Codon Use and the Rate of Divergence of Land Plant Chloroplast Genes

Brian R. Morton
Department of Botany and Plant Sciences, University of California, Riverside

Codon fitnesses for chloroplast genes were estimated using the relative synonymous codon use of psbA, which has a different pattern of codon use than other chloroplast genes and is the major translation product of the chloroplast. These estimates were used to calculate the codon adaptation index (CAI) of chloroplast genes from Marchantia polymorpha, Nicotiana tabacum, and Chlamydomonas reinhardtii. The genes with the highest CAI values in M. polymorpha correspond to those that are expressed at the highest levels. The rate of divergence between M. polymorpha and both C. reinhardtii and N. tabacum is inversely related to the CAI value of the M. polymorpha gene. The data suggest that selection is acting on the synonymous codon use of the highly expressed genes of the M. polymorpha chloroplast genome. The data set is inconclusive about N. tabacum genes, but, as there is a weaker correspondence between CAI value and expression level, it suggests that selection is not operating in this lineage.

Introduction

The rate of nucleotide substitution is affected by the rate of mutation and the degree of selective constraints. The role of selective constraints is supported by a great deal of evidence, mostly from studies on rates of nonsynonymous substitution. Studies on mammalian genes have shown a great deal of variation in the rate of nonsynonymous substitutions (Li et al. 1985) resulting from variation of selective constraints on amino acid replacements across genes. Nonsynonymous substitutions also occur at a lower rate than synonymous substitutions (Li et al. 1985).

Studies have shown that the rate of synonymous substitution also varies among genes, although not as much as nonsynonymous rates (Li et al. 1985). As with nonsynonymous substitutions, there is evidence that at least some of this variation is due to selective constraints. Rates of synonymous substitution have now been studied in several species and found to be related to codon use (Sharp and Li 1987b; Shields et al. 1988; Sharp and Li 1989; Sharp and Li 1991) and sequence composition (Moriyama and Gojobori 1992). For Escherichia coli and Salmonella typhimurium, the rate of synonymous substitution of a gene is related to codon bias (Sharp and Li 1987b; Sharp 1991). Codon use in E. coli, as well as in Saccharomyces cerevisiae, is correlated with tRNA abundance (Ikemura 1985), and the degree to which codon use is biased in a gene corresponds to the expression level of that gene (Sharp and Li 1987a). It has been proposed that selection at the codon level acts to increase translation efficiency (Ikemura 1985; Sharp and Li 1987a). Therefore, the variation in synonymous rate is taken as a reflection of the varying intensity of natural selection on synonymous substitutions at different loci as a result of different levels of gene expression (Sharp and Li 1987b; Sharp 1991). Similar evidence for selective differences between synonymous codons has been presented for Drosophila genes (Shields et al. 1988; Sharp and Li 1989).

The chloroplast genome of plants codes for about 80 protein-coding genes as well as for 4 rRNA and 30 tRNA genes (Clegg et al. 1991). Proteins coded by the chloroplast are involved in the light and dark reactions of photosynthesis as well as in transcription and translation. Gene content of chloroplasts is conserved among the higher plants, with a few exceptions (Downie and Palmer 1992; Wolfe et al. 1992). Codon use by chloroplast genes is biased toward a high representation of NNT and NNA codons and appears to reflect a mutational bias (Wolfe and Sharp 1988; Wolfe et al. 1992; Morton 1993). An exception to this pattern of codon use is the psbA gene. In Marchantia polymorpha the psbA gene has a bias toward, or higher representation of, the NNC codons in twofold-degenerate amino acids, the threefold-degenerate isoleucine, and the twofold-de-
generate group of serine (Umesono et al. 1988; Morton 1993). For each of these amino acids the only tRNA coded by the chloroplast genome is complementary to the NNC codon. Since psbA is the major translation product of the chloroplast (Mullet and Klein 1987), it has been proposed that the codon use of psbA is a result of selection for increased translation efficiency (Morton 1993) similar to the observation in unicellular organisms. This codon use is apparent in higher-plant psbA genes to a lesser extent and is very similar to the codon use followed by chloroplast genes of the green alga Chlamydomonas reinhardtii (Morton 1993).

Although all other chloroplast genes appear to have a codon use dominated by a mutation bias, the codon use by psbA suggests that selection affects codon use of chloroplast genes and may act on more than just a single locus. As in E. coli and Saccharomyces cerevisiae, the intensity of this selection would vary among genes, depending on expression level. To test this, the psbA gene was used to estimate codon fitness for chloroplast genomes. The codon adaptation index (CAI) uses estimates of codon fitness to measure the adaptedness of the codon use of a given gene (Sharp and Li 1987a). Estimating codon fitness by the relative synonymous codon use of highly expressed genes separately in E. coli and Saccharomyces cerevisiae, Sharp and Li (1987a) calculated the CAI of a large number of genes and showed that higher CAI values corresponded to higher expression levels. Sharp and Li (1987b) also demonstrated that the CAI of E. coli genes is inversely related to the rate of synonymous substitutions, evidence that there is selection on codon use.

