaps it is not surprising that RPs, which are small single-domain proteins, show high codon usage correlation with overall *E. coli* abundant codon usage, whereas the larger multidomain (and some dimeric) tRNA synthetases do not show this same bias, although they are also highly expressed. Perhaps there is greater need for these multidomain enzymes to fold properly and therefore a greater codon pair bias for directing the kinetics of translation (e.g. pause sites) to facilitate correct folding (see below).

Codon differences along gene length

Independent of gene size, we observed (Table 7) that the middle and final third (3′ end) of genes mutually entail the same levels of codon biases (quite similar codon frequencies), whereas the 5′ third codons possess significantly different codon frequencies. These differences may be important in translation initiation and/or early stages of translation. In this respect, a prominent example concerns encoding arginine with a major codon {CGN} versus a minor codon {AGR}. Specifically, the codons {AGR} are generally rare in *E. coli* genes and are primarily restricted to the amino end of genes (especially within the initial 25 bp), whereas {CGN} is predominant elsewhere in genes (see Chen and Inouye, 1994). However, these results do not apply to most other bacterial genomes (Karlin and Mrázek, 1998). For gene sizes up to 800 amino acids in eukaryotic genomes (human, yeast *C. elegans*), the middle and 3′ codon groups have quite similar codon frequencies, whereas the 5′ group projects a twofold increase in codon bias compared with the codon bias between the middle and 3′ groups. The comparisons are somewhat attenuated for large size genes. The strongest biases are with shorter genes (length 100–400 amino acids). However, the data of Irwin *et al.* (1995) suggest that highly biased slowly translated codon pairs (ribosome stalling sites) are more closely correlated with levels of expression than with protein length.

Codon usage along the E. coli *chromosome*

In dividing the *E. coli* genome into 10 equal lengths, we found a striking difference in site 3 G+C frequencies (S3) where genes near oriC carry 5% greater S3 frequencies than the genes about the ter region (Table 5). This distinction also applies to small (100–300 amino acids length) versus large (>800 amino acids length) genes where explicitly site 3 G+C frequencies are on average 58% for large genes compared with 53% for small genes (Table 6). The latter inequality does not hold for most other prokaryotic genomes. However, contrary to *E. coli*, for eukaryotic genomes including vertebrates, yeast and *C. elegans*, site 3 G+C frequency decrease significantly with increasing gene length (Table 6) (Duret *et al.*, 1995, Eyre-Walker, 1996).

Large genes in humans tend to have many introns but not so in the yeast genome. Do these codon differences (codon site 3 G+C differences) reflect on reading accuracy and in transcription and translation rates? Introns are relatively AT rich and genes in AT-rich regions putatively disengage the template strand more easily fomenting more rapid transcription. On the other hand, codons of more S3 types (high site 3 G+C frequencies) conceivably are more stable during transcription and translation. On this basis smaller genes appear to congregate in higher G+C isochores ensuring greater translational fidelity especially appropriate for human genes (Burge and Karlin, 1997).

Long genes putatively entail more pause sites and/or ribosomal stalling, perhaps merely by virtue of length. Concomitantly, reading mistakes could truncate transcripts. From this perspective C and G nucleotides in mRNA entail more stable codon–anticodon bonding interaction and more accurate translation, whereas A and T nucleotides are probably transcribed and translated more rapidly often facilitated by the wobble basepair. For example, different translational accuracies of Asn codons are recognized. Experimental evidence (reviewed in Parker, 1992) shows that *E. coli* Asn has an unusually high translation error rate for the codon AAU compared with AAC (see also Akashi, 1994).

Codon bias appears in *E. coli* to be correlated with gene length, unlike most other prokaryotic genomes (data not shown). Also, codon bias is often different at the beginning and end of genes. It has been suggested that translation pause sites, especially early in coding sequences, can inhibit translation initiation (Irwin *et al*., 1995). In this purview there appear to be conflicting selection pressures imposed by the constraints of initiation of ribosomal binding and translation fidelity in the early stages of translation.

How does codon bias of the average gene correlate with function? Function class codon usages most similar to the average gene include proteins of transport activity, nucleotide and amino acid biosynthesis, and central intermediary metabolism, whereas the most biased gene classes feature processes of translation, transcription, and energy metabolism (Table 8). These two divisions may distinguish needs during periods of fast cellular growth versus the stationary cell state. In this context, codon biases of eukaryotic genes differ most from prokaryotic genes.

Alien (highly special or possibly laterally transferred) genes are characterized in terms of extreme codon bias relative to the average *E. coli* genes and sufficiently high bias relative to RPs. These genes generally are of unknown function and either AT rich or GC rich. There are several clusters in this compilation. Extremes in codon bias might also be used for identifying pathogenicity islands and in developing gene classes reflecting differential expression levels and expression in untypical environments (e.g. anaerobic, stress conditions, nutritional changes).

