J Mol Evol (2003) 56:616–629
DOI: 10.1007/s00239-002-2430-1

JOURNAL OF MOLECULAR EVOLUTION

The Role of Context-Dependent Mutations in Generating Compositional and Codon Usage Bias in Grass Chloroplast DNA

Brian R. Morton

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

Received: 23 August 2002 / Accepted: 17 December 2002

Abstract. The influence of local base composition on mutations in chloroplast DNA (cpDNA) is studied in detail and the resulting, empirically derived, mutation dynamics are used to analyze both base composition and codon usage bias. A 4×4 substitution matrix is generated for each of the 16 possible flanking base combinations (contexts) using 17,253 noncoding sites, 1309 of which are variable, from an alignment of three complete grass chloroplast genome sequences. It is shown that substitution bias at these sites is correlated with flanking base composition and that the A + T content of these flanking sites as well as the number of flanking pyrimidines on the same strand appears to have general influences on substitution properties. The context-dependent equilibrium base frequencies predicted from these matrices are then applied to two analyses. The first examines whether or not context dependency of mutations is sufficient to generate average compositional differences between noncoding cpDNA and silent sites of coding sequences. It is found that these two classes of sites exist, on average, in very different contexts and that the observed mutation dynamics are expected to generate significant differences in overall composition bias that are similar to the differences observed in cpDNA. Context dependency, however, cannot account for all of the observed differences: although silent sites in coding regions appear to be at the equilibrium predicted, noncoding cpDNA has a significantly lower A + T content than expected from its own substitution dynamics, possibly due to the influence of indels. The second study examines the codon usage of low-expression chloroplast genes. When context is accounted for, codon usage is very similar to what is predicted by the substitution dynamics of noncoding cpDNA. However, certain codon groups show significant deviation when followed by a purine in a manner suggesting some form of weak selection other than translation efficiency. Overall, the findings indicate that a full understanding of mutational dynamics is critical to understanding the role selection plays in generating composition bias and sequence structure.

Key words: Transition — Transversion — Mutation bias — Codon usage bias — Chloroplast genome

Introduction

Understanding the degree to which codon usage has been shaped by selection, as opposed to mutation bias, has been the focus of much research (Antezana and Kreitman 1999; Gerber et al. 2001; Smith and Eyre-Walker 2001, 2002; Fay et al. 2002). There is now good evidence that highly expressed genes of certain genomes, particularly in unicellular organisms, are biased toward a set of "major" codons that are complementary to abundant tRNAs and that this bias is the result of selection to increase translation efficiency (Ikemura 1985; Li 1987; Andersson and Kurland 1990; Bulmer 1991; Sharp 1991; Akashi and

email: bmorton@barnard.edu

Eyre-Walker 1998). However, in the case of the weakly expressed genes from these same genomes and the genes from other genomes such as mammalian genomes, the role of selection remains unresolved. These genes show no clear evidence of adaptation for translation efficiency but there is evidence that some other selective pressure(s) may influence their codon usage (Antezana and Kreitman 1999; Llopart and Aguade 2000; Smith and Eyre-Walker 2001; Konu and Li 2002).

One difficulty with resolving the role of selection is that we lack a full understanding of mutation dynamics and, thus, of what sort of composition bias or codon usage to expect in the absence of selection. Such an understanding, though, is critical since studies of selection are frequently based on a comparative analysis of composition biases, particularly when a lack of polymorphism data precludes the application of some variation of the MacDonald–Kreitman (1991) test. For example, differences in composition between silent sites of genes and surrounding non-coding DNA have been used to infer selective constraints on codon usage (Carulli et al. 1993; Kliman and Hey 1994; Morton 1998; Percudani and Ottonello 1999) and comparisons of the codon usage at conserved and variable amino acid sites have been used to infer selective constraints for translation accuracy (Akashi 1994; Tautz and Nigro 1998; Labate et al. 1999). It has also been argued that deviations from Parity Rule 2 (PR2), which states that G = C and A = T, assuming that the two DNA strands have the same mutational biases, can be used to assess selection (Sueoka 1999, Sueoka and Kawanishi 2000).

These comparative analyses are based on an implicit assumption about mutation dynamics, which is that selection at the amino acid level has no influence on silent substitutions. This allows us to interpret significant differences in composition bias between noncoding DNA and silent sites of coding regions as evidence for selective constraints on the latter. However, if mutation dynamics are sufficiently complex, selection at the amino acid level can have an indirect influence on the composition bias of silent sites in coding regions (Morton 2001). Since selection on amino acid usage of a coding region constrains the composition of the first and second codon positions, third position silent sites occur within a limited, and biased, set of contexts. If mutations are context dependent (e.g., Blake et al. 1992; Hess et al. 1994; Morton 1995; Berg and Silva 1997), then direct comparisons of coding and noncoding DNA may be confounded by the fact that there are very different average contextual influences in the two sequences. Thus, the expectation that compositional properties will be similar across neutral sites, either within coding sequences or across both coding and noncoding DNA, would be unwarranted. Therefore, it is important that we develop a better understanding of complex mutational dynamics and how they may influence our analyses of sequence composition.

In this study we examine the importance of considering context dependency in analyzing codon usage in the flowering plant chloroplast genome (cpDNA). This genome is an excellent model for such a study. Although the highly expressed *psbA* and *rbcL* genes have a bias toward codons that are complementary to the tRNA population (Morton 1993, 1998), the remaining genes (referred to here as low-expression) have a distinctly different pattern of codon usage bias (Morton 2001). However, despite the fact that the codon usage of these genes shows no evidence for adaptation to the tRNA population, there is still some question about whether or not their codon usage has been influenced by some other form of selection. This is because there is a significant difference in composition between the silent sites of these genes and noncoding cpDNA from the same genome. The differences can be summarized as two general features. The first is that silent sites consistently have a higher A + T content than noncoding cpDNA and the second is that fourfold-degenerate sites show a strong skew toward pyrimidines (i.e., Y [pyrimidine] > R [purine]), a skew that is not observed in noncoding cpDNA. These differences are not limited to flowering plants but are consistent across all sequenced plastid genomes (Morton 2001).

Another reason for using the flowering plant chloroplast genome as a model is the fact that there is strong evidence that this genome's mutational dynamics are context dependent. Previous studies of noncoding cpDNA have shown that the degree of bias toward transitions is a function of the A + T content of flanking nucleotides (Morton 1995, 1997, 2000; Morton and Clegg 1995; Yang et al. 2002). The correlation between context and substitution bias is observed both within short noncoding regions (Morton and Clegg 1995; Yang et al. 2002) and across the entire genome (Morton 1995), indicating that it is a direct result of flanking nucleotides having an influence on mutation bias.

In the first part of this paper, substitutions in cpDNA are studied in depth to gain a better understanding of how they depend upon context. In the second part, the effect of context dependency on the molecular evolution of cpDNA is studied. This involves an application of the observed mutational dynamics toward two major goals: (1) to determine if differences in average context between silent sites of chloroplast coding sequences and noncoding cpDNA, coupled with a context-dependent mutation process, can generate the average compositional differences noted above and (2) to assess more accurately the role selection has played in gen-

erating the codon usage of low-expression chloroplast genes.

On the basis of the differences observed in noncoding cpDNA from three complete grass chloroplast genomes, a 4×4 substitution matrix is generated for each possible context, defined as the composition of the two flanking nucleotides. Each matrix is used to generate a Markov transition matrix for which the stationary vector is calculated to infer the equilibrium composition biases of sequences evolving under such a substitution scheme and to generate an expectation of context-dependent codon bias in the absence of selective pressure. These expected base composition and codon usage biases are then compared to observed biases in cpDNA.

