Selection at the Amino Acid Level Can Influence Synonymous Codon Usage: Implications for the Study of Codon Adaptation in Plastid Genes

Brian R. Morton

Department of Biological Sciences, Barnard College, Columbia University, New York, New York 10027

Manuscript received December 18, 2000
Accepted for publication June 27, 2001

ABSTRACT

A previously employed method that uses the composition of noncoding DNA as the basis of a test for selection between synonymous codons in plastid genes is reevaluated. The test requires the assumption that in the absence of selective differences between synonymous codons the composition of silent sites in coding sequences will match the composition of noncoding sites. It is demonstrated here that this assumption is not necessarily true and, more generally, that using compositional properties to draw inferences about selection on silent changes in coding sequences is much more problematic than commonly assumed. This is so because selection on nonsynonymous changes can influence the composition of synonymous sites (*i.e.*, codon usage) in a complex manner, meaning that the composition biases of different silent sites, including neutral noncoding DNA, are not comparable. These findings also draw into question the commonly utilized method of investigating how selection to increase translation accuracy influences codon usage. The work then focuses on implications for studies that assess codon adaptation, which is selection on codon usage to enhance translation rate, in plastid genes. A new test that does not require the use of noncoding DNA is proposed and applied. The results of this test suggest that far fewer plastid genes display codon adaptation than previously thought.

CODON bias is generally thought to result from the interplay of two forces, mutation bias and selection between synonymous codons (Li 1987; Akashi and Eyre-Walker 1998; Duret and Mouchiroud 1999). Since the time codon bias was first observed there has been a great deal of interest in determining the relative importance of these two forces as well as the factors that contribute to natural selection favoring one codon over a synonym. There is now strong evidence that in certain species selection discriminates between synonymous codons due to differences in translation efficiency (Ikemura 1985; Sharp 1991; Akashi 1995; Morton 1998, 2000). These differences result in codon adaptation in the highly expressed genes of these organisms, which is a bias toward those codons ("major" codons) that are complementary to abundant tRNAs (Ikemura 1985; Andersson and Kurland 1990; Bulmer 1991). Evidence for codon adaptation in highly expressed genes has now been advanced for a number of unicellular organisms (Ikemura 1985; Sharp 1991), Drosophila (Akashi 1994, 1995), and plastid genomes (Morton 1993, 1998, 2000).

A number of sequence comparison methods have been developed to infer whether or not codon usage of a gene or organism has been influenced by selection. Variations of the MacDonald-Kreitman test have proven effective in determining the influence of selection on the evolution of a particular sequence when polymorphism data are available (MacDonald and Kreitman 1991; Akashi 1999). In their absence a common approach is to compare compositional properties of silent sites to those of noncoding DNA. The way that mutation (M) and selection between synonymous changes (S_S) are commonly thought to generate codon bias can be represented very simply as

$$M + S_S \rightarrow \text{codon bias.}$$

This basic model leads to the common assumption that in the absence of selection on synonymous substitutions the composition of silent sites in protein-coding sequences will be affected only by the mutation bias and, therefore, the silent sites will all have the same equilibrium base frequencies and the same composition bias as neutral noncoding DNA.

This assumption has been the basis of many analyses. Comparisons between the composition of introns and silent sites of coding regions have been used in studies of codon usage in Drosophila (Carulli *et al.* 1993; Kliman and Hey 1994) and noncoding base frequencies have been used to generate expectations concerning codon frequencies in yeast (Percudani and Ottonello 1999). It has also been used to generate what has been referred to as parity rule 2 (PR2), which is the proposal that base frequencies at fourfold degenerate sites of coding sequences should be such that $f_A = f_T$ and $f_G = f_C$ in the absence of selective differences between synony-

Address for correspondence: Department of Biological Sciences, Barnard College, Columbia University, 3009 Broadway, New York, NY 10027. E-mail: bmorton@barnard.columbia.edu

mous codons (Chargraff 1979; Sueoka 1992, 1999), assuming that the mutation bias is equivalent in the two strands. The assumption is also implicit in the null hypothesis that the codon usage at conserved and variable amino acid sites should be equal, used by Akashi (1994) as the basis for a test of the influence of translation accuracy on codon usage.

Noncoding DNA has also been used to investigate codon adaptation in plastid genes. The composition of noncoding regions was used as an estimate of mutation bias so that it could be determined if the frequency of major codons in any given gene was significantly higher than expected from mutation bias alone (Morton 1998). For each gene, replicate random genes were generated with the same amino acid usage as the gene itself but with codon usage assigned randomly on the basis of resampling from the dinucleotide pool of noncoding regions from the same genome. The resulting distribution yielded an expected frequency of major codon usage in the absence of selection. The replicates also generated a variance so that selection could be inferred in cases where the observed level of major codon usage was significantly greater than the expectation. When this test was applied to plastid genes from a wide array of lineages it was found, generally, that >35% of genes from different algae and >10% of genes from land plants showed evidence for selection (Morton 1998).

The current work is a reconsideration of the analysis of plastid genes and others like it that utilize compositional properties. The possibility that a third factor, selection on amino acid composition of a protein (S_A), might influence the frequency at which synonymous codons are utilized is considered. Although the contribution of S_A to the relative usage of synonymous codons has been ignored previously, this work demonstrates that selection at the amino acid level could have a significant influence on the composition of synonymous sites. Therefore, the actual forces that interact to generate codon bias need to be represented as

$$M + S_S + S_A \rightarrow \text{codon bias.}$$

Although selection on amino acid usage has the potential to be a significant factor, it is essentially impossible to determine the magnitude of the influence because of the difficulties in evaluating the parameters involved. It does mean, though, that we cannot draw inferences concerning selective differences between synonymous codons by using the composition of noncoding DNA to estimate M: If we find that codon bias is significantly different from what is expected from M alone, we cannot conclude that $S_S \neq 0$, only that $S_S + S_A \neq 0$. As a result, the previous analysis of codon adaptation in plastid genes and other similar analyses were based on an invalid assumption and must be reconsidered.

This work is presented in two sections. The first section demonstrates that S_A has the potential to influence synonymous codon usage, even if all synonymous changes are neutral. In particular, it is shown that (1) the A + T content of a twofold degenerate site is a function of the strength of selection on amino acid usage at that residue and (2) selection at the amino acid level can give rise to asymmetry ($f_A \neq f_T$) at fourfold degenerate sites even when the mutation bias alone yields symmetry. Therefore, tests using compositional properties, such as the previous resampling test (Morton 1998) and PR2 (Sueoka 1999), to assess selection on silent changes must be treated with caution. In addition, the common method of comparing codon usage at conserved and variable sites to infer selective differences based on a need to translate certain codons more accurately (Akashi 1994) is invalid since the influence of S_A is different for conserved and variable sites. The second section of the article proposes and applies a novel recursive approach to assess codon adaptation of plastid genes. The results suggest that fewer plastid genes display codon adaptation than concluded previously.

