J Mol Evol (2000) 50:184–193
DOI: 10.1007/s002399910020

JOURNAL OF MOLECULAR EVOLUTION

Codon Usage in Plastid Genes Is Correlated with Context, Position Within the Gene, and Amino Acid Content

Brian R. Morton, Bernadette G. So*

Department of Biological Sciences, Barnard College, Columbia University, 3009 Broadway, New York, NY 10027, USA

Received: 21 July 1999 / Accepted: 5 November 1999

Abstract. Highly expressed plastid genes display codon adaptation, which is defined as a bias toward a set of codons which are complementary to abundant tRNAs. This type of adaptation is similar to what is observed in highly expressed *Escherichia coli* genes and is probably the result of selection to increase translation efficiency. In the current work, the codon adaptation of plastid genes is studied with regard to three specific features that have been observed in *E. coli* and which may influence translation efficiency. These features are (1) a relatively low codon adaptation at the 5′ end of highly expressed genes, (2) an influence of neighboring codons on codon usage at a particular site (codon context), and (3) a correlation between the level of codon adaptation of a gene and its amino acid content. All three features are found in plastid genes. First, highly expressed plastid genes have a noticeable decrease in codon adaptation over the first 10–20 codons. Second, for the twofold degenerate NNY codon groups, highly expressed genes have an overall bias toward the NNC codon, but this is not observed when the 3′ neighboring base is a G. At these sites highly expressed genes are biased toward NNT instead of NNC. Third, plastid genes that have higher codon adaptations also tend to have an increased usage of amino acids with a high G + C content at the first two codon positions and GNN codons in particular. The correlation between codon adaptation and amino acid content exists separately for both cytosolic and membrane proteins and is not related to any obvious functional property. It is suggested that at certain sites selection discriminates between nonsynonymous codons based on translational, not functional, differences, with the result that the amino acid sequence of highly expressed proteins is partially influenced by selection for increased translation efficiency.

Key words: Selection — Gene expression — Translation efficiency

Introduction

Essentially every protein-coding sequence analyzed to date displays a bias in synonymous codon usage. In many cases this codon bias reflects the genome composition bias and is probably due to a mutation bias, but there is strong evidence that selection discriminates between synonymous codons in some organisms. The most thoroughly studied example of this type of selection is *Escherichia coli.* Highly expressed genes from this bacterial species have a strong bias toward a specific subset of codons (Ikemura 1985; Sharp 1991) which are referred to as the major codons (Andersson and Kurland 1990). In each synonymous group there is usually a single major codon and this codon is complementary to the most abundant tRNA (Ikemura 1985). This observation that has led to the proposal that the bias toward major codons is an adaptation to increase translation efficiency of highly expressed genes (Ikemura 1985; Sharp 1991; Sharp and Matassi 1994). In contrast, *E. coli* genes

* *Present address:* Molecular Biology Institute, University of California, Los Angeles, CA 90024, USA
Correspondence to: Brian R. Morton; *E-mail:* bmorton@barnard.columbia.edu

with low levels of expression have a codon usage bias that appears to arise not from selection but from composition bias (Sharp 1991). Therefore, we distinguish among codon bias, any nonuniform usage of synonymous codons, and codon adaptation, a bias toward a defined set of major codons as a result of natural selection.

The codon usage of highly expressed *E. coli* genes has some interesting features in addition to an overall bias toward major codons. First, the bias toward major codons, that is, codon adaptation, is lower at the 5′ end of genes, particularly within the first 25 codons (Chen and Inouye 1990; Eyre-Walker and Bulmer 1993; Chen and Inouye 1994). Second, codon context is important. Codon usage at a particular site is influenced by flanking codons, presumably due to an effect of codon–codon interaction on translation (Murgola et al. 1984; Yarus and Folley 1985; Shpaer 1986; Gouy 1987). Third, the level of codon adaptation, measured by the Codon Adaptation Index (CAI) (Sharp and Li 1987), is correlated with amino acid usage (Lobry and Gautier 1994). Genes with high CAI values also have increased usage of amino acids coded by a GNN, and to a lesser extent an ANN, codon (Gutierrez et al. 1996). This correlation could be due to translational differences among nonsynonymous codons (Gutierrez et al. 1996) meaning that at some sites, selection may discriminate between amino acids based on differences in the translation efficiency of the two codons, not the functional properties of the amino acids themselves.

Plastid genes are similar to *E. coli* with regard to general features of codon usage bias. Highly expressed genes are biased toward a specific set of major codons (Morton 1993, 1996, 1998), while low-expression genes have a codon usage dominated by a high A + T content at third codon positions which matches the genome composition bias (Morton 1993). Since major codons match the plastid tRNA population, it has been proposed that selection for translation efficiency is generating a codon adaptation in highly expressed genes (Morton 1993, 1998).

Codon adaptation has been observed in every plastid genome studied to date but selection intensity varies among species. This variation is observed in both the number of genes that display codon adaptation and the level of this adaptation. Based on these two criteria we can roughly define three groups of species, those with strong, intermediate, and weak selection. The algae *Chlamydomonas reinhardtii* and *Cyanophora paradoxa* have strong selection with a large number of plastid genes in these species having high codon adaptation (Morton 1998). The green alga *Chlorella vulgaris*, the diatom *Odontella sinensis*, and the bryophyte *Marchantia polymorpha* have a selection intensity that is roughly intermediate; fewer genes show a noticeable bias toward major codons and the level of bias in these genes tends to be lower (Morton 1999). Finally, the flowering plants, such as *Oryza sativa*, have weak selection; only the most highly translated gene, *psbA*, shows a noticeable increase in the frequency of major codons, although the highly expressed *rbcL* gene does have a very slight increase (Morton 1998).

In the current work, plastid genes are examined for the additional codon usage features that have been observed in *E. coli*. It is shown that plastid genes have all three features described above. First, codon adaptation is lower at the 5′ end of coding sequences, particularly in those genes with a high overall codon adaptation. Second, codon usage bias is shown to vary among sites as a function of the composition of the 3′ flanking nucleotide. Finally, variation in the degree of codon adaptation among genes in the same genome is correlated with variation in amino acid content. The results demonstrate that codon usage bias of plastid genes is more complex than previously believed.