The results show that the substitution dynamics of cpDNA is strongly context dependent and give rise to variation in compositional properties, particularly A + T content and purine/pyrimidine skew, across contexts. In the comparison of silent sites from coding and noncoding cpDNA, it is found that the difference in A + T content, noted above, is not due to a deviation of silent coding sites from expected composition bias, but rather, results from noncoding cpDNA having a significantly lower A + T content than expected. It is suggested that the contribution of indels to the evolution of noncoding cpDNA is responsible for this deviation. In addition, the skew toward pyrimidines that is observed in coding sequences appears to be generated by the context dependency of mutations occurring at sites that are biased toward pyrimidine-rich contexts as a result of the structure of the genetic code and the average amino acid composition of chloroplast proteins. The latter finding draws into question the use of Parity Rule 2 (Sueoka 1999) as an analytical tool for selection.

When context effects are accounted for in an analysis of codon usage in low-expression genes, it is found that sites followed by a pyrimidine match the predicted codon usage. In contrast, several codon groups deviate strongly from expectation when followed by a purine. The pattern seems to indicate some form of weak selection on codon usage, possibly related to mRNA structure. Overall, the findings indicate that mutational dynamics can be much more complex than commonly assumed and may lead to compositional patterns that are rarely considered but that could complicate inferences about selection on synonymous sites.

Materials and Methods

The complete chloroplast genome sequences from *Zea mays* (NC_001666), *Triticum aestivum* (NC_002762), and *Oryza sativa* (NC_001320) were obtained from GenBank and protein-coding sequences extracted using the information in each file. A total of 31 genes, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psaA*, *psaB*, *petA*, *petB*, *petD*, *atpA*, *atpB*, *atpF*, *atpH*, *atpI*, *ndhD*, *ndhE*, *ndhF*, *rpoA*, *rps2*, *rps4*, *rps7*, *rps8*, *rps11*, *rps14*, *rps19*, *rpl2*, *rpl14*, *rpl16*, *rpl23*, and *rpl33* was combined for codon usage analyses. These represent most of the protein-coding regions of the grass genome. The remaining genes were excluded since they are affected by significant indel mutations that could bias the analysis. Each of these genes has a codon usage that is not statistically different from that predicted from the cumulative codon pool as tested by resampling (Morton 2001). The genes *psbA* and *rbcL* were excluded since they are by far the most prominent translation products in the chloroplast and there is strong evidence that they are under selection to adapt their codon usage to the tRNA content (Morton 2001).

All noncoding sequences that had the same two flanking genes in each of the three genomes were extracted and aligned. The regions analyzed were those that met the criteria outlined previously to decrease the likelihood that nonhomologous regions were being compared (Morton 1995). A total of 77 noncoding regions was analyzed with 17,253 aligned sites, 1309 of which (7.6%) were variable. Given this level of sequence divergence, it was assumed that multiple hits at a site would not have a significant effect. The strand given in the GenBank file is referred to as the plus strand. The total base composition of these regions was calculated from the *Zea mays* sequence. A list of the noncoding regions, defined by flanking genes, is available upon request.

Substitution Matrices. Substitution matrices for the plus strand of noncoding cpDNA were generated using *Oryza* as the outgroup in the three taxa tree (Duvall and Morton 1996). The structure of the matrices is such that each row represents the original nucleotide and each column represents the extant nucleotide. For each site, the original nucleotide was taken as the inferred state in the *Zea–Triticum* ancestor with *Oryza* as the outgroup. Nucleotide state change was then inferred along both branches that diverge from the *Zea–Triticum* ancestor, one leading to *Zea* and the other to *Triticum*. As a result, those sites with the nucleotide N observed in all three sequences were scored as two entries into row N, column N, one entry for each branch along which the nucleotide state was conserved. Variable sites, where a change from nucleotide N to nucleotide N_1 was inferred along one of the two branches, were scored as one entry into row N, column N_1, for the substitution, and one entry into row N, column N, for the branch along which the nucleotide was conserved. The two entries for each site were then added to the matrix corresponding to the context observed at that site, where context is defined as the composition of the 5′ and 3′ flanking bases on the plus strand. Contexts are indicated as [5′ base_3′ base], where the underscore indicates the site of the mutation, such that T_C indicates those sites flanked by a 5′ T and a 3′ C. Only sites within a conserved context were scored. Overall, a separate matrix was developed for each of the 16 possible contexts, and these matrices summarize the substitution dynamics on the plus strand. These 16 matrices are referred to as the specific context matrices.

Those specific context matrices that share generic context features were combined to examine any general trends in the variation in substitution bias and to compare the results with the observations of earlier studies. These two generic features are A + T content (A + T context) and the arrangement of pyrimidines (Y context), both of which have been shown to be associated with variation in the ratio of transitions to transversions in cpDNA (Morton 1995; Morton et al. 1997; Yang et al. 2002). The A + T context is the number of flanking base pairs that are AT. There are three possibilities for this context: A + T = 0 (no flanking AT base pairs), A + T = 1 (one flanking AT base pair), and A + T = 2 (each flanking site is an AT base pair). The pyrimidine (Y) context is defined as the number of flanking pyrimidines on the plus strand, which can be zero, one, or two. To examine the effect of these two features separately, Y context is analyzed separately within each

A + T context. For example, the matrices from the specific contexts A_A, A_T, T_A, and T_T were combined to make up the A + T = 2 generic context. Within this, A_A represents the Y = 0 context, T_T represents the Y = 2 context, and the remaining two make up the Y = 1 context. In general, these combined matrices are referred to as generic context matrices.

Markov Transition Matrices and Sequence Evolution. For each substitution matrix, whether for a specific or generic context, a Markov transition matrix was generated by dividing each element in the matrix by the sum of the element's row. For a Markov matrix Π the stationary vector ϕ was calculated by generating Π^t for large t, each row of which will equal ϕ (Cox and Miller 1965), and then verifying ϕ by making sure that it meets the definition $\phi = \phi\Pi$. Since ϕ gives the equilibrium base frequencies under that substitution regime, we can calculate equilibrium A + T. In addition, we can calculate Y–R skew, defined by Eq. (1).

$$100\% * ((T + C) - (G + A))/(G + A + T + C) \quad (1)$$

For every Π we can also determine the GC → AT pressure, calculated as the total GC → AT rate divided by the total AT → GC rate, for both transitions and transversions ($\gamma 1$ and $\gamma 2$, respectively, as defined by Morton [2001]). These parameters are informative about the relative A + T content of noncoding DNA (i.e., fourfold-degenerate sites) and twofold-degenerate sites in a coding sequence. Mutation bias alone will result in an A + T content at neutral twofold-degenerate sites in protein-coding sequences that is greater than the A + T content of noncoding sequences if the GC → AT pressure is stronger for transitions than it is for transversions (i.e., $\gamma 1 > \gamma 2$ [Morton 2001]). This inequality can contribute to an increased A + T content at silent sites within coding sequences relative to noncoding DNA, as observed in cpDNA (see above; Morton 2001), without any selective pressure on silent changes.