MODELS OF SEQUENCE EVOLUTION AND CODON BIAS IN THE ABSENCE OF SELECTION ON CODON USAGE

First we use a simplified evolutionary model to demonstrate the manner in which selection at the amino acid level can influence the frequency with which synonymous codons are utilized. Because of the parameters involved it is not possible to determine if selection on amino acid usage is actually influencing the compositional pattern of silent sites in specific genes. Despite this, we are able to conclude that the composition bias of a neutral site in any protein-coding sequences cannot be generally expected to match either the composition of neutral noncoding sites or the composition of other silent sites in protein-coding sequences. The influence of amino acid selection on codon usage is examined first for twofold degenerate sites alone and then for all codon groups using a codon-codon transition matrix.

Twofold degenerate sites: The primary approach to studying the evolution of DNA sequences is to model it as a Markov process (Lewontin 1989; Goldman and Yang 1994; Muse and Gaut 1994; Gu and Li 1998; Lio and Goldman 1998; Yang *et al.* 1998; Yang and Nielsen 2000) in which the dynamics of change are represented by an $n \times n$ transition matrix $\mathbf{P}$, where $\mathbf{P}_{ij}$ is the probability of changing from state i to state j during a given time interval. Usually, the observed base (or codon) frequencies are assumed to be at equilibrium and then incorporated into $\mathbf{P}$, which can be used to derive measures for the comparison of homologous DNA sequence data. In this study we instead generate matrices as a function of mutation and selective pressure and then calculate the stationary state of the process, which will correspond to the equilibrium frequencies of either nucleotide or codon composition. The stationary vector

(ϕ) of a Markov transition matrix **P** is defined by the relationship

$$\phi = \phi\mathbf{P}. \quad (1)$$

The stationary vector for a nucleotide mutation model can be represented as (π_G, π_A, π_T, π_C). From any initial vector the process will tend toward ϕ and then remain at that composition. The stationary vector of a matrix **P** can be determined by calculating $\mathbf{\Pi} = \mathbf{P}^t$ for $t \rightarrow \infty$. The matrix **Π** is composed of n rows (for an $n \times n$ **P** matrix) of ϕ (Cox and Miller 1965). Therefore, given any transition matrix we can solve the equilibrium base frequencies or codon frequencies if it is a codon transition matrix.

We start with the nucleotide mutation matrix, **N**, defined as

$$\mathbf{N} = \begin{pmatrix} & \alpha\mathbf{1} & \beta\mathbf{1} & \beta\mathbf{2} \\ \alpha\mathbf{2} & & \beta\mathbf{4} & \beta\mathbf{3} \\ \beta\mathbf{3} & \beta\mathbf{4} & & \alpha\mathbf{2} \\ \beta\mathbf{2} & \beta\mathbf{1} & \alpha\mathbf{1} & \end{pmatrix}.$$

Only off-diagonals are given in **N**, and the values of the diagonal elements are such that each row sums to 1. All nucleotide matrices presented are in the order G, A, T, C. This general matrix allows us to address a number of possible mutation schemes and has the advantage that it yields symmetrical equilibrium frequencies ($\pi_A = \pi_T$ and $\pi_G = \pi_C$). It is also important that **N** is not necessarily time reversible, which is an assumption built into many common mutation models to simplify some calculations. However, there is no biological reason for making this assumption (Lio and Goldman 1998) so **N** is not constrained to be reversible.

The equilibrium A + T content for **N** can be determined from the equilibrium base frequencies

$$\pi_A = \pi_T = \frac{1}{2}\left(\frac{\alpha 1 + \beta 1}{\alpha 1 + \alpha 2 + \beta 1 + \beta 3}\right). \quad (2)$$

If we now express $\alpha 2$ and $\beta 3$ as functions of $\alpha 1$ and $\beta 1$, respectively, such that $\alpha 2 = \gamma 1\ \alpha 1$ and $\beta 3 = \gamma 2\ \beta 1$, then $\gamma 1$ and $\gamma 2$ measure the GC ↔ AT mutation pressure for transitions and transversions, respectively. The equilibrium A + T composition can now be expressed by

$$\mathrm{A} + \mathrm{T} = \left(\frac{\alpha 1 + \beta 1}{\alpha 1(1 + \gamma_1) + \beta 1(1 + \gamma 2)}\right). \quad (3)$$

This mutation model allows us to investigate how the A + T content of twofold degenerate sites is dependent on the strength of amino acid selection. The basic idea is that, if the GC ↔ AT pressure is different for transitions and transversions, then sites at which selection limits substitutions to transitions will have a different A + T content than neutral sites.

Let us assume that sites are undergoing mutations as represented by **N** and that all synonymous changes are neutral. We consider only the third codon position of twofold degenerate amino acids, at which either a purine or a pyrimidine is favored depending on the amino acid coded. (Changes at the first and second codon positions are dealt with below.) Ignoring positive selection, the influence of selection at the amino acid level can be modeled using the parameter s to represent the reduction in the frequency of transversions at twofold degenerate sites due to purifying selection. When $s = 1$ there is no selection and when $s = 0$ there are no transversions due to absolute constraints. Using this parameter we can write the substitution matrix for twofold degenerate sites as

$$\mathbf{N_S} = \begin{pmatrix} & \alpha\mathbf{1} & s\beta\mathbf{1} & s\beta\mathbf{2} \\ \alpha\mathbf{2} & & s\beta\mathbf{4} & s\beta\mathbf{3} \\ s\beta\mathbf{3} & s\beta\mathbf{4} & & \alpha\mathbf{2} \\ s\beta\mathbf{2} & s\beta\mathbf{1} & \alpha\mathbf{1} & \end{pmatrix}.$$

The equilibrium A + T content for sites evolving by $\mathbf{N_S}$ is a function of the strength of selection at the amino acid level as given by

$$\mathrm{A} + \mathrm{T} = \left(\frac{\alpha 1 + s\beta 1}{\alpha 1(1 + \gamma 1) + s\beta 1(1 + \gamma 2)}\right). \quad (4)$$

The relationship between A + T content and selection on the amino acid level is not straightforward but, rather, depends on the relative values of $\gamma 1$ and $\gamma 2$ as follows:

$\gamma 1 = \gamma 2$: In this case, Equation 4 reduces to A + T = $1 / (1 + \gamma 1)$ and s does not influence A + T content.

$\gamma 1 < \gamma 2$: Sites with lower s values (stronger selection) have a higher equilibrium A + T content.

$\gamma 1 > \gamma 2$: Sites with lower s values (stronger selection) have a lower equilibrium A + T content.