Materials and Methods

Complete plastid genome sequences were downloaded from GenBank and protein-coding sequences greater than 350 nucleotides in length were extracted as described previously (Morton 1998). Three species were selected in order to provide a comparison of genomes under strong, intermediate, and weak selection on codon usage. The three chosen were *Cyanophora paradoxa*, a species with strong selection on codon usage of plastid genes, *Marchantia polymorpha*, which has an intermediate selection intensity, and *Oryza sativa*, which has extremely weak selection on codon usage (Morton 1998).

In addition to the comparison of the three complete genomes, plastid genes with high levels of codon adaptation from three other species were studied. These included the genes with relatively high codon adaptation (based on a level defined below) from the algae *Chlorella vulgaris*, *Chlamydomonas reinhardtii* and *Odontella sinensis* which have intermediate to strong selection on codon usage overall (Morton 1999). Sequences from *Chlorella* and *Odontella* were extracted from the complete genome sequences while individual *Chlamydomonas* gene sequences were taken directly from GenBank (see the list of genes in Morton 1998).

For every gene the Codon Adaptation Index (CAI) (Sharp and Li 1987) was measured as described previously (Morton 1998). This statistic measures the adaptation of a particular gene (or any given codon usage table) based on a defined matrix of codon fitness values. In this case, codon fitness values were determined from the relative codon usage in a set of the most highly expressed plastid genes (see Morton 1998). Therefore, CAI will reflect the degree of bias toward major codons.

To correlate the variation in both codon adaptation and amino acid usage among genes, we developed two additional measures of codon adaptation. Both of these new measures are independent of amino acid composition, which is not necessarily the case for CAI; although CAI is largely independent of amino acid composition there are circumstances in which it might be influenced by protein composition. For the first measure, referred to as the Binary CAI (BCAI), every codon with a fitness of 0.85 or greater was reassigned a fitness of 1 and all others were reassigned a fitness of 0. The BCAI for a gene was then calculated from these new fitness values using the formula for CAI but with equal weighting for each of the 18 amino acids with multiple codons. The value for BCAI can range from 0 (no major codons) to 18 (only major codons used) and is not dependent on amino acid usage. The second

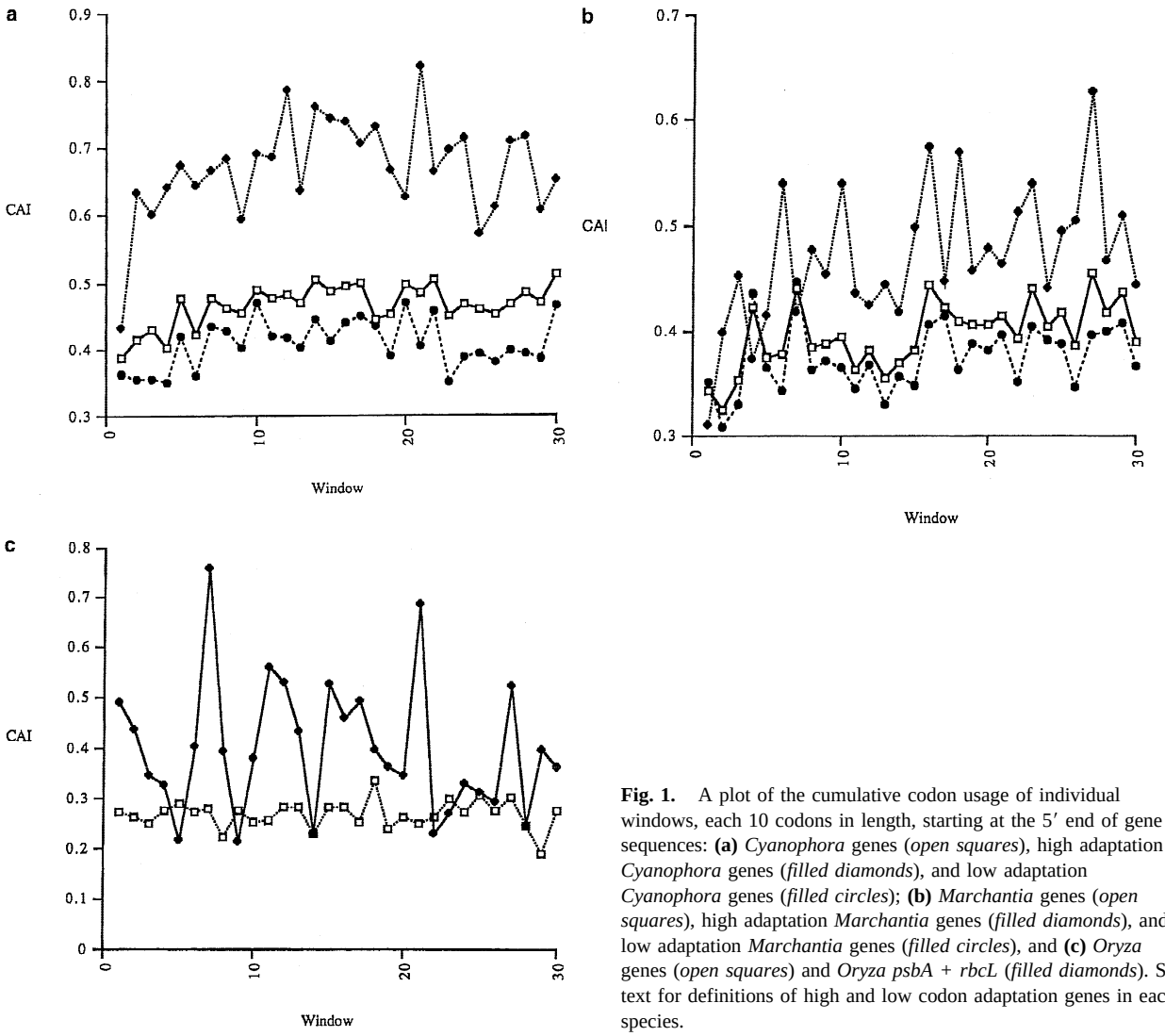

Fig. 1. A plot of the cumulative codon usage of individual windows, each 10 codons in length, starting at the 5′ end of gene sequences: **(a)** *Cyanophora* genes (*open squares*), high adaptation *Cyanophora* genes (*filled diamonds*), and low adaptation *Cyanophora* genes (*filled circles*); **(b)** *Marchantia* genes (*open squares*), high adaptation *Marchantia* genes (*filled diamonds*), and low adaptation *Marchantia* genes (*filled circles*), and **(c)** *Oryza* genes (*open squares*) and *Oryza psbA + rbcL* (*filled diamonds*). See text for definitions of high and low codon adaptation genes in each species.

measure involved calculating 18 new CAI values for each gene, referred to as XCAI values. Each of the 18 XCAI values for a given gene was calculated in the same manner as CAI with the sole exception being that one of the amino acids with multiple codons was excluded. Therefore, for any given gene, the content of an amino acid and the XCAI value calculated by excluding that amino acid are independent measures.