Strand Symmetry Tests. The equivalence of substitution parameters for the two DNA strands was tested in two ways. The first test, the overall strand symmetry test, involves a comparison of matrices with complementary contexts, such as C_C and G_G, since a site in the C_C context on the plus strand is in the G_G context on the minus strand, and vice versa. If substitutions are occurring in the same manner on both strands, then the substitution bias measured for the plus strand in the C_C context would be the same as the substitution bias observed if we examined changes along the minus strand in the same context. Since the latter is the same as the complements of the changes that occur on the plus strand in the G_G context (for example, A → C changes on the plus strand in the G_G context also represent T → G changes in the C_C context on the minus strand), we can test for overall equivalence of substitution pattern by comparing the complementary matrices. The null hypothesis, applied to complementary matrices, is that complementary changes, such as the parenthetical example above, occur at the same rate. A rejection of the null hypothesis would indicate that substitution dynamics of the two strands are not equivalent, meaning that the substitution matrices derived from the plus strand cannot be applied to cpDNA in general. The following comparisons were made: (a) the C_C matrix to the complement of the G_G matrix, (b) C_A to the complement of T_G, (c) C_T to the complement of A_G, (d) T_C to the complement of G_A, (e) G-T to the complement of A_C, and (f) T_T to the complement of A_A. Each comparison involved comparing the two complementary rows using a χ^2 test and the null hypothesis was rejected if any single comparison was significant. The matrices for the contexts C_G, G_C, A_T, and T_A are self-complementary so the symmetry comparison discussed next also tests this null hypothesis.

The second test, referred to here as the context symmetry test, is concerned with the skew generated by substitution dynamics at specific sites instead of overall equivalence of substitution dynamics. Consider all sites that exist in a specific context, such as G_G, on the plus strand. We want to know whether or not at these sites, the two strands, which exist in a C_C context on the minus strand, have the same substitution bias. If they do, then we would observe no skew, that is, we would observe A = T and G = C, at these sites at equilibrium. If, instead, there is significant inequality in substitution bias between the two strands at these sites, due to the fact that they can be in different contexts, then we will observe skew at equilibrium. For example, if G → A changes are occurring at a significantly higher rate than C → T changes in the G_G matrix, then the two strands have different substitution biases at these sites. As a result, these sites will evolve to have a higher frequency of A than T, and we will observe a higher frequency of GAG trinucleotides than GTG trinucleotides on the plus strand. Note that if the two strands have the same overall substitution parameters, tested by the overall strand symmetry test above, then the same skew will be observed on the minus strand, meaning that the plus strand will also be biased toward CTC over CAC trinucleotides. Therefore, this test addresses whether or not sites evolving under the substitution regime specified by a given matrix will display skew (A <> T and/or G <> C) at equilibrium.

The null hypothesis is that, for any matrix, substitution rates in the G row are equal to the complementary substitution rates in the C row (e.g., G → A, G → T, and G → C, compared to C → T, C → A, and C → G, respectively). In addition, substitution rates within the A row are equal to the complementary substitution rates within the T row. The test was performed using a χ^2 test with three degrees of freedom on the observed number of substitutions. A rejection of the null hypothesis means that at least one substitution type occurs at a higher frequency on one strand relative to the other and we would expect to observe skew in this context at equilibrium.

Expected Context-Dependent Codon Usage Bias. For each fourfold-degenerate codon group (CTN, TCN, CGN, ACN, GTN, GCN, CCN, and GGN), four separate expected codon usage tables were generated as a function of each 3′ flanking base. Contexts are written with an underscore representing the site of substitution, in this case the third codon position, and a "|" indicating the codon boundary. Thus, GG_|G represents the third position of glycine codons that are followed by a G. Expected codon usage is simply the stationary vector from the appropriate specific context matrix; for AC_|G sites, codon usage is determined by the stationary vector of the C_G substitution matrix. Significant differences between the expected table and the observed codon usage in that context were assessed in two ways. The first was a standard χ^2 test with three degrees of freedom. The second utilized a resampling of the matrix itself. Ten thousand random matrices, each with the same number of entries in every row as the empirical matrix, were generated by sampling with replacement from the empirical matrix. The sampling was performed for each row, and each row is treated as a multinomial so the probability of sampling an event was equal to the observed frequency of that event within the row. For example, if the empirical matrix has entries of 140 G → G, 20 G → A, 20 G → T, and 20 G → C changes in the row representing G as the original state, then the probability of sampling G → G is 70%, while the probability for each base substitution is 10%. For each random matrix the stationary vector was calculated, and the 95% confidence interval for each base was calculated from these 10,000 vectors. This interval is the range within which 95% of the vectors were observed to have an equilibrium frequency of that base, with equal numbers falling outside each end of the range. The rationale of this test is that, using the observed frequency of each evolutionary "transition" as our estimate of the probability of that event, it provides an estimate of how sampling error in the sub-

Table 1. Features of the specific context substitution matrices from the plus strand of noncoding cpDNA

Context[a]	T_s rate[b]	T_v rate	T_s/T_v	A + T[c]	Y–R skew[d]
G_G	0.166	0.051	3.2	76.9	−43.5
G_C	0.126	0.061	2.1	70.7	NS
C_G	0.145	0.091	1.6	64.7	−39.4
C_C	0.177	0.046	3.9	53.0	29.5
G_A	0.105	0.052	2.0	65.1	−33.1
A_G	0.174	0.080	2.2	78.7	−23.7
G_T	0.140	0.067	2.1	70.3	NS
A_C	0.138	0.068	2.0	77.9	−30.0
C_A	0.135	0.080	1.7	63.0	19.9
T_G	0.107	0.076	1.4	66.5	NS
C_T	0.193	0.076	2.5	76.8	78.7
T_C	0.071	0.070	1.0	76.5	41.4
A_A	0.076	0.091	0.8	78.8	−33.4
A_T	0.095	0.075	1.3	81.2	NS
T_A	0.073	0.093	0.8	80.5	NS
T_T	0.077	0.103	0.7	85.5	29.5

[a] Context is given as 5′ flanking base and 3′ flanking base, with the underscore indicating the site of substitution (see text).
[b] Rate is given for transitions (T_s) and transversions (T_v).
[c] Predicted A + T content at equilibrium for sites that exist in each context.
[d] Predicted Y–R skew (see text) at equilibrium.

stitution matrix itself could affect our estimate of equilibrium composition bias. In the test, if a base is observed at a frequency that lies outside of its 95% confidence interval, it is scored as a significant deviation from expected usage given the empirical matrix.

To generate expected codon usage for twofold-degenerate sites, the parameters $\gamma 1$ and $\gamma 2$ were calculated for every matrix as discussed above, and the equilibrium A + T content of twofold-degenerate sites was calculated following Morton (2001). Since every twofold-degenerate group involves either an A or G, or a T or C, expected A + T content yields expected codon usage. Expected codon usage was generated for every twofold-degenerate codon group as a function of the 3′ flanking base by using the γ parameters from the appropriate matrix.

Results

Context-Dependent Variation in Substitution Properties of cpDNA. The overall strand symmetry test, which compares complementary matrix pairs (see Materials and Methods), showed no significant difference between the two strands of DNA (data not shown). Therefore, the plus and minus strands display the same substitution pattern and the substitution matrices derived from the plus strand analysis are applicable to both.

Substitutions on the plus strand vary noticeably as a function of context. The rates of both transition and transversion changes as well as the ratio of these two main substitution types are presented for each specific context matrix in Table 1. In addition, this table shows the predicted equilibrium A + T content for each matrix, which is the expected A + T content of sites evolving in that particular context, as well as the Y–R skew (Eq. 1) for those matrices that showed a significant deviation in the context symmetry test (see Materials and Methods). All of these features, or results, of the substitution process vary among contexts and indicate that the substitution process is context dependent. Perhaps the most dramatic variation is the different skew that is expected at equilibrium in each context, ranging from a strong skew toward pyrimidines (such as in the C_T context) to a strong skew toward purines (such as in the context G_G).