Therefore, as long as $\gamma 1 \neq \gamma 2$, the strength of selection on amino acid use influences the equilibrium A + T content. Differences in composition between sites with different levels of constraints at the amino acid level will be most pronounced when the relative rate of transversion mutations is higher and negligible when transversions are relatively rare. Interestingly, **N** is time reversible if and only if $\gamma 1 = \gamma 2$ so the general condition under which selection will influence the A + T content at twofold degenerate sites is when **N** is not time reversible.

Some implications of Equation 4 are shown in Figure 1, which plots the equilibrium A + T content of constrained twofold degenerate sites ($s < 1$) as a fraction of the equilibrium A + T content of unconstrained sites ($s = 1$) that have the same mutation matrix. (This ratio is referred to as AT_2:AT_{NC}.) Figure 1a demonstrates how AT_2:AT_{NC} varies with respect to $\gamma 2$ and with variation in

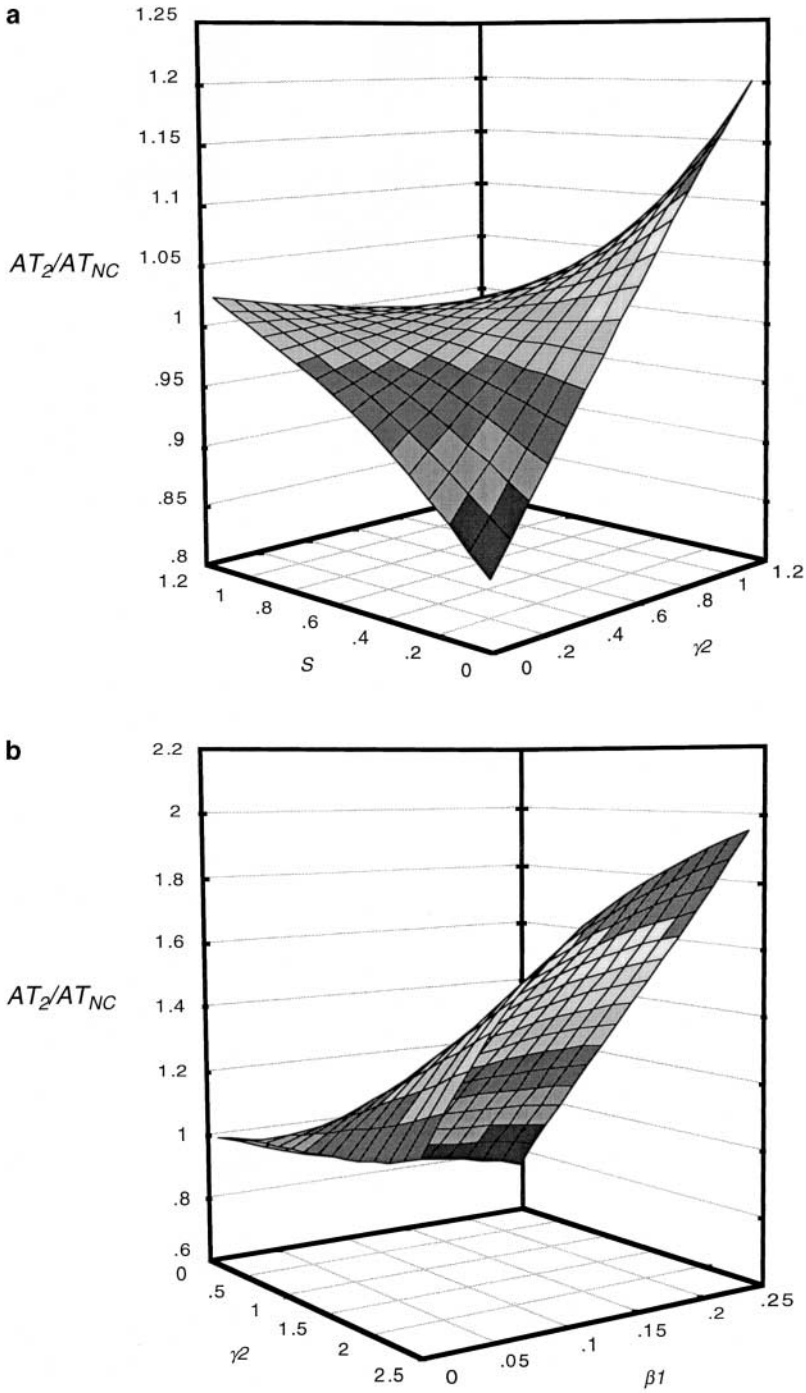

FIGURE 1.—The ratio of the equilibrium A + T content at twofold degenerate sites (AT_2) to the equilibrium A + T content of unconstrained sites (AT_{NC}) is plotted over a range of parameter values. The A + T contents were generated by applying Equation 4 in the text. In a, the mutation parameters $\alpha 1$ and $\alpha 2$ were set to 0.1 and 0.05 (meaning that $\gamma 1 = 0.5$) while $\beta 1$ was set to 0.2. Calculations were performed over a range of values of s (selective constraints, see text for an explanation of the parameters) and $\gamma 2$ (GC $\leftrightarrow$ AT pressure for transversions). The selective pressure on twofold degenerate sites varies from no selective constraints ($s = 1$) to absolute constraints ($s = 0$). The AT_{NC} values were calculated with $s = 1$ (no selective constraints). All calculations were performed over the range $\gamma 2 = 0.01$ to $\gamma 2 = 1.0$. In b, the mutation parameters $\alpha 1$ and $\alpha 2$ were set to 0.1 and 0.05, respectively, and calculations were made over the ranges $\beta 1 = 0.01$–0.2 and $\gamma 2 = 0.01$–2.0. For each point the ratio is the A + T content when $s = 0$ (constrained twofold degenerate sites) divided by the A + T content of unconstrained sites. Note that the deviation of this ratio from 1 is strongest when the relative transversion rate in the mutation matrix ($\beta 1$) is higher.

the s parameter for the constrained sites. When $\gamma 2 < \gamma 1$, AT_2:AT_{NC} decreases as selection strength increases but when $\gamma 2 > \gamma 1$ the A + T content increases with increasing selection strength. With these parameters, at the extremes of the range shown, fully constrained twofold degenerate sites are ~20% higher or lower in A + T content than are unconstrained sites.

Varying the value of $\beta 1$ while holding $\alpha 1$ and $\alpha 2$ constant does not alter the general shape of the graph in Figure 1a but AT_2:AT_{NC} approaches 1.0 at the extremes as $\beta 1$ decreases so that the graph becomes progressively flatter as $\beta 1$ approaches 0 (data not shown). In general, the degree of deviation of AT_2:AT_{NC} from 1.0 increases as $\beta 1$ increases relative to $\alpha 1$ since in these cases transversions, which do not affect the composition of constrained twofold degenerate sites, will have more influence on the equilibrium composition of neutral DNA. This is demonstrated by Figure 1b.