To examine codon adaptation along the length of gene sequences, sets of genes were analyzed cumulatively. Every gene in the set under analysis was divided into nonoverlapping windows 10 codons in length, beginning with the start codon. Codons within the same window from each gene were tabulated to generate a cumulative codon usage table for that window. A CAI value was then calculated separately for each window based on the cumulative codon usage. In addition, a random table generation was performed for each window using the approach described previously (Morton 1998). For every window, 500 replicate codon usage tables with the same amino acid usage as the observed table were generated at random using the base frequencies from noncoding regions of the same genome (Morton 1998). This approach makes the assumption that noncoding sequences accurately reflect composition bias in the absence of selection, even though there are certainly noncoding sites that are affected by selection. A total of 23,207, 20,499, and 36,123 noncoding sites were used for *Cyanophora*, *Marchantia*, and *Oryza*, respectively. Since the plastid genome codes for less than 100 genes, many of which are organized into operons, it is unlikely that selection acts on more than a relatively small proportion of all of these sites. Therefore, the assumption that these sequences accurately reflect the genome composition bias is probably valid. For each random table, CAI was calculated and from the 500 replicates a distribution generated. The mean of this distribution is the expected CAI value given genome composition bias (taking noncoding regions as an estimate of this bias). An observed CAI that was more than two standard deviations above the mean was considered evidence that composition bias alone cannot account for cumulative codon adaptation of that window. All codon usage analyses were performed using programs written by the authors in Pascal on a G3 Power Macintosh.

Results

Codon Adaptation Along the Length of Genes. The codon adaptation scans from three species are shown in Fig. 1. The three species are *Cyanophora* (Fig. 1a), *Marchantia* (Fig. 1b), and *Oryza* (Fig. 1c). To compare genes with relatively high and relatively low codon adaptation levels within each species, three plots are presented for *Cyanophora* and *Marchantia* and two for *Oryza*. Each plot

represents a different set of genes and shows the CAI calculated from the cumulative codon usage of those genes in each of 30 nonoverlapping windows starting at the 5′ end. For *Cyanophora* the three plots represent all genes, high codon adaptation genes (8 genes with CAI values over 0.60), and low codon adaptation genes (41 genes with CAI values under 0.45). The three plots from *Marchantia* are also for all genes, high codon adaptation genes (7 genes with CAI values over 0.425) and low codon adaptation genes (all 33 remaining genes). In the case of *Oryza*, the cumulative codon adaptation of all genes is plotted as well as a plot for the two high codon adaptation genes, *rbcL* and *psbA* (Morton 1998). A low adaptation plot was excluded for *Oryza* since it is essentially identical to the full plot (data not shown).

The results of the codon adaptation scans in Fig. 1 are quite similar to what has been observed in *E. coli*, where high adaptation genes have decreased adaptation at the 5′ end (Eyre-Walker and Bulmer 1993). In high adaptation genes from both *Cyanophora* and *Marchantia* the first window—which represents the first 10 codons—has a noticeably lower codon adaptation than any other window. On the other hand, there is no such decrease at the 5′ end of any set of genes from *Oryza*, a species which has very weak selection. It is also noticeable that for all three species there is some fluctuation in adaptation along the length of coding sequences but the significance of this is difficult to judge. The coding content of the three genomes is quite low, being well under 100 protein-coding genes, and it is apparent that the degree of fluctuation decreases with increasing number of genes; the greatest fluctuation is seen in the plot from *Oryza* that includes only two genes, *psbA* and *rbcL*, while the plots from each species that include all genes show much less variation. Despite this fluctuation, the decrease in codon adaptation at the 5′ end of high adaptation genes from both *Cyanophora* and *Marchantia* (Figs. 1a and b) and the similarity to the decrease at the 5′ end of high adaptation *E. coli* genes is striking.

We also tested high codon adaptation genes from three other species of algae to determine if there was a similar decrease in adaptation at the 5′ end. Plots of CAI along the length of high adaptation genes from *Chlamydomonas*, *Chlorella* and *Odontella* are shown in Fig. 2. These three species were chosen since they all show evidence for at least an intermediate selection for codon usage (Morton 1998, 1999). In all three cases either the first window (10 codons) or the first two windows (20 codons) have a lower CAI than any other section of the gene.

These results indicate that plastid genes with high overall codon adaptation tend to have a relatively low level of adaptation over the first 10 to 20 codons. This decreased codon adaptation at the 5′ end of genes with relatively high CAI values is not, however, a decrease below the expected codon adaptation given the genome