Codon choices and co-translational folding

There are increasing studies concerned with correlations between codon usage in a gene sequence and protein structure (e.g. Thanaraj and Argos, 1996; Netzer and Hartl, 1997; 1998). The analysis of Thanaraj and Argos (1996) argues *the rare codon hypothesis* for domains and secondary structures where the use of repetitive rare codons might reduce translation rate and induce translation pauses, allowing protein domains and suitable secondary structures to fold into native structural conformation. Along these lines, there are apparently subtle differences in prokaryotic and eukaryotic translation mechanisms, e.g. the important role of chaperonins in prokaryotes but not in eukaryotes and the potentially important activity of co-translational folding in eukaryotes but not in prokaryotes, Netzer and Hartl (1998). In prokaryotes, codon choice is presumably influenced by structure via evolutionary selection for the most accurately translated sequences at structurally important locations. In eukaryotes, related to co-translational folding, propitious codon choice can facilitate domain folding and reduce the probability of misfolding.

Acknowledgements

Samuel Karlin is supported in part by NIH Grants 5R01GM10452-34, 5R01HG00335-10, and NSF Grant DMS9704552. Allan M. Campbell is supported in part by NIH Grant 5R01GM51117-29.

References

Akashi, H. (1994) Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translation accuracy. *Genetics* **136:** 927–935.

Andersson, S.G.E., and Kurland, C.G. (1990) Codon preferences in free-living microorganisms. *Microbiol Rev* **54:** 198–210.

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

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

Burge, C., and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78–94.

Chen, G.-F.T., and Inouye, M. (1994) Role of the AGA–AGG codons, the rarest codons in global gene expression in *Escherichia coli*. *Genes Dev* **8:** 2641–2652.

Claverie, J.-M. (1997) Computational methods for the identification of genes in vertebrate genome sequences. *Hum Mol Genet*. **6:** 1735–1744.

Deschavanne, P., and Filipski, J. (1995) Correlation of GC content with replication timing and repair mechanisms in weakly expressed *E. coli* genes. *Nucleic Acids Res* **23:** 1350–1353.

Dong, H., Nilsson, L., and Kurland, C.G. (1996) Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J Mol Biol* **260:** 649–663.

Duret, L., Mouchiroud, D., and Gautier, C. (1995) Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J Mol Evol* **40:** 308–317.

Eyre-Walker, A. (1996) Synonymous codon bias is related to gene length in *Escherichia coli*: Selection for translational accuracy? *Mol Biol Evol* **13:** 864–872.

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

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

Goldman, E., Rosenberg, A.H., Zubay, G., and Studier, F.W. (1995) Consecutive low-usage leucine codons block translation only when near the 5′ end of a message in *Escherichia coli*. *J Mol Biol* **245:** 467–473.

Irwin, B., Heck, J.D., and Hatfield, G.W. (1995) Codon pair utilization biases influence translational elongation step times. *J Biol Chem* **270:** 22801–22806.

Karlin, S., and Mrázek, J. (1996) What drives codon choices in human genes? *J Mol Biol* **262:** 459–472.

Kurland, C.G. (1993) Major codon preference: Theme and variations. *Biochem Soc Trans* **21:** 841–846.

Lawrence, J.G., and Ochman, H. (1997) Amelioration of bacterial genomes: Rates of change and exchange. *J Mol Evol* **44:** 383–397.

Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13:** 660–665.

Maynard-Smith, J., and Smith, N.H. (1996) Synonymous nucleotide divergence: What is ‘saturation’? *Genetics* **142:** 1033–1036.

Medigue, C., Rouxel, T., Vigier, P., Henaut, A., and Danchin, A. (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* **222:** 851–856.

Mrázek, J., and Karlin, S. (1998) Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci USA* **95:** 3720–3725.

Netzer, W.J., and Hartl, F.U. (1997) Recombination of protein domains facilitated by cotranslational folding in eukaryotes. *Nature* **388:** 343–349.

Netzer, W.J., and Hartl, F.U. (1998) Protein folding in the cytosol: Chaperonin-dependent and independent mechanisms. *Trends Biochem Sc.* **23:** 68–73.

Parker, J. (1992) Variations in reading the genetic code. In *Transfer RNA in Protein Synthesis* Hatfield, D.S., Lee, B.J., Pistle, R.M. (eds). CRC Press, pp. 191–267.

Riley, M. (1993) Functions of the gene products of *Escherichia coli*. *Microbiol Rev* **57:** 862–952.

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

Sharp, P.M., and Matassi, G. (1994) Codon usage and genome evolution. *Curr Opin Genet Dev* **4:** 851–860.

Thanaraj, T.A., and Argos, P. (1996) Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci* **5**: 1973–83.

VanBogelen, R.A., Abshire, K.Z., Pertsemlidis, A., Clark, R.L., and Neidhardt, F.C. (1996) Gene-protein database of *Escherichia coli* K-12, edn 6, In *Escherichia coli and Salmonella: Cellular and Molecular Biology*, 2nd edn, Neidhardt, F.C. (ed). American Society for Microbiology Press, Washington, D.C. pp. 2067–2117.