In real sequences the selective constraints on amino acid composition will vary from site to site, meaning that the expected A + T content at twofold degenerate sites will also vary from site to site, if the mutation dynam-

ics meet the required conditions that $\gamma 1 \neq \gamma 2$. This variation can exist among sites within a gene or as a general property among genes with different overall selective constraints. Therefore, the composition of noncoding sequences cannot be used as an estimation of the expected A + T content of twofold degenerate sites of protein-coding sequences. Nor can twofold degenerate sites with different selective constraints at the amino acid level be expected to evolve to the same A + T content, even when all synonymous changes are selectively neutral.

This last point has particular implications for studies that consider the influence of selection for translation accuracy on codon usage (Akashi 1994; Tautz and Nigro 1998; Labate *et al.* 1999). The approach is to compare the codon usage of conserved amino acid sites with those that vary among the taxa studied, the former assumed to be on average more important for protein function and, as a result, potentially under stronger selective pressure to be translated accurately. The null hypothesis proposed is that the two types of sites have the same expected codon frequencies (Akashi 1994). However, conserved and variable sites, by definition, have different levels of amino acid constraints and as a result could differ substantially in codon usage without selection for translation accuracy or any other selective difference between synonymous codons. Therefore, this test for translation accuracy is inappropriate.

The limitation of the model presented is that it deals only with changes at the third codon position. Despite this limitation, the model provides a good basis for predicting how selection should influence A + T content at twofold degenerate sites as a function of mutation dynamics, primarily the parameters $\gamma 1$ and $\gamma 2$. As long as the degeneracy of most sites changes at a rate much lower than the synonymous substitution rate, then the relationship between mutation dynamics and A + T content should not be significantly affected. This is dealt with in the next section when more complex codon-codon transition models are considered.

Codon-codon transition models: We now consider how selection at the amino acid level could affect composition at neutral fourfold degenerate sites. The approach taken was to develop a simple model of sequence evolution that incorporates only mutation bias and selection between nonsynonymous replacements and then to calculate the equilibrium codon frequencies for the model. The influence of amino acid selection will be assessed by measuring the AT asymmetry, or skew ($[f_T - f_A]/[f_T + f_A]$), produced when the underlying mutation model itself generates no skew. The possibility for skew at fourfold degenerate sites lies in the fact that two different synonymous codons can mutate to different nonsynonymous codons. For example, if we compare nonsynonymous mutations from the glycine codons GGA and GGT, GGA is a single mutation from AGA (Arg), CGA (Arg), TGA (Stop), GAA (Glu), GCA (Ala), and GTA (Val), while GGT is a single mutation from AGT (Ser), CGT (Arg), TGT (Cys), GAT (Asp), GCT (Ala), and GTT (Val). Differences in the probability of fixation of these replacements, as well as differences in the probability of the forward and reverse changes, could lead to significant differences in the equilibrium frequencies of the two codons.

Different 61 × 61 Markov matrices for the sense codons were generated by using various mutation matrices based on **N** (above) and a second matrix called the amino acid acceptance matrix (**A**). In the matrix **A**, $\mathbf{A}_{ij}$ is the probability that a change from amino acid i to amino acid j that is generated by a nucleotide mutation is "accepted" by selection. Although we are using the term acceptance, the value does not actually represent the probability of a nonsynonymous mutation being fixed, but, rather, is the rate of that replacement as a fraction of the rate of neutral evolution. Therefore, an acceptance of 1 would mean that all such replacements are neutral and would occur at the neutral substitution rate. An acceptance value of 0.1 for two amino acids would mean that the rate of replacement across sites from amino acid i to amino acid j is one-tenth the neutral rate. Mutations to stop codons are assumed to be lethal ($\mathbf{A} = 0$) so that the matrix is 61 × 61 instead of 64 × 64. For simplicity we assume that the same matrix applies across all sites.

From the **N** and **A** matrices we can generate the 61 × 61 Markov transition matrix, **C**, for codon transitions. In this matrix the probability of changing from codon x to codon y ($x \neq y$) is given by

$$\mathbf{C}_{x,y} = \mathbf{N}_{ij} \times \mathbf{A}_{nm}. \quad (5)$$

In Equation 5, $\mathbf{N}_{ij}$ is the base mutation required to go from codon x to codon y and $\mathbf{A}_{nm}$ is the probability of accepting a change from amino acid n, the amino acid coded by codon x, to amino acid m, the amino acid coded by codon y. Setting $\mathbf{A}_{ii} = 1$ for all i leads to a model in which there is no selective difference between any synonymous codons. For any two codons that differ by two nucleotide changes $\mathbf{C}_{x,y} = 0$, so only single base mutations are considered, and all diagonals are set to make the rows sum to 1 so that it is a Markov matrix. The equilibrium codon frequencies are determined by calculating ϕ, which is any row of $\mathbf{C}^t$ for large values of t as described above. This vector can be used to calculate skew for any fourfold degenerate codon group by the formula $(f_T - f_A)/(f_T + f_A)$.

The main difficulty is that there is an infinite possible parameter space and the matrices are probably unknowable, particularly since the parameters will vary across time and from site to site. Therefore, the approach taken was not to search the parameter space exhaustively but, rather, to examine a few biologically reasonable examples to determine if the equilibrium codon frequencies are such that $f_T \neq f_A$. The A + T content at twofold degenerate sites was also calculated for every

TABLE 1

Mutation parameters used to generate the codon-codon transition matrices

Matrix	α2	β1	β2	β3	β4	γ1	γ2	Eq. A + T[a] (%)	T_s: T_v
$\mathbf{N}_1$	0.05	0.02	0.001	0.05	0.001	0.25	2.5	69	~3.5:1
$\mathbf{N}_2$	0.1	0.05	0.05	0.001	0.001	0.5	0.02	71	~3:1
$\mathbf{N}_3$	0.1	0.1	0.02	0.001	0.001	0.5	0.01	75	~2.5:1
$\mathbf{N}_4$	0.05	0.05	0.01	0.02	0.005	0.25	0.4	78	~3:1

[a] Equilibrium A + T content of noncoding sequence with this mutation matrix.

parameter set to verify that Equation 4 applies when changes can occur at all three codon positions.

Since this work is ultimately concerned with codon adaptation in plastid genes, the mutation matrices used were chosen such that they give rise to noncoding sequences that have properties consistent with plastid noncoding DNA. These are a high A + T content and a bias toward transitions (Morton 1995). Four different mutation matrices were used ($\mathbf{N}_1$, $\mathbf{N}_2$, $\mathbf{N}_3$, and $\mathbf{N}_4$), which are summarized in Table 1, with $\alpha 1 = 0.2$ in each matrix (only the relative values of **N** are important, not the scale). The equilibrium A + T contents of noncoding sequences with these mutation parameters are within the range observed in plastid DNA and, since they are all based on **N**, each results in symmetrical equilibrium base frequencies. The T_s:T_v ratios for the matrices are all ~3:1, the approximate ratio observed in flowering plant chloroplast DNA (Morton 1995).