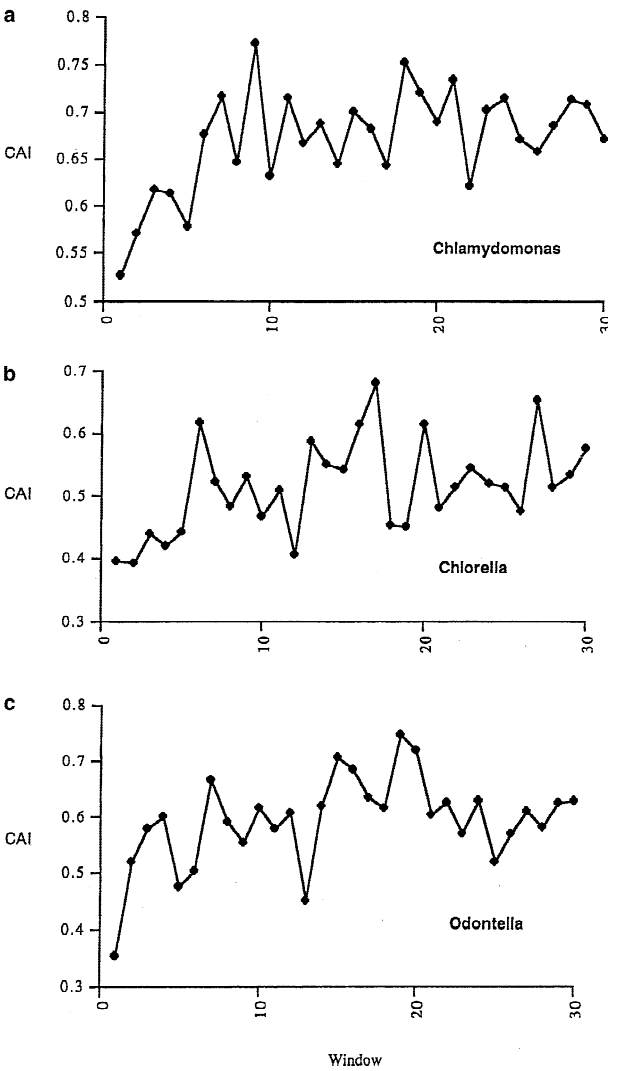

Fig. 2. A plot of the cumulative codon usage starting at the 5′ end of gene sequences for **(a)** *Chlamydomonas* genes with CAI values over 0.70, **(b)** *Chlorella* genes with CAI values over 0.45, and **(c)** *Odontella* genes with CAI values over 0.50.

composition bias. In high adaptation genes from each species, a random generation test (see Materials and Methods) showed that even within the first window, CAI was significantly higher than expected (data not shown). Therefore, although lower than the rest of the gene, there is still evidence that codon usage is under selection to increase the usage of major codons over the level generated by composition bias, similar to the findings for *E. coli* (Eyre-Walker and Bulmer 1993).

Codon Usage and Composition of the 3′ Neighboring Sites. A comparison of codon usage to the composition of the 3′ neighboring codon reveals that codon content is important for most twofold degenerate codon groups with a pyrimidine at the third position (NNY codon groups). In each of the codon groups AAY, TTY, TAY, GAY, and CAY, the major codon is the C-terminating codon (NNC codon) and highly expressed

Table 1. Major codon usage and context in low and high bias *Cyanophora* genes

	High adaptation genes[a]				Low adaptation genes[b]			
	3′ base				3′ base			
Codon	G	A	T	C	G	A	T	C
CAT	8	6	3	1	17	19	27	15
CAC	9	6	6	3	5	0	2	3
GAT	45	17	20	9	35	53	39	21
GAC	11	11	11	9	4	6	5	7
AAT	16	6	6	2	50	89	111	38
AAC	16	18	14	11	4	11	12	5
TAT	20	13	9	6	36	51	59	36
TAC	13	14	10	7	6	6	4	4
TTT	24	7	3	1	77	103	114	32
TTC	14	18	22	15	3	15	22	8
Total T	113	49	41	19	215	315	350	142
Total C	63	67	63	45	22	38	45	27

[a] Defined as genes with a CAI value greater than 0.6.
[b] Defined as genes with a CAI value less than 0.4.

Table 2. Major codon usage of twofold degenerate groups and context in the *Cyanophora psbA* gene

	3′ base			
Codon	G	A	T	C
CAT	1	0	0	0
CAC	2	3	3	1
GAT	3	0	1	0
GAC	0	3	0	0
AAT	4	0	0	0
AAC	3	5	6	3
TAT	1	2	0	0
TAC	1	3	2	3
TTT	3	4	1	1
TTC	3	5	6	6
Total T	12	6	2	1
Total C	9	19	17	13

Table 3. Major codon usage of twofold degenerate groups and context in high CAI genes of different organisms

		3′ base			
Organism and CAI range	Codon type[a]	G	A	T	C
Chlamydomonas	NNT	60	33	18	9
CAI >0.70	NNC	117	102	110	85
Chlorella	NNT	70	49	42	22
CAI >0.45	NNC	45	44	33	41
Odontella	NNT	149	102	55	36
CAI >0.5	NNC	85	58	97	55
Marchantia	NNT	64	51	28	21
CAI >0.45	NNC	16	26	27	19
Oryza	NNT	36	29	17	15
psbA + rbcL	NNC	15	26	19	16

[a] Cumulative usages of NNT and NNC codons are given for the codon groups AAY, TAY, TTY, GAY, and CAY (see text).

plastid genes have a bias toward these codons. Other genes display a strong bias toward the NNT codons, which is consistent with the composition bias of plastid genomes (Morton 1998).

The bias in highly expressed genes toward the NNC codons, however, turns out to be dependent on context. A comparison of the codon usage of high and low adaptation genes from *Cyanophora* is given in Table 1, where codon usage is broken down as a function of the composition of the 3′ nucleotide. Although high adaptation genes show an overall bias toward NNC codons, when the 3′ neighboring base (the first position of the 3′ codon) is a G the bias is toward NNT codons. The bias toward NNT in this context is not as strong as in low expression genes but it is significantly different than other contexts in high adaptation genes (a heterogeneity test gives $\chi^2 = 31.8$, $p < 0.001$). Overall for NNY codons upstream of a G, 64% (113 of 176) are NNT in high bias genes, while at all other sites, only 38% (109 of 284) are NNT. In low bias genes, 89% (1022 of 1154) of all NNY sites have an NNT codon. This decreased usage of NNC relative to NNT at sites upstream from a G is observed in each of the codon groups (Table 1).

This context effect is observed within individual genes and in plastid genes from other species. In the *psbA* gene, which has the strongest adaptation of all *Cyanophora* cyanelle genes, the decreased bias toward NNC upstream of GNN codons is observed (Table 2). In addition, high adaptation genes from every plastid genome show the same context dependency; in each case the bias toward NNC codons is weakest upstream from a G and tends to be strongest upstream of a C (Table 3). *Chlamydomonas* shows the weakest context dependency but it is still observed; 117 of 177 NNY codons (66%) are NNC upstream from a G but in other contexts the proportion is 83% (297 of 357). A heterogeneity test shows that this difference is significant ($\chi^2 = 20.3$, $p < 0.001$).