To examine some of the general trends in the context variation and to compare the findings to previously published work on substitution bias and context in cpDNA, matrices from specific contexts were combined into generic contexts. These are the generic contexts that were studied in earlier work (Morton 1995; Morton et al. 1997; Yang et al. 2002). Tables 2a and 2b give sample matrices: Table 2a shows the substitution matrices from the plus strand as a function of flanking base A + T content (the A + T = 0, A + T = 1, and A + T = 2 contexts), and Table 2b shows the substitutions in the A + T = 2 context, which was chosen since it has the highest number of sites, broken down by Y context. Some of the variation among these generic context matrices, along with the substitutions occurring in the different Y contexts within the A + T = 0 and A + T = 1 contexts, is demonstrated by the general properties that are summarized in Table 3. Variation across A + T context shows trends that are consistent with those described previously for cpDNA sequences (Morton 1995, 2000; Yang et al. 2002). Although there is variation between specific contexts within each generic context, it is seen in Table 3 that as the A + T content of the flanking bases increases, the rate of transitional changes decreases, with the

Table 2a. Plus-strand noncoding cpDNA substitution matrices in different A + T contexts

Context	T_o G	A	T	C
0 A + T				
G	505	32	3	4
A	15	883	5	5
T	4	9	871	28
C	8	2	22	522
1 A + T				
G	2341	123	32	26
A	85	5060	30	35
T	39	18	5146	75
C	21	23	121	2183
2 A + T				
G	2314	78	50	26
A	41	5427	34	40
T	30	38	5486	30
C	30	58	89	2459
Total				
G	5160	233	85	56
A	141	11370	69	80
T	73	65	11503	133
C	59	83	232	5164

Table 2b. Noncoding cpDNA substitution matrices for sites in the A + T = 2 context, analyzed as a function of the number of flanking pyrimidines on the plus strand

Context[a]	T_o G	A	T	C
2 A + T, 0Y (A_A)				
G	922	26	18	12
A	15	1893	5	7
T	12	9	1140	9
C	5	12	16	437
2 A + T, 1Y (A_T & T_A)				
G	1010	36	25	9
A	22	2306	14	18
T	12	18	2390	12
C	16	13	43	1086
2 A + T, 2Y (T_T)				
G	382	16	7	5
A	4	1228	15	15
T	6	11	1956	9
C	9	33	30	936

[a] The context is expressed in terms of A + T and Y (see text) and the specific contexts that represent each case are also indicated.

result that the T_s:T_v ratio decreases and the overall substitution rate decreases. The data in Table 3 also indicate that G and C undergo a higher rate of change, transitions in particular, than do A and T but that it is a decrease in the rate of change from A and T that is largely responsible for the overall rate decrease. A consequence of this is that the A + T content of flanking bases has an influence on the A + T content of a site, as seen by the fact that the equilibrium A + T content predicted from each matrix is higher when the flanking bases themselves have a higher A + T content (Table 3).

Along with A + T context, Y context also has an influence on substitution bias and the resulting equilibrium composition. Substitutions occurring at sites that are flanked on both sides by a pyrimidine on the same strand lead to equilibrium compositional biases that tend to be skewed toward pyrimidines. The amount of Y–R skew expected at equilibrium is given for those generic context matrices that gave a significant result for the context symmetry test (see Materials and Methods), indicating that the two strands have significantly different substitution patterns at the sites that have that context on the plus strand. When both flanking nucleotides are a purine there is a strong negative skew (i.e., toward purines), and when they are both pyrimidines there is a strong positive skew in the expected equilibrium composition, a skew that is also apparent in the matrices themselves (Table 2b). In contrast, neither the single pyrimidine generic contexts nor the generic A + T matrices show any significant difference between the two strands (Table 3). A prominent feature that contributes to this aspect of the equilibrium composition is the difference in the rate of Y → R as opposed to R → Y transversions. When both flanking nucleotides are purines we see a much higher rate of Y → R transversions, while the opposite relationship is observed when both flanking nucleotides on a strand are pyrimidines (Table 3). In general, Table 3 indicates that there appear to be some basic trends in the relationship between context and substitution bias, although it is also clear from Table 1 that the full nature of the relationship is complex.

If we assume that the vast majority of substitutions in noncoding sites are the result of fixation of neutral mutations by random genetic drift, then the substitution matrices from noncoding cpDNA are an accurate representation of mutational dynamics in cpDNA. The context dependency is likely, then, to arise from a combination of the influence of the 5′ neighbor on polymerase misincorporations (e.g., Petruska and Goodman 1985) and an influence of local base composition on the efficiency of the mismatch repair process (e.g., Radman and Wagner 1986). Even if the substitution matrices from noncoding cpDNA represent a significant influence of selection, this selection could be a general pressure across both noncoding and coding sequences. In either case, if we assume that silent sites of coding sequences are at equilibrium, we can use these predicted equilibrium compositions to test for selective constraints that are specific to these sites. The one difficulty this method faces, of course, is if future studies show significant selective constraints on noncoding sequences that do not exist on coding sequences. We assume here that substitution dynamics from non-

Table 3. Features of the plus strand substitutions occurring in generic contexts

A + T context[a]	Y context[b]	T_s rate	T_v rate	T_s/T_v	GC rate[c]	AT rate	Total rate	A + T[d]	Y–R skew[e]
0 A + T	All	0.0332	0.0137	2.4	0.0647	0.0363	0.0469	66.5	NS
1 A + T	All	0.0263	0.0146	1.8	0.0710	0.0269	0.0409	73.4	NS
2 A + T	All	0.0147	0.0189	0.78	0.0649	0.0191	0.0335	81.0	NS
Total	All	0.0214	0.0165	1.3	0.0680	0.0240	0.0379	75.9	NS
		R → Y rate	Y → R rate						
0 A + T	0Y	0.0143	0.0370	3.2	0.0692	0.0302	0.0448	76.9	−43.5
0 A + T	1Y	0.0305	0.0453	1.8	0.0690	0.0370	0.0491	71.5	NS
0 A + T	2Y	0.0297	0.0163	3.9	0.0530	0.0415	0.0458	53.0	29.5
1 A + T	0Y	0.0076	0.0175	1.85	0.0661	0.0259	0.0383	71.7	−25.2
1 A + T	1Y	0.0167	0.0195	1.68	0.0686	0.0337	0.0455	68.4	NS
1 A + T	2Y	0.0225	0.0061	2.03	0.0729	0.0207	0.0373	80.3	63.5
2 A + T	0Y	0.0104	0.0232	0.84	0.0615	0.0184	0.0322	78.8	−33.4
2 A + T	1Y	0.0192	0.0164	1.03	0.0634	0.0200	0.0339	80.3	NS
2 A + T	2Y	0.0251	0.0197	0.74	0.0705	0.0185	0.0343	85.5	29.5

[a] Number of AT base pairs immediately flanking the site.
[b] Number of pyrimidines on the plus strand immediately flanking the site.
[c] Rate of change from either GC or AT is given.
[d] Equilibrium A + T content determined from the stationary vector of the matrix.
[e] The equilibrium pyrimidine skew (Eq. 1) is given for those matrices with significant asymmetry (NS, not significantly different from zero skew).

coding cpDNA can be used to predict equilibrium composition in the absence of selective constraints.

Variation in mutation dynamics in different contexts can influence molecular evolution. Of particular interest here is that silent sites of coding sequences exist in a limited set of contexts and thus could evolve, on average, very differently than noncoding DNA just because of context-dependent mutation. This has two implications that are investigated here. First, it complicates direct comparisons of compositional properties between the silent sites from coding sequences and noncoding DNA. Differences in overall composition bias, as observed in cpDNA (Morton 2001), may reflect a difference in average context rather than an influence of selection. Second, if we wish to determine the codon usage generated by mutation bias alone, which is necessary to assess any role selection may have played, then we need to consider context. Different codon groups can have different second position nucleotides, meaning that they may have very different equilibrium composition (i.e., codon) biases. We also need to consider codon usage as a function of the 3′ flanking nucleotide. These two points are considered below in applications of the observed substitution dynamics to sequence evolution.