Estimating the amino acid acceptances, as defined above, is problematic. Obviously, no matrix exists that could apply to every site in a gene at all times, nor does one that applies to all proteins exist, but it is possible to get a rough estimate of the relative probabilities of fixation of different amino acid changes across all sites. To accomplish this, we utilized one empirically based transition matrix for amino acid changes and one matrix based on a calculation of chemical similarity. The empirical matrix was derived from a study of plastid genomes (Adachi *et al.* 2000, Table 3). Each cell in their table represents the probability of amino acid *i* being replaced by amino acid *j* during a given period of time. To convert these to acceptances as defined above, every cell in the matrix was divided by the maximum observed value and then multiplied by the arbitrary value 0.5. The chemical similarity matrix was taken from Grantham (1974), where the chemical distances were converted to acceptance values by dividing each by the maximum distance plus 1.0 and subtracting one-half this value from 0.5 so that the acceptances range from 0.002 (chemically most distant) to 0.49 (chemically most similar). For both matrices, then, the maximum acceptance was arbitrarily set to ~50%, meaning that the highest rate of amino acid replacement is one-half the rate of neutral changes, a choice that is discussed below.

From the **N** and **A** matrices the full transition matrix was calculated as described above using Equation 5 and then the stationary vector of codon frequencies was determined. All calculations were performed on a Power Macintosh G3 using a Pascal program written by the author. The stationary vector was verified by applying Equation 1 and from ϕ the frequency of each codon was calculated within each synonymous group. The stationary frequencies of the NNT and NNA codons of each fourfold degenerate codon group are given in Table 2. For Leucine, Serine, and Arginine the codon frequencies represent the proportion relative to all six synonymous codons. For each codon group, the skew between A and T at fourfold degenerate sites is also given (only A and T are considered since the sequences are A + T rich). In addition, Table 2 gives the average A + T content of twofold degenerate sites along with the predicted influence of selection on the basis of Equation 4. The average is not weighted by the frequency of occurrence of each synonymous group.

The results in Table 2 demonstrate that mutation bias and selection on amino acid usage are capable of generating skew at fourfold degenerate sites in the absence of selective differences between synonymous codons. Almost every codon group is skewed at equilibrium in the models presented and in some cases this skew is quite strong, >20%. The degree of the skew, as well as the direction ($f_T > f_A$ or $f_A > f_T$), varies quite a bit among codon groups depending on the mutation model and the acceptance matrix but there is skew in each case. Due to the enormous parameter space and the variation across sites and time, demonstrating that the skew observed in any given gene is the result of amino acid selection and mutation bias is probably an impossible task. Despite this, we can draw a more limited conclusion: The general assumption that fourfold degenerate sites of coding sequences should follow Chargaff's second parity rule in the absence of selective constraints on synonymous changes is not valid.

Table 2 also supports the model presented above regarding twofold degenerate sites. For each combination of nucleotide mutation model and amino acid acceptance matrix the prediction is met. When $\gamma 1 < \gamma 2$ ($\mathbf{N}_1$ and $\mathbf{N}_4$) we observe that the equilibrium A + T content

TABLE 2

Equilibrium composition of silent sites under different nucleotide mutation models

		Adachi matrix				Grantham matrix			
	Codon	N_1	N_2	N_3	N_4	N_1	N_2	N_3	N_4
Leu[a]	CTA	0.12	0.18	0.18	0.13	0.15	0.16	0.15	0.14
	CTT	0.15	0.16	0.15	0.14	0.16	0.16	0.15	0.14
	Skew[b]	11.92	−5.46	−7.97	3.18	2.66	−1.41	−1.63	0.56
Ser	TCA	0.27	0.23	0.22	0.27	0.25	0.22	0.23	0.26
	TCT	0.27	0.23	0.21	0.27	0.30	0.20	0.19	0.28
	Skew	0.55	−0.13	−0.21	0.15	9.73	−4.91	−9.09	2.59
Arg	CGA	0.29	0.044	0.018	0.17	0.26	0.11	0.07	0.17
	CGT	0.29	0.064	0.033	0.17	0.19	0.14	0.12	0.15
	Skew	−0.23	18.01	29.58	−0.64	−16.13	12.25	26.76	−4.85
Pro	CCA	0.35	0.36	0.37	0.39	0.35	0.36	0.39	0.39
	CCT	0.34	0.36	0.37	0.39	0.34	0.35	0.36	0.39
	Skew	−1.56	−0.14	0.14	0.01	−2.32	−1.67	−4.14	0.24
Thr	ACA	0.33	0.39	0.42	0.38	0.34	0.36	0.39	0.39
	ACT	0.36	0.32	0.33	0.40	0.36	0.35	0.36	0.40
	Skew	4.55	−9.26	−12.44	2.13	3.64	−1.87	−3.22	0.64
Val	GTA	0.32	0.43	0.48	0.37	0.34	0.36	0.38	0.39
	GTT	0.36	0.29	0.28	0.41	0.35	0.35	0.36	0.39
	Skew	5.70	−20.72	−26.10	4.26	0.99	−1.24	−1.73	0.28
Ala	GCA	0.34	0.38	0.40	0.39	0.33	0.36	0.39	0.39
	GCT	0.35	0.34	0.35	0.39	0.36	0.35	0.37	0.39
	Skew	0.44	−5.70	−7.38	0.81	3.45	−0.79	−1.29	0.49
Gly	GGA	0.34	0.36	0.38	0.39	0.33	0.35	0.38	0.39
	GGT	0.35	0.35	0.37	0.39	0.35	0.35	0.36	0.39
	Skew	1.55	−1.21	−1.13	0.37	1.94	0.16	−1.58	0.55
Twofold deg. (%)	Avg.[c]	78.4	67.3	67.9	79.7	72.7	69.5	71.8	78.8
	NC[d]	69	71	75	78	69	71	75	78
	Pred.[e]	Increased	Decreased	Decreased	Increased	Increased	Decreased	Decreased	Increased

[a] The frequency of each codon is given relative to all synonymous codons of that group. Only the NNA and NNT codons are shown.
[b] Measured by $(f_T - f_A)/(f_T + f_A) \times 100\%$.
[c] Unweighted average A + T content of twofold degenerate codon groups.
[d] Equilibrium A + T content of noncoding (NC) DNA (see Table 1).
[e] Predicted relationship between the A + T content of twofold degenerate sites relative to the A + T content of noncoding DNA based on Equation 4 (text).