Although significant heterogeneity in codon usage among different contexts is observed for the NNY codon groups, this is not observed in other codon groups. Table 4 shows the codon usage of fourfold and sixfold degenerate codon groups from high adaptation genes of *Cyanophora* relative to the 3′ neighboring base composition. None of the codon groups show significant heterogeneity. Further, no significant heterogeneity is observed in the fourfold and sixfold degenerate groups from high adaptation genes of *Marchantia* and *Oryza*, nor is any observed in the NNR codon groups of any species (data not shown). Therefore, only in the NNY codon groups is a correlation observed between codon usage and context.

Table 4. The effect of context on codon usage of fourfold and sixfold degenerate codon groups from high adaptation genes of *Cyanophora*

	3′ base				
Codon	G	A	T	C	Heterogeneity test
CTG	0	0	2	0	
CTA	5	2	0	0	
CTT	8	7	6	8	
CTC	0	0	0	0	
TTG	1	0	0	0	
TTA	66	49	26	24	$\chi^2 = 20.09$ NS
TCG	0	0	0	0	
TCA	0	2	1	1	
TCT	31	15	18	11	
TCC	6	7	1	3	
AGT	7	5	9	4	
AGC	7	7	8	2	$\chi^2 = 12.18$ NS
CGG	0	0	0	0	
CGA	0	0	0	0	
CGT	49	11	24	13	
CGC	3	1	0	1	
AGG	0	0	0	0	
AGA	6	6	2	6	$\chi^2 = 11.28$ NS
CCG	0	2	3	0	
CCA	23	13	7	9	
CCT	14	5	8	10	
CCC	0	0	0	0	$\chi^2 = 11.95$ NS
ACG	0	0	0	0	
ACA	10	9	7	0	
ACT	44	23	22	17	
ACC	3	2	2	2	$\chi^2 = 6.52$ NS
GTG	0	1	0	0	
GTA	42	19	15	9	
GTT	37	16	18	11	
GTC	0	0	0	0	$\chi^2 = 4.68$ NS
GCG	1	4	5	3	
GCA	45	26	23	18	
GCT	52	35	29	18	
GCC	0	0	0	0	$\chi^2 = 6.28$ NS
GGG	2	1	0	0	
GGA	4	5	1	5	
GGT	56	56	36	25	
GGC	4	0	1	4	$\chi^2 = 14.39$ NS

Codon Bias and Amino Acid Content. Comparisons were made between codon adaptation, measured by CAI, and amino acid composition for all genes from *Cyanophora*, *Marchantia*, and *Oryza*, as examples of plastid genomes with strong, intermediate, and weak selection (see Materials and Methods). In Table 5 the correlation coefficient between CAI and the proportion of each amino acid is given for the three species. For *Cyanophora* the correlation between the frequency of each of Leu, Val, Ala, Gly, Ile, Glu, Asp, Asn, Lys, and codon adaptation is significant at the 5% level using Fisher's *Z* transformation (Bliss 1967). In the case of *Marchantia*, which is under less stringent codon bias selection than *Cyanophora*, six amino acids show a significant correlation (Table 5). All six are amino acids that have significant correlations in *Cyanophora* and the relative degree of correlation among the six amino acids is quite consistent. For *Oryza* there is no significant correlation between the proportion of any amino acid in a gene and codon adaptation.

To test whether or not the results in Table 5 are affected by an influence of amino acid composition on the calculation of CAI, two additional comparisons were made. The first involved a comparison of the variation in both BCAI (which is independent of amino acid usage; see Materials and Methods) and amino acid content. The second approach was to calculate 18 different XCAI values for each gene, where each XCAI excluded 1 of the 18 degenerate amino acids from consideration (see above). The correlation between the variation among genes in the content of a specific amino acid and the XCAI values calculated excluding that amino acid, was then determined. The results for both BCAI and XCAI are presented in Table 5. Although there are a few cases in which we find a significant correlation in one comparison but not the other, the results are very similar for all three measures of codon adaptation. Therefore, the correlation between amino acid content and CAI is not generated by an influence of amino acid usage on the calculation of codon adaptation.

A more general comparison shows that amino acids coded by more G + C-rich codons, and GNN codons in particular, tend to be positively correlated with codon bias while those coded by A + T-rich codons tend to be negatively correlated. All five amino acids coded by a GNN codon have a significant positive correlation and there is a significant correlation between G + C content of the first two codon positions and codon adaptation in *Cyanophora* but not *Oryza* (Table 6). The correlation is strongest between codon adaptation and first codon position G content as well as between codon adaptation and C content at the second codon position. As a result, genes with a relatively high codon adaptation in *Cyanophora* have a very different amino acid content than those with lower adaptation, which is demonstrated by a Square Plot (Fig. 3) adapted from Foster et al. (1997).

Since many highly expressed plastid genes are membrane proteins, such as photosystem proteins and those involved in electron transport, a correlation between codon adaptation and amino acid content may be a secondary result of functional constraints on membrane as opposed to cytosolic proteins. To examine this possibility, two sets of genes from *Cyanophora* were studied separately. The first consisted of genes coding for products that act in the cytosol (such as RuBisCO and ribosomal proteins) and genes coding for membrane proteins (including photosystem I and II proteins as well as ATP synthase subunits). Both sets showed a significant correlation between G + C content and codon bias ($r = 0.602$ for cytosolic proteins and $r = 0.680$ for membrane proteins).