General Composition Biases in Silent Coding Sites and Noncoding cpDNA. A comparison of the contexts in which sites occur in noncoding cpDNA to those of fourfold-degenerate sites in low-expression chloroplast genes shows that they are quite different (Table 4). While roughly half the sites in each case occur in an A + T = 1 context, the fourfold-degenerate coding sites are observed much more frequently in an A + T = 0 context. As for Y context, the Y = 0 and Y = 2 contexts are equally frequent in noncoding cpDNA in all A + T contexts, as expected for symmetrical sequences. However, a much higher proportion of fourfold-degenerate coding sites exists in the Y = 2 context than in the Y = 0 context. These differences in average context are a result of the

Table 4. The proportion of fourfold-degenerate sites observed within various generic contexts in both coding and noncoding cpDNA

A + T context	Y context	Coding	Noncoding[a]
0 A + T	0Y	7.6%	2.7%
	1Y	20.5%	4.0%
	2Y	9.5%	2.6%
1 A + T	0Y	7.6%	12.2%
	1Y	28.5%	21.0%
	2Y	14.0%	12.0%
2 A + T	0Y	0	12.5%
	1Y	5.9%	20.0%
	2Y	6.4%	12.9%
All	0Y	15.2%	27.5%
All	1Y	54.9%	44.9%
All	2Y	29.9%	27.6%
0 A + T	All	37.5%	9.4%
1 A + T	All	50.1%	45.2%
2 A + T	All	12.3%	45.4%

[a] All noncoding sites are considered fourfold degenerate.

Table 5. General composition features in coding and noncoding cpDNA

	Noncoding DNA		Fourfold-degenerate sites in coding sequences		All silent sites in coding sequences	
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
G	3388	2587.2	456	462.1	726	679.4
A	7169	7852.2	1192	1242.6	1922	2027.3
T	7200	8157.8	1536	1466.0	2714	2710.3
C	3368	2534.3	633	639.3	1028	968.0
A + T	68.0%	75.8%	71.5%	71.1%	72.6%	74.2%
Skew	0.05%	1.2%	13.6%	10.5%	n/a[a]	n/a

[a] Not calculated since it depends on the frequencies of amino acids coded by twofold-degenerate codon groups.

Table 6. GC → AT pressures for transitions and transversions in each substitution matrix

	GC → AT[a]	
Context	T_s	T_v
G_G	2.06	NA[b]
G_C	2.71	1.72
C_G	2.83	0.35
C_C	1.74	0.63
G_A	2.16	0.80
A_G	5.29	2.18
G_T	2.27	2.94
A_C	4.15	2.42
C_A	2.25	1.28
T_G	3.69	0.69
C_T	3.64	1.89
T_C	3.48	2.55
A_A	3.91	3.16
A_T	4.42	3.89
T_A	7.31	1.86
T_T	8.92	3.34

[a] GC-to-AT pressure is calculated as the rate of GC → AT change divided by the rate of AT → GC change. The value is given separately for transitions only (T_s) and transversions only (T_v).
[b] Requires division by zero.

fact that, while sites in noncoding cpDNA occur in all contexts randomly, only two of the eight fourfold-degenerate codon groups have a T at the second position (GTN and CTN), while none have an A at the second position and only two (GGN and CGN) have a purine at the second position. Furthermore, constraints in coding sequences result in a much higher G + C content at the first codon position, the positions that are the 3′ flanking bases of fourfold-degenerate sites, relative to noncoding cpDNA (Morton 1997).

Given these differences in average context and the variation in mutational biases observed across contexts, we would expect the average compositional bias to be very different for noncoding cpDNA and fourfold degenerate sites in coding sequences. This was tested by considering each specific context separately and generating an expected equilibrium base composition in *Zea mays* cpDNA. The expected equilibrium base composition is simply the stationary vector of the substitution matrix for that context multiplied by the number of sites that occur in that context. The overall expected base composition was then calculated by summing across all 16 contexts. This calculation was performed separately for both fourfold- and twofold-degenerate sites of low-expression genes, as well as for the noncoding regions of *Z. mays*.

Table 5 compares the expected and observed overall equilibrium composition. The observed compositional differences between coding and noncoding cpDNA noted previously are apparent in Table 5. First, silent sites in coding sequences have a higher A + T content than noncoding cpDNA. The most striking feature, though, is that this difference in A + T content is not predicted by the substitution dynamics but is due to the fact that noncoding cpDNA is not at the equilibrium predicted from its own substitution dynamics. In contrast, the coding sites are extremely close to the expected composition. Either the noncoding sequences have not reached equilibrium, which seems unlikely given the match observed for coding sites, or there are other mutational factors, such as indels, that contribute significantly to the evolution of noncoding sequences. The observation that noncoding cpDNA has a lower-than-expected A + T content is true for all specific contexts, not just in the cumulative base composition, although it is particularly strong in the contexts where both flanking base pairs are AT (data not shown). The discrepancy between the observed and the expected A + T content of noncoding cpDNA is discussed further below. The second general difference between noncoding and coding cpDNA is the skew observed at fourfold-degenerate sites in the latter. The data in Table 5 indicate that the bias toward two-Y contexts in coding sequence (Table 4) leads to the expectation that there will be skew at these sites.

Another feature that is apparent from Table 5 is that twofold-degenerate sites have a slightly higher A + T content, both observed and expected, than

Table 7a. Observed and expected codon usage for fourfold-degenerate groups when the 3′ flanking base is a pyrimidine

	3′ T		3′ C	
	Obs.	Exp.	Obs.	Exp.
CTG	6	8.2	4	7.5
CTA	30	35.7	16	12.1
CTT	71	68.2	39	34.3
CTC	19	14.1	6	10.9
TCG	4	3.4	9	8.1
TCA	13	14.3	10	12.3
TCT	38	39.6	29	28.0
TCC	15	12.7	22	15.9
CGG	2	6.9	5	6.5
CGA	21	20.7	20	26.1
CGT	37	28.4	24	20.5
CGC	6	8.8	17	12.8
ACG	10	4.9	4	10.1
ACA	23	20.4	26	15.3
ACT	53	56.6	40	41.9
ACC	14	18.1	17	19.7
GTG	6	7.7	12	11.1
GTA	43	33.7	27	17.9
GTT	58	64.3	48	50.7
GTC	12	13.3	9	16.2
GCG	6	6.5	12	16.0
GCA	33	27.1	25	31.1
GCT	75	75.3	84	66.4
GCC	19	24.1	17	31.3
CCG	4	3.5	8	7.8
CCA	14	14.5	18	11.8
CCT	36	40.2	31	32.2
CCC	17	12.9	10	15.2
GGG	18	17.1	22	14.8
GGA	62	51	48	59.0
GGT	62	70.3	50	46.3
GGC	21	24.8	29	28.9

Table 7b. Observed and expected codon usage for fourfold-degenerate groups when the 3′ flanking base is a purine

	3′ G		3′ A	
	Obs.	Exp.	Obs.	Exp.
CTG	19	28.8	11	10.3
CTA	41	41.4	27	26.2
CTT	47	37.9	42	49.5
CTC	15	14.0	14	8.0
TCG	15	20.8	13	13.3
TCA	20	54.4	29	36.1
TCT	37	15.4	30	39.3
TCC	36	17.3	44	27.4
CGG	8	18.2	14	15.0
CGA	34	38.5	24	47.0
CGT	26	14.1	43	16.5
CGC	12	9.3	8	10.3
ACG	17	27.0	18	16.9
ACA	49	70.6	47	45.7
ACT	42	20.0	57	49.8
ACC	32	22.4	25	34.7
GTG	37	42.0	17	14.5
GTA	80	60.3	54	36.8
GTT	39	55.4	39	69.6
GTC	22	20.5	22	11.2
GCG	27	37.1	20	20.5
GCA	75	96.8	60	55.4
GCT	66	27.5	77	60.3
GCC	24	30.7	21	42.0
CCG	20	24.3	8	13.6
CCA	30	63.5	32	36.7
CCT	42	18.0	50	40.0
CCC	35	20.2	28	27.8
GGG	60	47.4	20	34.1
GGA	65	100.5	96	106.7
GGT	56	36.8	69	37.6
GGC	28	24.2	17	23.4

fourfold-degenerate sites. The reason for this is that transitions have a stronger GC → AT pressure than do transversions in all but one specific context, as seen in the summary of GC → AT pressures for different matrices (Table 6). As a result, twofold-degenerate sites, which undergo primarily transition changes, have a higher A + T content than sites that can also undergo transversions. (Note that the parameters tend to increase in value along with increasing A + T context, resulting in the higher A + T content in these contexts seen in Table 3.)