TABLE 3

Plastid genomes used in this analysis

Species	Group	Accession no.	Noncoding DNA[a]
Cyanophora paradoxa	Chlorophyta	U30821	23,207
Chlorella vulgaris	Chlorophyta	AB001684	54,952
Odontella sinensis	Chromophyta	Z67753	16,682
Porphyra purpurea	Rhodophyta	U38804	24,301
Marchantia polymorpha	Pinophyta	X04465	20,499
Pinus thunbergii	Gymnosperm	D17510	30,697
Epifagus virginiana	Anthophyta	M81884	41,607
Nicotiana tabacum	Anthophyta	Z00044	43,909
Oryza sativum	Anthophyta	X15901	36,123

[a] Number of nucleotides of noncoding DNA extracted from each genome.

of twofold degenerate sites is greater than the A + T content of noncoding DNA while when $\gamma 2 < \gamma 1$ ($\mathbf{N}_2$ and $\mathbf{N}_3$) the equilibrium A + T content of twofold degenerate sites is lower than the content of noncoding DNA.

The results given are robust with respect to the choice of 50% as the maximum acceptance for amino acid changes. Although the choice is necessarily arbitrary, variation of this maximum value from 20 to 100% of the neutral rate had no effect on the general conclusion, but as the maximum value approaches zero (no amino acid changes), codon frequencies of fourfold degenerate groups equilibrate to the same frequency as noncoding DNA (data not shown). This is to be expected since, when nonsynonymous substitutions are extremely rare, the "gain" and "loss" of the codons of fourfold degenerate groups occurs essentially only through synonymous changes. Therefore, skew should be produced only when amino acid selection is not too strong. The fact that selection intensity on amino acid changes can influence codon usage at fourfold degenerate sites again draws into question the comparison of codon usage at conserved and variable amino acid sites. Skew was also generated using the PAM250 matrix (Dayhoff *et al.* 1978) as the basis for the acceptance matrix (data not shown).

Conclusions: Although the model employed is necessarily a simplification of the substitution process, the factors ignored should not affect the overall conclusion that selection at the amino acid level can influence synonymous codon usage. The model does not account for variation in mutation dynamics as a function of flanking base composition, which is known to occur in chloroplast DNA (Morton 1995), nor does it deal with variation in selection on amino acid usage across sites and time, which certainly exists in any protein-coding sequence. It is not clear how such variations would affect equilibrium composition but it is unlikely that they would completely eliminate the influence observed. It may be possible to pursue this in future work. Despite these difficulties, though, the negative examples are sufficient to demonstrate that selection at the amino acid level cannot simply be ignored. Given this, we must conclude that the previous approach to measuring the influence of selection on plastid gene codon usage (Morton 1998) was not appropriate. In addition, the model makes it clear that analyses of composition data from silent sites in protein-coding regions are not as simple as commonly assumed.

DETECTING CODON ADAPTATION IN PLASTID GENES

The work presented above demonstrates that selection at the amino acid level can potentially influence codon usage. Here we look at evidence that our previous analysis (Morton 1998) was affected by such selection and then explore an alternative test for codon adaptation in plastid genes. The genes to be analyzed here are all protein-coding sequences >350 nucleotides in length from each of the complete genome sequences listed in Table 3, which are the genomes analyzed previously (Morton 1998, 2000). The issue we want to address is whether or not selection is acting on any individual gene to specifically increase major codon usage, that is, if the gene displays codon adaptation. For each gene the codon adaptation index (CAI), the most common measure of the degree to which a gene uses major codons, was measured using the method of Sharp and Li (1987) and the codon fitness assignments described previously (Morton 1998). In addition, from each genome, all noncoding regions >30 nucleotides in length were extracted to use in the calculations. The total amount of this noncoding DNA is given in Table 3 for each genome.

Amino acid selection and the previous resampling test: Composition data from these genes strongly support the possibility that the previous test for codon adaptation was inappropriate. This is seen most clearly in those plastid genes with a relatively low frequency of major codons, defined here as genes with CAI values <0.4, which are the majority of plastid genes. The cumulative codon usage of these genes deviates from what

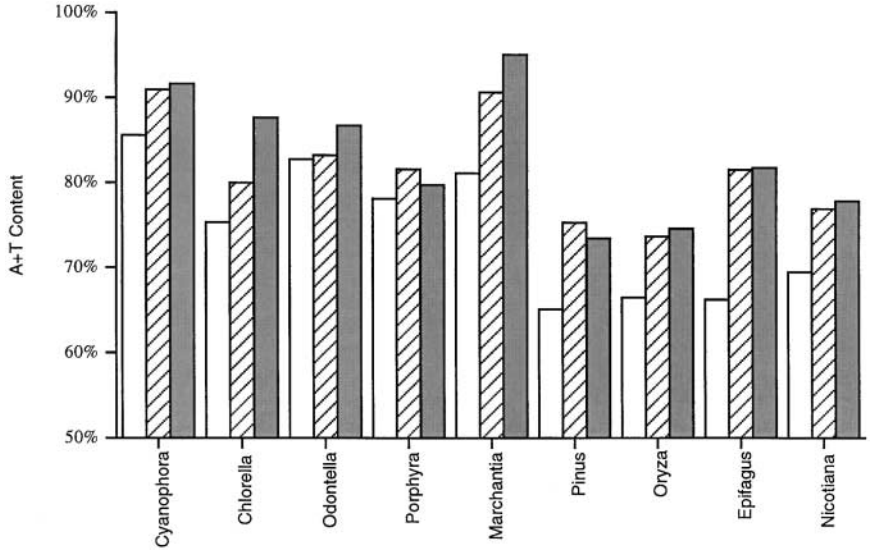

FIGURE 2.—Expected and observed A + T content of twofold degenerate sites for plastid genes with CAI values <0.4. The expected A + T content of twofold degenerate sites for each species (open bars) is based on the dinucleotide frequencies of noncoding sequences. For every amino acid the expected number of each codon is based on the total number of occurrences of the amino acid multiplied by the frequency of the third position nucleotide in noncoding DNA, conditional on the composition of the 5′ base. For example, the expected number of TTT codons (which, along with TTC, codes phenylalanine) is the total number of phenylalanine residues multiplied by the proportion T/(T + C) in noncoding DNA when the 5′ flanking base is a T. The overall expected frequency is based on the sum of these values across all amino acids. The observed A + T content at twofold degenerate sites of protein-coding sequences is shown for sites with either a pyrimidine (hatched bars) or a purine (shaded bars). In every species, the twofold degenerate sites have a higher A + T content than does noncoding DNA.

would be observed if silent site composition matched the composition of noncoding DNA, but the deviation is not due to an increase in major codon usage in every codon group. Instead, the deviation is similar to what could result from amino acid selection as seen above; in every species twofold degenerate sites from genes with CAI < 0.4 show a higher A + T content than expected (Figure 2) while fourfold degenerate sites show strong skew, even though noncoding DNA is asymmetric (Figure 3). These deviations cannot be due to selection for major codons because, while in some codon groups major codons are used more frequently than predicted from noncoding DNA, in other codon groups the deviation is due to a decreased frequency of major codons. Specifically, in the codon groups CAY, GAY, TTY, TAY, and AAY, selection for translation efficiency favors the codon with a C at the third position (MORTON 1993, 1998) but the genes with CAI values <0.4 show a higher than expected frequency of T at the third position (Figure 2). This would not be the case if there were selection for translation efficiency since a high frequency of C at the third position of these codon groups is the most notable feature of plastid genes with strong codon adaptation (MORTON 1993, 1998) and selection should not act to decrease the frequency of major codons.