Table 5. Coefficient of correlation between codon adaptation and amino acid frequency in *Cyanophora, Marchantia,* and *Oryza*[a]

	Cpa			*Mpo*			*Osa*		
Amino acid	CAI	BCAI	XCAI	CAI	BCAI	XCAI	CAI	BCAI	XCAI
Leu (CTN, TTR)	−0.296	−0.349	−0.385	NS	NS	NS	NS	NS	NS
Ser (TCN, AGY)	NS	NS	NS	NS	NS	NS	NS	NS	NS
Arg (CGN, AGR)	NS	NS	NS	NS	NS	NS	NS	NS	NS
Pro (CCN)	NS	NS	NS	NS	NS	NS	NS	NS	NS
Thr (ACN)	NS	NS	NS	NS	NS	NS	NS	NS	NS
Val (GTN)	0.373	0.340	0.318	0.399	0.315	0.335	NS	NS	NS
Ala (GCN)	0.678	0.602	0.654	0.513	0.370	0.469	NS	NS	NS
Gly (GGN)	0.293	0.452	0.336	0.345	0.408	0.447	NS	NS	NS
Ile (ATA, ATY)	−0.671	−0.606	−0.598	−0.461	NS	NS	NS	NS	NS
His (CAY)	NS	NS	NS	NS	NS	NS	NS	NS	NS
Gln (CAR)	NS	NS	NS	NS	NS	NS	NS	NS	NS
Glu (GAR)	0.213	NS	NS	NS	NS	NS	NS	NS	NS
Asp (GAY)	0.329	0.209	0.241	0.399	NS	0.299	NS	NS	NS
Asn (AAY)	−0.407	−0.358	−0.271	NS	NS	NS	NS	NS	NS
Lys (AAR)	−0.408	−0.537	−0.468	−0.436	−0.518	−0.599	NS	NS	−0.318
Tyr (TAY)	NS	NS	NS	NS	NS	NS	NS	NS	NS
Cys (TGY)	NS	NS	NS	NS	NS	NS	NS	NS	NS
Phe (TTY)	NS	NS	NS	NS	NS	NS	NS	NS	NS
Met (ATG)	0.321	0.278	0.321	NS	NS	NS	NS	NS	NS
Trp (TGG)	NS	NS	NS	NS	NS	NS	NS	NS	NS

[a] Correlation coefficients for the relationship between codon adaptation and amino acid content. Three measures of adaptation were used, CAI, BCAI, and XCAI (see text). Correlations are given for *Cyanophora* (*Cpa*), *Marchantia* (*Mpo*), and *Oryza* (*Osa*). Values that are not significant at the 5% level are marked NS.

Discussion

All three features of codon usage bias that have been described for *E. coli* genes are shown here to exist in plastid genes. First, plastid genes with high overall codon adaptation have a noticeably lower codon adaptation at the 5′ end, particularly the first 10 to 20 codons (Figs. 1 and 2). Second, codon context is significant; some major codons are not used preferentially in certain contexts (Tables 1 and 3). Finally, amino acid content is significantly correlated with variation in codon adaptation in genomes under selection for codon usage (Tables 5 and 6, Fig. 3).

Low Codon Adaptation at the 5′ End. In the high adaptation genes from *Cyanophora* and *Chlamydomonas,* which have the highest CAI values of all plastid genes that have been studied to date (Morton 1998), there is a marked decrease in the cumulative codon adaptation at the 5′ end of the sequences (Figs. 1a and 2). This decrease is not observed in plastid genes with low overall adaptation (Fig. 1c) but is found in high adaptation genes from other species (Figs. 1b and 2). In general, genes with high overall codon adaptation display a relatively low adaptation over the first 10–20 codons.

It has been noted that rare codons are used at relatively high frequency over the first 25 codons of *E. coli* genes (Chen and Inouye 1990, 1994), which shows up as a decreased codon adaptation at the 5′ end of these genes, particularly those with high overall codon adaptation (Eyre-Walker and Bulmer 1993). It has been suggested that the increased representation of rare codons, particularly AGG, at the 5′ end is part of a cellwide gene regulation mechanism in which expression could be manipulated by altering tRNA levels (Chen and Inouye 1990, 1994). Alternatively it has been proposed that selection on some factor other than codon usage, such as mRNA structure or ribosome binding, acts at the 5′ end of genes and is stronger than selection for translation efficiency at degenerate sites within this region (Eyre-Walker and Bulmer 1993).

The available data do not appear to support one model over the other. Comparing observed codon adaptation of

Table 6. Correlation coefficients for the comparison of G + C content of the first two codon positions and codon bias[a]

	Cpa[a]	*Osa*
1st position		
G	0.573	NS
C	NS	NS
G + C	0.563	NS
2nd position		
G	0.304	NS
C	0.544	NS
G + C	0.564	NS
1st + 2nd positions		
G	0.605	NS
C	0.511	NS
G + C	0.632	NS

[a] Correlation coefficients are given. Values that are not significant at the 5% level are marked NS. *Cpa, Cyanophora; Osa, Oryza.*

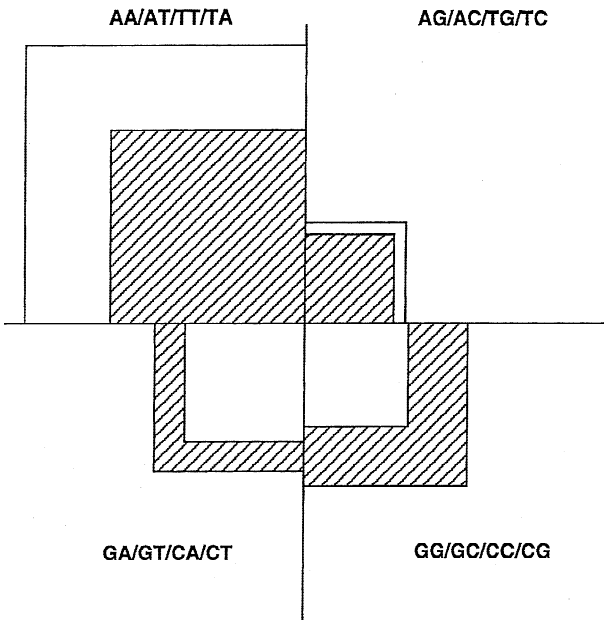

Fig. 3. A square plot (Foster et al. 1997) of high adaptation (CAI values over 0.50; *hatched*) and low adaptation (CAI values less than 0.40; *open*) *Cyanophora* cyanelle genes. The four regions represent four sets of amino acids with different A + T compositions over the first two codon positions. These compositions are indicated in each region. The area plotted in each region represents the frequency of the codons that region represents and, therefore, the amino acids coded by those codons. High adaptation genes use a noticeably higher frequency of GC/GG/CC/CG codons but a lower frequency of AA/AT/TT/TA codons.

plastid genes to an expectation generated from random sequences based on noncoding base frequencies (see Materials and Methods) shows that, although codon adaptation at the 5′ end of high adaptation genes is lower than other parts of the sequence, it is still significantly higher than expectation. This is found in all plastid gene sets in which cumulative codon adaptation is relatively low at the 5′ end (data not shown) as well as in *E. coli* (Eyre-Walker and Bulmer 1993). This observation that major codons are utilized at a frequency that is higher than would be expected in the absence of selection seems inconsistent with a system to down-regulate genes. On the other hand, the presence of rare codons at the 5′ end of a gene is known to decrease the expression level of that gene (Chen and Inouye 1994), indicating that the potential exists for rare codons to be involved in some manner of gene regulation. As a result, the reason for the decreased codon adaptation remains unclear.