Overall, the results in Tables 4, 5, and 6 tell us that the general compositional differences between silent sites of low-expression chloroplast genes and noncoding cpDNA (Morton 2001) cannot be taken as evidence for selection on the former. Without a careful consideration of the context-dependent mutational process, we have no way to assess this comparison. The results illustrate why previous attempts to detect an influence of selection on codon usage by using such comparisons (e.g., Morton 1998, 2001) were incorrect. Instead, we must analyze expected codon usage of chloroplast genes as a function of local context.

Context-Dependent Substitutions and Codon Usage in Low-Expression Chloroplast Genes. The availability of context-dependent substitution models from noncoding cpDNA allows us to make better predictions of the codon usage that arises from mutation bias alone than was possible with simpler models. This section looks at the codon usage bias of the low-expression chloroplast genes listed under Materials and Methods to determine if mutation dynamics alone are sufficient to explain their codon bias. Expected context-dependent codon usage was calculated separately for each codon group based on the number of occurrences immediately upstream of each of the four bases (the 5′ flanking base is defined by the base at the second position of the codon group; see Ma-

Table 8. Observed and expected codon usage for twofold-degenerate groups as a function of the composition of the 3′ flanking base

	3′ Y		3′ R	
	Obs.	Exp.	Obs.	Exp.
CAT	72	75.7	89	91.4
CAC	23	19.3	24	21.6
CAG	21	23.1	54	36.3
CAA	80	77.9	118	135.7
GAG	33	33.3	65	47.3
GAA	113	112.7	153	172.7
GAT	93	93.9	122	131.2
GAC	27	26.1	40	30.8
AAT	95	98.9	129	138.5
AAC	31	27.1	42	32.5
AAG	30	27.4	67	49.9
AAA	90	92.6	176	193.1
TAT	101	97.3	116	115.8
TAG	22	25.7	27	27.2
TTT	132	172.2	176	189.3
TTC	84	43.8	66	52.7

terials and Methods). The observed and expected codon usages are shown in Table 7a for sites with a downstream pyrimidine and in Table 7b for sites with a downstream purine. Agreement between observed and expected codon usage was assessed with a χ^2 test with three degrees of freedom and a resampling test (see Materials and Methods). Those cases with a significant difference ($p < 0.05$) for both tests are boxed.

For fourfold-degenerate sites upstream from a pyrimidine, codon usage matches the expected bias caused by context-dependent mutations in 15 of 16 cases (Table 7a). This match between observed and expected exists in spite of the fact that the bias, usually in the form of a pyrimidine skew, is quite strong in many cases. In addition, the observed bias is quite different from what would be expected if we were to use noncoding base composition as a basis (see Table 5) or if we were to predict codon usage using all substitutions in noncoding cpDNA without regard to context (see row headed "Total" in Table 3). Overall, the context-dependent substitution model predicts codon usage bias remarkably well for those sites immediately upstream from a pyrimidine.

The results are quite different for sites with a downstream purine (Table 7b), for which 9 of the 16 cases show highly significant deviation. A few observations suggest that this deviation is a result of some form of weak selective pressure and is not simply due to an inaccuracy in the substitution models. First, the resampling test that generates the 95% confidence interval accounts for sampling error in the matrices. Second, there is no consistent deviation by context: codon groups that have the same base at the second position show different results. For example, CT_|R (where _ represents the degenerate position and | represents the codon boundary) matches expected codon use well but GT_|R shows a strong bias toward GTA|R, a bias not predicted from the substitution matrix that would lead to a bias toward GTT|R. Since the degenerate sites of CT_|R and GT_|R are in the same context, this difference cannot be explained as a result of an inaccurate substitution model being used to predict codon usage. A similar difference is noted among TC_|R, GC_|R; and AC_|R. Third, there is an interesting correlation in many cases between the contexts with a 3′ G and those with a 3′ A, which are tested with different substitution matrices. The significant bias toward GTA noted above occurs in both contexts (GTA|G and GTA|A), as does the significant bias toward CGT (Table 7b). Most interesting is the case of TC_|R, in which the codon TCC, which is not common upstream from a pyrimidine, is highly represented upstream from both G and A, a result not expected from substitutions occurring in either context. In contrast, neither ACC nor GCC is observed at a high frequency in the same contexts so this "preference" for C in the C_G and C_A contexts is limited to the TCC codon. Fourth, analysis of substitutions from noncoding cpDNA as a function of the composition of three flanking nucleotides, two 5′ and one 3′, shows no difference from the matrices presented here (data not shown), so the results in Tables 7a and 7b do not appear to be a result of more complex context dependency. Finally, the match between the observed and the expected codon usage when there is a 3′ pyrimidine indicates that there is not a general violation of our assumption that sequences are at equilibrium. Taken together, these points suggest that deviations from expected codon usage at sites upstream from a purine are due in part to some form of selective pressure.

The observed and expected codon usages for twofold-degenerate amino acids are listed in Table 8, excluding cysteine, which is extremely rare. A χ^2 test was used to test significance in each case. The resampling test was not applied since the confidence intervals apply to the whole matrix, not to a subset of substitutions. We can take this approach to estimating codon usage at twofold-degenerate sites if we assume that transversion changes (i.e., nonsynonymous) at the third position are relatively rare. The results are shown for downstream purines and pyrimidines, not specific bases, since, in terms of significant deviation, they were the same for G and A as well as for T and C (data not shown). Cases for which a significant deviation is observed are boxed. As with the fourfold-degenerate groups, there is a relationship

Table 9. Codon usage at phenylalanine sites as a function of the 3′ base

	3′ G		3′ A		3′ T		3′ C	
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
TTT	120	106.6	56	82.7	68	100.2	64	72.0
TTC	26	39.4	40	13.3	53	20.8	31	23.0

between deviation from expectation and a 3′ purine. Twofold-degenerate groups match expected codon usage extremely well when there is a downstream pyrimidine; ignoring TTY for the moment, codon usage is not significantly different than expected. Once again, though, when the downstream base is a purine, three of the codon groups, the three that have a purine at the third position, fail to match precisely the expected codon bias, although the difference is not remarkable.

The codon group that stands out is the one coding phenylalanine (TTY). Codon usage in this group is significantly different than expected when the downstream base is either a pyrimidine or a purine. Unlike the other twofold-degenerate codon groups, in this case the deviation occurs most strongly when there is a downstream A or T. For both bases there is a significant underrepresentation of TTT (Table 9). With a total of 124 TTT and 93 TTC codons upstream from an A or T, the observed 42.9% TTC is in marked contrast to the 19% G + C content expected in high-A + T contexts (Table 3) and the 15.7% expected given the higher GC → AT pressure of transitions in this context (Tables 6 and 9). This difference between observed and expected seems to be more consistent with some sort of selection against UUUU and UUUA runs in the mRNA than with mutation bias.