Given the composition data, it is probable that the resampling test was misleading with respect to codon adaptation. This can be demonstrated by repeating the resampling procedure based on noncoding dinucleotide frequencies as described (MORTON 1998) and comparing observed and expected results without regard to

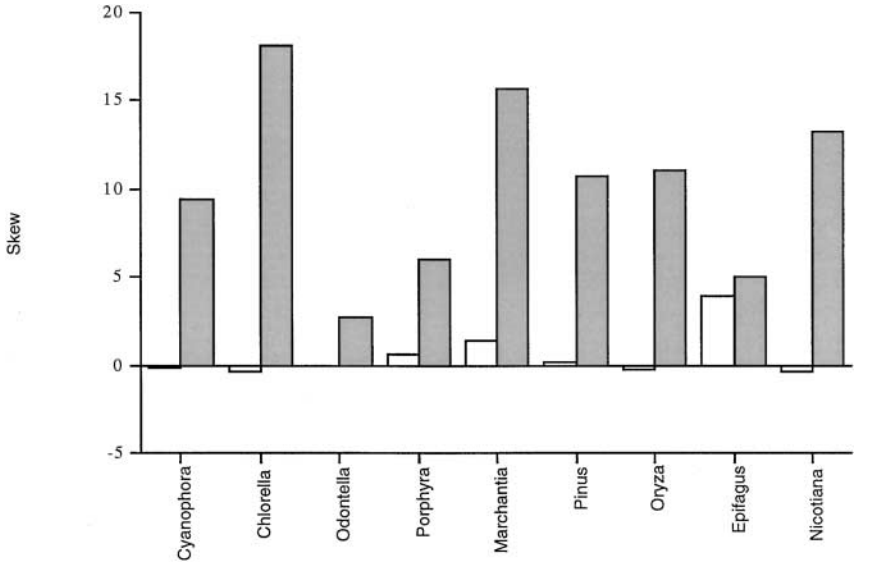

FIGURE 3.—AT skew, calculated by $(f_T - f_A)/(f_T + f_A) \times 100\%$, in noncoding sequences (open bars) and at fourfold degenerate sites of protein-coding sequences with CAI values <0.4 (shaded bars).

TABLE 4

Number of genes with a codon adaptation index value greater than expected given noncoding DNA dinucleotide frequencies

Organism	Total genes	No. of genes with CAI greater than expected[a]	%
Cyanophora	81	81	100
Chlorella	53	49	92.4
Odontella	72	68	94.4
Porphyra	116	103	88.8
Marchantia	33	33	100
Pinus	34	31	91.2
Epifagus	19	17	89.5
Nicotiana	31	27	87.1
Oryza	40	36	90.0

[a] The expected CAI is the mean CAI of the 500 randomly generated codon tables. The results show the number of genes with an observed CAI greater than expected without respect to statistical significance.

statistical significance. When the observed CAI value of each gene is compared to the mean CAI of the random codon tables (the expected CAI), we find a positive deviation for almost every gene (Table 4), meaning that the overall frequency of major codons is higher than expected in almost every plastid gene. This deviation cannot be explained solely by codon adaptation in every gene as noted already and, further, the deviation in many of the genes (Figures 2 and 3) could potentially be generated by selection at the amino acid level. Therefore, even though we cannot demonstrate conclusively that selection at the amino acid level is responsible, given the complexity of the parameters involved, the evolutionary models presented above and the composition data presented here are enough to draw this approach into question. Even in the case of a gene with a CAI value significantly higher than expected, conclusions cannot be drawn concerning codon adaptation or even selective differences between synonymous codons.

It should be noted that the results in Table 4 and Figures 2 and 3 do not appear to be a result of different mutation dynamics in transcribed and nontranscribed DNA. When the calculations are repeated using only noncoding DNA within 30 nucleotides of a start or stop codon and are therefore assumed to be transcribed noncoding DNA, or with intron sequence data where available, the results are not affected (data not shown).

Recursive resampling approach: The new approach begins with the assumption, based on the argument just presented, that the general codon usage features observed in plastid genes with low CAI values (Figures 2 and 3) are not due to selection favoring major codons. Given this assumption we can use a recursive resampling approach to determine which genes are significantly different from the "cumulative" in their major codon usage. The null hypothesis is that all genes that lack codon adaptation show the same pattern of codon usage bias, but in this case we do not require it to match noncoding DNA. A corollary of this hypothesis is that resampling from the cumulative codon pool of genes lacking codon adaptation can explain the codon usage of any individual gene without codon adaptation. Therefore, if a gene has a significantly higher frequency of major codons than predicted from this resampling we can reject the hypothesis and infer that selection for codon adaptation acts on this gene.

To start, all of the protein-coding genes of a genome were combined to generate a cumulative codon pool for that species. For every gene, 500 replicate copies

TABLE 5

Genes from each species with a significantly large CAI value as determined by the recursive resampling test

Species	Genes	No. rejected[a]	Genes rejected	Significant CAI values: noncoding[b] (%)
Cyanophora	81	22 (27)	*apcA apcB apcD apcE apcF atpA atpB cpcA cpcB petA petB petD psaA psaB psaF psbA psbB psbC psbD rbcL rbcS tufA*	75
Chlorella	53	15 (28)	*atpA atpB atpI petB petD psaA psaB psbA psbB psbC psbD rbcL rps3 rps12 tufA*	63[c]
Odontella	72	11 (15)	*atpA atpB psaA psaB psbA psbB psbC psbD rbcL rbcS tufA*	48
Porphyra	116	9 (8)	*apcA atpB cpcA cpcB cpcC petB psbA rbcL tufA*	37
Marchantia	33	5 (15)	*atpA psaA psbA psbC rbcL*	50
Pinus	34	1 (3)	*psbA*	24
Epifagus	19	0	*None*	—
Nicotiana	31	2 (7)	*psbA rbcL*	11
Oryza	40	2 (5)	*psbA rbcL*	8

[a] The number of genes for which the null hypothesis was rejected (see text). Percentages are in parentheses.
[b] Results from Morton (1998) using noncoding DNA to estimate an expected level of codon adaptation. Epifagus was not included in that study.
[c] Results from Morton (2000) using noncoding DNA to estimate an expected level of codon adaptation.

with the same amino acid usage as the gene itself were then generated randomly by drawing, with replacement, from the combined codon pool. The average (expected) CAI of the random set was then calculated along with the standard deviation. For a gene with an observed CAI more than three standard deviations greater than expected, the null hypothesis was rejected and the gene was then set aside. Following the first run, all genes that had been set aside were taken to have significant codon adaptation and excluded from further analysis. The remaining genes were then combined and the process was repeated so that the codons from those genes rejected in the previous step were not used in the resampling process. These steps were continued until a run was complete without any genes found to have a significantly greater CAI than expected.