Context and the Bias Toward Major Codons. The data presented here indicate that particular major codons are not used preferentially at all sites but, rather, only in specific contexts. In high adaptation plastid genes, each NNY twofold degenerate codon group shows an overall bias toward the NNC codon, while in low adaptation genes the bias is toward NNT (Morton 1993). In high adaptation genes, however, NNY sites with a 3′ neighboring G have a significantly lower bias toward NNC (in the case of *Chlamydomonas*) or a bias toward NNT codons (Tables 1 and 3). The same influence of a 3′ G on NNC codon usage is observed in high adaptation plastid genes from the red alga *P. purpurea,* black pine, and tobacco, species not presented in Table 3 (data not shown). Overall, every high adaptation plastid gene examined shows the same codon context effect.

One possible explanation for this pattern of codon usage is an avoidance of CpG sites. Some CpG sites are methylated in mammalian DNA and this dinucleotide is highly underrepresented in mammalian genomes (Bulmer 1986; Karlin and Mrazek 1996). Although it is not clear whether or not CpG sites are methylated in plastid sequences, CpG avoidance fails to explain the codon context of plastid sequences. First, NNC codons of fourfold degenerate groups are used at the same frequency upstream of all bases (Table 4 and data not shown), which does not suggest a general composition feature. Second, CpG is not generally avoided in plastid noncoding regions, as it is in mammalian DNA. For example, in *Cyanophora,* the G content of noncoding regions is 8.6% but CpG makes up 10.9% of CpN dinucleotides and 1.15% of all dinucleotides (relative to the expected proportion of 0.91%). Therefore, an avoidance of CpG is not a general compositional feature but is specific to NNY codon groups.

Since this codon context dependency is observed in high adaptation genes, the most likely explanation, given the general model for selection on codon usage, is that context is related to translation efficiency. In addition, since context dependence is limited to certain codon groups (Table 4), neighboring base influence is not consistent. In *E. coli,* codon context is known to have an effect on translation (Murgola et al. 1984; Shpaer 1986) but the details of how this might happen for plastid genes are uncertain. Interestingly, NNC codons are major codons for the NNY codon groups in *E. coli* but the same sort of context dependency is not observed (data not shown). This means that codon context effects must be different in the two systems, suggesting that they are a property of the overall translation apparatus.

Regardless of the underlying cause, the significant context specificity of NNC major codons indicates that a simple measure of major codon usage might not be sufficient. It may be necessary to adjust measures of codon adaptation in order to account for context of major codons. In addition, it suggests that site-by-site selection may be more complex than usually assumed, which may help explain the observation that rare codons appear to be conserved at specific sites of highly expressed bacterial genes (Maynard Smith and Smith 1996).

Codon Bias and Amino Acid Content. It has long been known that the base composition bias of a genome, or genomic region, influences the amino acid content of coding sequences (Sueoka 1961; Jukes and Bhushan

1986; D'Onofrio et al. 1991; Porter 1995; Foster et al. 1997). More recently, a correlation between codon adaptation and amino acid content has been noted in *E. coli* (Lobry and Gautier 1994; Gutierrez et al. 1996). The correlation has been explained as a result of selection for increased translation efficiency, such that highly expressed genes tend to have increased proportions of certain amino acids because the codons coding these are translated more efficiently (Gutierrez et al. 1996). The basic notion is that a number of sites within a protein may be able to tolerate any of a given set of amino acids and selection will favor one that is coded by a codon that is translated more efficiently.

The evidence presented here suggests that a similar phenomenon occurs in plastid genes. In the *Cyanophora* genome there is a significant correlation between codon adaptation and the relative frequency of certain amino acids. High codon adaptation genes tend to have higher frequencies of Ala, Val, Gly, and Asp and lower frequencies of Ile, Leu, Lys, and Asn (Table 5). None of the amino acids within each group share any obvious functional properties; for example, alanine and valine are apolar, glycine is uncharged, and aspartic acid has a charged side group. However, the codons coding these amino acid sets do share an obvious property: they all have a G at the first position. More generally there is a significant decrease in amino acids with A + T at the first two codon positions in high adaptation genes (Table 6). A correlation is observed separately, and with roughly the same correlation coefficient, in both cytosolic and membrane protein-coding genes. Alternative measures of codon adaptation that are fully independent of amino acid content (BCAI and XCAI) give very similar results, demonstrating that the correlations are not an artifact of an influence of amino acid content on the calculation of CAI.

The same correlation between amino acid content and codon adaptation is observed in *Marchantia* to a weaker extent but is not observed in *Oryza.* Therefore, the existence of a correlation itself is associated with strong selection on codon usage to increase translation efficiency. It is proposed here that the amino acid composition of plastid genes that are under strong codon adaptation selection is influenced to some degree by translation considerations. Codons that are rich in G + C at the first two positions, and GNN codons in particular, are favored over other codons at those sites which are neutral with regard to the functional differences between the amino acids.