Overall, context dependency can account for a major part of the codon usage bias in the low-expression genes. However, the fact that there is a noticeable pattern to the deviation observed in Tables 7a, 7b, 8, and 9 suggests that the codon usage of these genes is not free from selective constraints. The bias toward TCC, CGT, and GTA upstream from a purine, as well as CCT upstream from a G, and the high frequency of TTC upstream from an A or T cannot be explained by the observed substitution patterns in cpDNA.

Discussion

The matrices generated from the alignment of noncoding regions of three grass chloroplast genomes show that substitution dynamics in cpDNA are correlated with flanking base composition. In general, variation in both local A + T composition and pyrimidine content along one strand is correlated with variation in substitution bias and, as a result, with expected equilibrium composition bias. This substitution pattern most likely reflects mutation patterns, and the variation in substitution bias across contexts is probably due to an influence of local base composition on polymerase misincorporation and/or mismatch repair. Flanking base composition has been demonstrated to influence both of these processes in *E. coli*. Biochemical studies have shown that the base immediately 5′ from a site can affect misincorporation bias (Petruska and Goodman 1985; Mendelman et al. 1989) and the local A + T composition affects the relative efficiency of the repair of different mismatches (Jones et al. 1987; Radman and Wagner 1986). Similar effects might explain the pattern of context dependency that is indicated by the cpDNA substitution matrices.

The influence of context on mutation dynamics introduces to studies of molecular evolution the problem that mutational dynamics will vary across sites and across time at a specific site (Morton 1995; Berg and Silva 1997). Codon–codon transition matrices (e.g., Lewontin 1989; Muse and Gaut 1994; Yang and Nielsen 2000) account for this to some degree but those developed to date do not consider the influence of the 3′ flanking base. The influence of context on substitution bias also makes it critical to consider base composition bias and codon usage bias as functions of context. This is particularly important if we wish to make inferences regarding selection by contrasting sequence compositions.

Two lines of analysis were pursued here to consider how context dependency might affect our understanding of molecular evolution in the flowering plant chloroplast genome. The first considered how context dependency affects the direct comparison of the base composition of noncoding cpDNA to the composition of silent sites in coding sequences. The second involved a more detailed consideration of codon usage in low-expression genes. These are discussed separately below.

Comparing Silent Sites from Coding and Noncoding DNA. If mutations are context- dependent, then comparisons of base composition across sites, such as the application of Parity Rule 2 (PR2; Sueoka 1999), comparisons of intron, or spacer, and exon composition (e.g., Carulli et al. 1993), and the use of noncoding DNA composition to predict "neutral" codon usage (Morton 1998), are all problematic and should be treated with caution. If the sites being compared differ in their average context, then we cannot assume that they should evolve to the same equilibrium composition bias.

This problem is apparent in the chloroplast genome. The difference in skew at fourfold-degenerate

sites of coding and noncoding cpDNA (Morton 200l) (Table 5) would be interpreted by PR2 as evidence for selection. However, since fourfold-degenerate sites of chloroplast protein-coding genes tend to occur in contexts that are G + C-rich, and/or have a high pyrimidine content on the coding strand, the mutational dynamics of cpDNA are expected to generate a significant skew toward pyrimidines in the absence of selection (Table 5). In general, the argument that fourfold-degenerate sites within genes should show no skew if noncoding DNA is symmetrical ignores the possibility of context dependency, so inferring selection from PR2 (Sueoka 1999; Sueoka and Kawanishi 2000) could be misleading.

Another difficulty with comparing composition bias is apparent in the different A + T contents of silent sites from coding and noncoding cpDNA (Table 5). One factor that may contribute to this difference is that the A + T content of noncoding cpDNA is much lower than the expected equilibrium A + T content given the substitutions occurring within noncoding sequences. This deviation is very interesting and warrants further study. It is quite likely that many noncoding spacers deviate from the expected equilibrium composition due to a prevalence of indels. It has been shown that a segment of DNA, roughly 1 kb in length, is present in the middle of a spacer region in the chloroplast genome of some grasses (Morton and Clegg 1993). This region contains a pseudogene and two ORFs (Morton and Clegg 1993) and has a lower A + T content than the surrounding spacer DNA (Morton and Clegg 1995). Given this composition, it may be a coding region duplication that was somehow inserted into a spacer. The segment appears to have lost any function since many sections of this region have subsequently been lost by deletion in descendent lineages (Morton and Clegg 1993). This observation suggests that the A + T content of noncoding cpDNA could be influenced by more than just base substitutions.

In addition to the contribution of indels to the evolutionary process, we must consider the possibility that noncoding cpDNA has simply not reached equilibrium. Mutational dynamics may have changed recently enough that there has been insufficient time for cpDNA to evolve to equilibrium. Such a scenario is difficult to test, but it does not seem to be as good an explanation as a contribution of indels to molecular evolution, since it fails to provide a good explanation of why coding sequences match expected composition bias so well. Of course, the analysis of base composition also assumes that noncoding cpDNA is essentially neutral. However, we must keep in mind the possibility that a number of noncoding sites are constrained by selection. With the growing evidence for functional noncoding RNA in some organisms (Klein et al. 2002; Storz 2002), we cannot ignore the possibility that functional constraints contribute to the deviation from expected composition bias.

Finally, we also need to consider the possibility that a certain fraction of sites in any given context may have been in that context for only a short period of time due to a substitution at a flanking site, which would affect our calculation of the equilibrium composition bias within a particular context. The potential effect of this was tested by simulating the evolution of a DNA sequence 100,000 nucleotides in length and undergoing substitutions in a context-dependent manner defined by the empirical matrices generated from noncoding cpDNA. The sequence was mutated randomly until roughly 100 substitutions had occurred per site. Composition bias was calculated by context at random times during this process. The composition bias in each context was never significantly different than expected (data not shown) so it does not appear that fluctuating contexts have a significant impact on our calculations. However, the complex sequence evolution possible with context dependency will need to be studied in much greater depth in future work.

Implications for Testing Selection on Translation Accuracy. A comparison that is closely related to those discussed above is the test for selective pressures created by translation accuracy optimization (Akashi 1994). The idea of this test is to compare codon usage of conserved (i.e., ''critical'') amino acid sites to codon usage at variable amino acid residues. An implicit assumption underlying the comparison of variable and conserved residues is that amino acid selection does not influence codon usage. However, the complex mutational dynamics observed in the chloroplast genome invalidate this assumption. One of the consistent features of the matrices presented here is that the GC $\rightarrow$ AT pressure of transitions ($\gamma 1$ from Morton [2001]) is consistently greater than the GC $\rightarrow$ AT pressure of transversions ($\gamma 2$; see Table 6). When $\gamma 1$ and $\gamma 2$ are unequal, the GC $\rightarrow$ AT pressure at twofold-degenerate sites can be indirectly influenced by selection at the amino acid level. This is because the third positions of conserved twofold-degenerate codons are limited to transitions and, therefore, have a different GC $\rightarrow$ AT pressure than noncoding sites, and fourfold-degenerate sites in general, at which both transitions and transversions have occurred. They will also have a different GC $\rightarrow$ AT pressure than the third positions of variable twofold-degenerate codons, since the latter will have undergone some transversion changes.

The demonstration that $\gamma 1$ and $\gamma 2$ can be very different draws into question the comparison of composition bias at different sites, such as done in the test for selective pressures on translation accuracy. In

general, this test should never be applied unless it has been clearly established that the GC → AT pressure of both transitions and transitions is equivalent. Otherwise, the null hypothesis of this test is flawed. If mutation dynamics were also context- dependent, as is the case in cpDNA, then it would also be necessary to limit comparisons to sites within the same context.