The results of this test are given in Table 5, which indicates the genes from each species for which the null hypothesis is rejected. The proportion of genes found to have significant codon adaptation in the previous study is also given in Table 5 for comparison. In every species, a minority of genes was found to have a significantly higher frequency of major codons than expected from the cumulative codon usage. Essentially every gene for which the null hypothesis is rejected is known from studies of protein translation levels to be highly expressed (see Morton 1998). These are primarily the core photosystem I (designated by psa) and photosystem II (psb) genes, major subunits of the ATPase (atp), and, in those species that code them, the phycobilisomes (cpc) and allophycocyanins (apc). Of particular interest is that the null hypothesis is rejected for *psbA*, the most highly translated plastid gene, in every case. Overall, the number of genes inferred to have codon adaptation is much lower using the current approach, particularly in the algae and the liverwort Marchantia. In addition, the parasitic Epifagus (not included in the previous study), which has lost all photosynthesis genes, including *psbA*, from its genome (Wolfe *et al.* 1992), has no gene that is significantly different from the cumulative pool. Using a less stringent approach of setting aside genes that were two standard deviations or more above the mean gave very similar results.

As an estimate of codon adaptation, the current approach is likely to be more accurate than the previous test and may prove promising in other genomes. However, although this approach allows us to infer which plastid genes show evidence for selection on translation efficiency, the null hypothesis assumes homogeneous codon usage by those genes lacking codon adaptation. It is not clear how much variation among genes could exist in the absence of selection for translation efficiency, but the potential does exist that such variation could render the null hypothesis invalid. This can be investigated in future studies on factors other than codon adaptation that influence codon usage. In the case of plastid genes, though, the fact that a small number of highly expressed genes show statistically significant CAI values relative to the cumulative codon pool is strong evidence that selection acts on these genes specifically to increase major codon usage.

I thank two anonymous reviewers for their extremely helpful comments. This work was supported in part by grant MCB-9727906 from the National Science Foundation.

LITERATURE CITED

Adachi, J., P. J. Waddell, W. Martin and M. Hasegawa, 2000 Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. **50:** 348–358.

Akashi, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics **136:** 927–935.

Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076.

Akashi, H., 1999 Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene **238:** 39–51.

Akashi, H., and A. Eyre-Walker, 1998 Translational selection and molecular evolution. Curr. Opin. Genet. Dev. **8:** 688–693.

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

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

Carulli, J. P, D. E. Krane, D. L. Hartl and H. Ochman, 1993 Compositional heterogeneity and patterns of molecular evolution in the Drosophila genome. Genetics **134:** 837–845.

Chargraff, E., 1979 How genetics got a chemical education. Ann. NY Acad. Sci. **325:** 345–360.

Cox, D. R., and H. D. Miller, 1965 *The Theory of Stochastic Processes.* Chapman & Hall, New York.

Dayhoff, M. O., R. Schwartz and B. C. Orcutt, 1978 Suppl. 3 in *Atlas of Protein Sequence and Structure*, Vol. 5, edited by M. O. Dayhoff. National Biomedical Research Foundation, Washington, D.C.

Duret, L., and D. Mouchiroud, 1999 Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. USA **96:** 4482–4487.

Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11:** 725–736.

Grantham, R., 1974 Amino acid difference formula to help explain protein evolution. Science **185:** 8662–8864.

Gu, X., and W.-H. Li, 1998 Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. Proc. Natl. Acad. Sci. USA **95:** 5899–5905.

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

Kliman, R. M., and J. Hey, 1994 The effects of mutation and natural selection on codon bias in the genes of Drosophila. Genetics **137:** 1049–1056.

Labate, J. A., C. H. Biermann and W. F. Eanes, 1999 Nucleotide variation at the *runt* locus in *Drosophila melanogaster* and *Drosophila simulans.* Mol. Biol. Evol. **16:** 724–731.

Lewontin, R. C., 1989 Inferring the number of evolutionary events from DNA coding sequence differences. Mol. Biol. Evol. **6:** 15–32.

Li, W.-H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. **24:** 337–344.

Lio, P., and N. Goldman, 1998 Models of molecular evolution and phylogeny. Genome Res. **8:** 1233–1244.

Macdonald, J. H., and M. Kreitman, 1991 Adaptive evolution at the Adh locus in Drosophila. Nature **351:** 652–654.

Morton, B. R., 1993 Chloroplast DNA codon use: evidence for selection at the *psbA* locus based on tRNA availability. J. Mol. Evol. **37:** 273–280.

Morton, B. R., 1995 Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions. Proc. Natl. Acad. Sci. USA **92:** 9717–9721.

Morton, B. R., 1998 Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J. Mol. Evol. **46:** 449–459.

Morton, B. R., 2000 Codon bias and the context dependency of nucleotide substitutions in the evolution of plastid DNA. Evol. Biol. **31:** 55–103.

Muse, S. V., and B. S. Gaut, 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to chloroplast genome. Mol. Biol. Evol. **11:** 715–724.

Percudani, R., and S. Ottonello, 1999 Selection at the wobble position of codons read by the same tRNA in Saccharomyces cerevisiae. Mol. Biol. Evol. **16:** 1752–1762.

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

Sueoka, N., 1992 Directional mutation pressure, selective constraints, and genetic equilibria. J. Mol. Evol. **34:** 95–114.

Sueoka, N., 1999 Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A=T and G=C. J. Mol. Evol. **49:** 49–62.

Tautz, D., and L. Nigro, 1998 Microevolutionary divergence pattern of the segmentation gene *hunchback* in Drosophila. Mol. Biol. Evol. **15:** 1403–1411.

Wolfe, K. H., C. W. Morden, S. C. Ems and J. D. Palmer, 1992 Rapid evolution of the plastid translational apparatus in a nonphotosynthetic plant: loss or accelerated sequence evolution of tRNA and ribosomal protein genes. J. Mol. Evol. **35:** 304–317.

Yang, Z., and R. Nielsen, 2000 Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17:** 32–43.

Yang, Z., R. Nielsen and M. Hasegawa, 1998 Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. **15:** 1600–1611.

Communicating editor: G. A. Churchill