It is interesting that, while this feature of codon usage is similar to *E. coli,* there are some fundamental differences. In *E. coli,* codons with a purine at the first position, and particularly GNN codons, are used at a higher frequency in high codon adaptation genes than in low codon adaptation genes (Gutierrez et al. 1996). To examine this further, we used the high and low adaptation

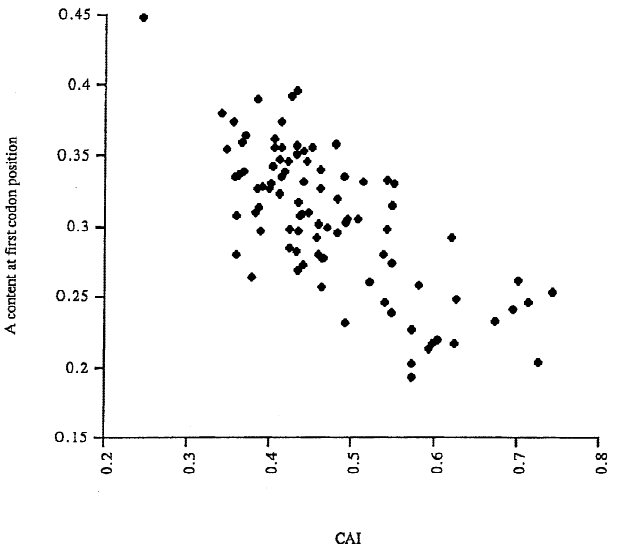

Fig. 4. A comparison of the first codon position A content and codon adaptation of *Cyanophora* cyanelle genes.

gene sets from Lobry and Gautier (1994) and compared RNN and YNN codon proportions in the two sets. For every possible pair of amino acids, excluding sixfold degenerate groups, we calculated the value (RNN)/(RNN + YNN) based on the cumulative amino acid usage within both sets of genes. The value for the low adaptation genes was then subtracted from the value for the high adaptation genes and significance measured by assuming two binomial distributions. If a statistically significant and positive difference was observed, the RNN amino acid was considered "preferred" in high adaptation genes, while if there was a significant negative difference, the YNN amino acid was considered "preferred." There is no pair of amino acids, regardless of chemical similarity, in which the RNN/(RNN + YNN) value is smaller in high adaptation genes than in low adaptation genes (data not shown). On the other hand, there are 70 pairwise amino acid comparisons for which the RNN/(RNN + YNN) value is significantly larger in high adaptation genes than low adaptation genes. Of these, there are 29 pairs for which an ANN codon is present at a higher relative frequency in high adaptation genes. Therefore, there is a general relative preference for RNN codons over YNN codons in high adaptation genes.

These findings for *E. coli* are in contrast to high adaptation plastid genes, which show a decreased usage of every ANN codon with the exception of ATG which codes methionine (Table 5). Overall, though, there is a strong negative correlation in *Cyanophora* between CAI and the frequency of ANN codons ($r = -0.721$; Fig. 4). This is in contrast to a correlation coefficient of +0.055 from *E. coli* (Gutierrez et al. 1996). Therefore, while high adaptation plastid genes share an increased relative usage of GNN codons with *E. coli* high adaptation genes, the plastid genes differ in that they tend to have a de-

creased usage of ANN codons. This suggests that the mechanism underlying the influence on translation efficiency is different in plastids and *E. coli*.

Conclusions. A number of recent studies have made it evident that plastid genes have several features of codon usage that are shared with *E. coli* genes. It has been established previously that, as is the case in *E. coli*, highly expressed plastid genes, particularly in algae, are strongly biased toward a set of major codons, apparently to increase translation efficiency (Morton 1993, 1998, 1999). The current work shows that plastid genes share additional features, specifically a decreased adaptation at the 5′ end, codon context, and a correlation between codon adaptation and amino acid usage. Overall, it is now apparent that codon usage is more complex than previously believed for plastid genes, a finding that has potential implications for gene manipulation. The finding that both codon context and amino acid content are related to codon adaptation suggests that a simple evaluation of major codon versus minor codon is not sufficient. In addition, these two features are of potential importance to the manipulation of any sequences that are to be introduced into plastids and expressed. Future studies that directly examine how these codon usage features influence translation efficiency will help both to test the translation model of selection on codon usage and to determine how most effectively to manipulate coding sequences in order to optimize expression in transformed plants.

Acknowledgments. This work was supported in part by NSF Grant MCB-9727906. We would like to thank Adam Eyre-Walker and two anonymous reviewers for useful suggestions regarding the analyses presented.

References

Andersson SGE, Kurland CG (1990) Codon preferences in free-living microorganisms. Microbiol Rev 54:198–210

Bliss CI (1967) Statistics in biology, Vol 1. McGraw–Hill, New York

Bulmer M (1986) Neighboring base effects on substitution rates in pseudogenes. Mol Biol Evol 3:322–329

Chen GT, Inouye M (1990) Suppression of the negative effect of minor arginine codons on gene expression; Preferential usage of minor codons within the first 25 codons of the *Escherichia coli* genes. Nucleic Acids Res 18:1465–1473

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

D'Onofrio G, Mouchiroud D, Aissani B, Gautier C, Bernardi G (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J Mol Evol 32:504–510

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

Foster PG, Jermiin LS, Hickey DA (1997) Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol 44:282–288

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

Gutierrez G, Marquez L, Marin A (1996) Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translational efficiency. Nucleic Acids Res 24: 2525–2527

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–35

Jukes TT, Bhurshan V (1986) Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes. J Mol Evol 24:39–44

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

Lobry JR, Gautier C (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res 22:3174–3180

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

Morton BR (1993) Chloroplast DNA codon use: Evidence for selection at the *psbA* locus based on tRNA availability. J Mol Evol 37:273–280

Morton BR (1996) Selection on the codon bias of *Chlamydomonas reinhardtii* chloroplast genes and the plant *psbA* gene. J Mol Evol 43:28–31

Morton BR (1998) Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J Mol Evol 46:449–459

Morton BR (1999) Codon bias and the context dependency of nucleotide substitutions in the evolution of plastid DNA. Evol Biol (in press)

Murgola EJ, Pagel FT, Hijazi KA (1984) Codon context effects in missense suppression. J Mol Biol 175:19–27

Porter TD (1995) Correlation between codon usage, regional genomic nucleotide composition, and amino acid composition in the cytochrome P-450 gene superfamily. Biochim Biophys Acta 1261:394–400

Sharp PM (1991) Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium:* Codon usage, map position, and concerted evolution. J Mol Evol 33:23–33

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

Sharp PM, Matassi G (1994) Codon usage and genome evolution. Curr Opin Gen Dev 4:851–860

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

Sueoka N (1961) Compositional correlation between deoxyribonucleic acid and protein. Cold Spring Harbor Symp Quant Biol 26:35–43

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