Context dependency can also generate problems for a comparison of conserved and variable amino acid sites since different fourfold-degenerate codon groups can differ significantly in their bias (Tables 7a and 7b). This is similar to the problem of fluctuating context over time mentioned above in reference to noncoding DNA. However, instead of taking an average across sites as discussed above, in this case we would specifically be comparing those sites with conserved contexts (depending on the evolution of the 3′ codon) to sites that have changed contexts due to a change at a flanking site, since at least half of the variable sites are expected to have undergone a change at the second codon position. For example, if we were to take a random selection of GCN codons upstream of a purine and change them to GTN codons, their codon usage would be significantly different than that of GTN codons that had evolved to equilibrium. Even if they are not of sufficient number to affect the overall bias, as was the case in the simulation study mentioned above, the specific comparison of these sites to sites in a conserved context is a different matter. How significantly this might affect such a comparison would be difficult to measure, but it raises questions about making the comparison without considering such a possibility.

Codon Usage in the Chloroplast Genome. Context dependency was also considered in an analysis of codon usage in low-expression chloroplast genes. The equilibrium compositions predicted by the Markov matrix for each specific flanking base context were used to predict neutral codon usage for each codon group as a function of the 3′ base. Although the codon usage bias matches the expected usage in many contexts (Tables 7a, 7b, and 8), there are some significant deviations from the predicted codon usage that occur in a pattern that is suggestive of some action of selection. First, excluding TTY, which is discussed separately below, 12 of the 13 contexts in which a significant deviation is observed include a 3′ purine. Second, the deviation in a particular synonymous group almost always occurs for both a 3′ G and a 3′ A, and involves a bias toward the same codon in both cases (Tables 7a and 8). Third, as noted above, there are differences between synonymous groups that have the same context for the third position, such as TCN, ACN, and GCN. The strong overrepresentation of TCC|R, but the lack of a similar bias toward ACC|R and GCC|R, even though the context is the same, indicates that significantly different substitutions have occurred at the third positions of these different synonymous groups. Finally, the case of TTY shows a distinct deviation, in which TTC is observed much more frequently than expected when the 3′ flanking nucleotide is an A or T (Table 9). Taken together, these four features suggest that weak selection has played a role in generating the codon usage pattern of low-expression chloroplast genes.

We must consider the possibility that our assumption that sequences are at equilibrium is violated, such as would be the case if sequences have not had sufficient time to evolve to equilibrium following a shift in mutation dynamics. However, it is very striking that codon usage matches the predicted usage when upstream from a pyrimidine. Given this, it is proposed that the deviations correlated with a 3′ flanking purine are due to a weak selection, possibly due to constraints on mRNA structure. Overall, though, the selection appears to influence codon choice at a minority of sites; if we were to change just roughly 5% of the codons in Tables 7a, 7b, and 8 (about 300 of 6283), we would have a match with expected codon usage. Although this is not meant to be an accurate measure of the number of sites under selective constraint, it suggests that such constraints affect codon usage at a small proportion of sites. Therefore, it is proposed that the data indicate that the codon usage of these genes is not entirely free of selective constraints and that a full understanding of the influence of selection must account for the complex mutation dynamics of this genome.

Summary and Conclusions. The results presented here suggest that mutation dynamics of cpDNA are influenced by context in a complex manner and that this complexity can confound comparisons of the compositional bias of noncoding DNA and silent sites of protein-coding sequences. In general, such comparisons appear to be more problematic than often assumed. In addition, an analysis of the codon usage bias in low-expression chloroplast genes indicates that weak selection may be acting. The complexity of mutations observed here should be tested in other genomes since it raises a number of important questions about the evolutionary process at the molecular level.

References

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

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

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

Antezana MA, Kreitman M (1999) The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J Mol Evol 49:36–43

Berg OG, Silva PJN (1997) Codon bias in *Escherichia coli*: The influence of codon context on mutation and selection. Nucleic Acids Res 25(7):1397–1404

Blake RD, Hess ST, Nicholson-Tuell J (1992) The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. J Mol Evol 34:189–200

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

Carulli JP, Krane DE, Hartl DL, Ochman H (1993) Compositional heterogeneity and patterns of molecular evolution in the Drosophila genome. Genetics 134:837–845

Cox DR, Miller HD (1965) The theory of stochastic processes. Chapman and Hall, NewYork

Duvall MR, Morton BR (1996) Molecular Phylogenetics of Po aceae: an expanded analysis of RbCl sequence data. Mol Phylogenet Evol 5:352–358

Fay JC, Wyckoff GJ, Wu C-I (2002) Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 415:1024–1026

Gerber AS, Loggins R, Kumir S, Dowling TE (2001) Does non-neutral evolution shape observed patterns of DNA variation in animal mitochondrial genomes? Annu Rev Genet 35:539–566

Hess ST, Blake JD, Blake RD (1994) Wide variations in neighbor-dependent substitution rates. J Mol Biol 236:1022–1033

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

Jones M, Wagner R, Radman M (1987) Repair of a mismatch is influenced by the base composition of the surrounding nucleotide sequence. Genetics 115:605–610

Klein RJ, Misulovin Z, Eddy SR (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. Proc Natl Acad Sci USA 99:7542–7547

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

Konu O, Li MD (2002) Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents. J Mol Evol 54:35–41

Labate JA, Biermann CH, Eanes WF (1999) Nucleotide variation at the *runt* locus in *Drosophila melanogaster* and *Drosophila simulans*. Mol Biol Evol 16:724–731

Lewontin R (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

Llopart A, Aguade M (2000) Nucleotide polymorphism at the RpII215 gene in *Drosophila subobscura*: Weak selection on synonymous mutations. Genetics 155:1245–1252

MacDonald JH, Kreitman M (1991) Adaptive evolution at the Adh locus in *Drosophila*. Nature 351:652–654

Mendelman LV, Boosalis MS, Petruska J, Goodman MF (1989) Nearest neighbor influences on DNA polymerase insertion fidelity. J Biol Chem 264:14415–14423

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 (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 BR (1997) The influence of neighboring base composition on substitutions in plant chloroplast coding sequences. Mol Biol Evol 14:189–194

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 (2000) Codon bias and the context dependency of nucleotide substitutions in the evolution of plastid DNA. Evol Biol 31:55–103

Morton BR (2001) Selection at the amino acid level can influence synonymous codon usage: Implications for the study of codon adaptation in plastid genes. Genetics 159:347–358

Morton BR, Clegg MT (1993) A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near *rbcL* in the grass family (Poaceae). Curr Genet 24:357–365

Morton BR, Clegg MT (1995) Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. J Mol Evol 41:597–603

Morton BR, Oberholzer VM, Clegg MT (1997) The influence of specific neighboring bases on substitution bias in noncoding regions of the plant chloroplast genome. J Mol Evol 45:227–231

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–724

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

Petruska J, Goodman MF (1985) Influence of neighboring bases on DNA polymerase insertion and proofreading fidelity. J Biol Chem 260:7533–7539

Radman M, Wagner R (1986) Mismatch repair in *Escherichia coli*. Annu Rev Genet 20:523–538

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

Smith NGC, Eyre-Walker A (2001) Synonymous codon bias is not caused by mutation bias in G + C-rich genes in humans. Mol Biol Evol 18:982–986

Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415:1022–1024

Storz G (2002) The expanding universe of noncoding RNAs. Science 296:1260–1263

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

Sueoka N, Kawanishi Y (2000) DNA G + C content of the third codon position and codon usage biases of human genes. Gene 261:53–62

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

Yang Y-W, Tai P-Y, Chen Y, Li W-H (2002) A study of the phylogeny of *Brassica rapa*, *B. nigra*, *Raphanus sativus* and their related genera using non-coding regions of chloroplast DNA. Mol Phylogenet Evol 23:268–275

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