This study examines M. polymorpha and Nicotiana tabacum chloroplast genes for evidence of similar selective forces, with regard to the suggestion that psbA codon use is a result of selection. Wolfe and Sharp (1988) measured rates of divergence, for chloroplast genes, between M. polymorpha and N. tabacum and noted that there was no correlation with codon bias. The method of measuring codon bias, however, was not defined. Degree of codon bias measured simply by deviation from uniform use of synonymous codons may not effectively measure codon selection. In chloroplast genes, it is the shift in codon preference that selection appears to produce (Morton 1993); a measure of strict bias may not detect such an effect. A better estimate would be to measure adaptation by CAI to a specific codon use. The assumption is made here that the codon use of psbA, as the most highly expressed gene and one with a noticeably different codon use, represents the pattern favored by natural selection, and its relative codon use can be used to estimate fitness. If selection is acting on codon use of chloroplast genes, highly expressed genes should have greater CAI values and the rate of synonymous substitution should be correlated with CAI as observed in other studies. The CAI values and rates of divergence were measured for a set of longer chloroplast genes to determine whether selection is acting on more than just the psbA locus. The data are consistent with selection acting on codon use of M. polymorpha chloroplast genes.

**Material and Methods**

All gene sequences were extracted from the complete chloroplast genome sequence for both Marchantia polymorpha (Ohyama et al. 1986) and Nicotiana tabacum (Shinozaki et al. 1986). The genes psbA, psbB, petA, rbcL, atpB, psbC, and psbD from Chlamydomonas reinhardtii were taken from GenBank release 72 by using the GCG package (Devereux et al. 1984). Sequences were aligned using the BestFit program of the GCG package. The numbers of synonymous and nonsynonymous substitutions per site were estimated using the method of Li et al. (1985).

The CAI (Sharp and Li 1987a) is an estimate of adaptedness of a gene when codons are assigned a fitness. In studies on Escherichia coli, codon fitnesses are estimated by the relative synonymous codon use of codons from highly expressed genes (Sharp and Li 1987a). The relative synonymous use of a codon is represented by \( R_{ij} \) as calculated by equation (1).

\[
R_{ij} = \frac{n_{ij}}{n_{max}}.
\]

In equation (1), \( n_{ij} \) is the number of occurrences of codon \( j \) of amino acid \( i \) in the gene, and \( n_{max} \) is the number of occurrences of the most highly represented codon of amino acid \( i \). Codon fitnesses were estimated separately for each genome by taking the relative synonymous codon use of a codon from the psbA gene. As suggested by Sharp and Li (1987a), codons not used by the psbA gene were assigned 0.5 occurrences in the gene, for purposes of fitness estimates. The amino acid lysine is not coded by the C. reinhardtii psbA gene, so equal fitness estimates of 1.0 were assigned to the two synonymous codons. The CAI of a gene of length \( L \), excluding methionine and tryptophan residues, was then calculated by equation (2), using the fitness estimates \( R_{ij} \) from the psbA gene of the same genome.

\[
\text{CAI} = \exp \left[ \frac{1}{L} \sum_{i=1}^{18} \sum_{j=1}^{18} n_{ij} \ln(R_{ij}) \right].
\]

The degeneracy of amino acid \( i \) is represented by \( n_i \). The codon bias index (CBI) (Morton 1993) is a measure of deviation from uniform use of synonymous
codons, for a given codon use table, and is calculated by equation (3).

\[
\text{CBI} = \frac{1}{L-18} \left( \sum_{i=1}^{R} X_i \left( \sum_{j=1}^{n_i} (1-R_{ij})^2 \right) \right).
\]

The number of occurrences of amino acid \(i\) is represented by \(X_i\). Calculations were limited to genes >650 nt in length. Although the CAI is not dependent on length, shorter genes are subject to sampling problems (Sharp and Li 1987a). To minimize this, the 650-nt cutoff was chosen for comparisons of sequence divergence and codon adaptation. CAI values for chloroplast genes <650 bp in length were calculated, and the averages are presented in table 1. All rate estimates and CAI and CBI values were calculated by programs written by the author using THINK Pascal for the Macintosh.

**Results**

The chloroplast genes used in this study are listed in table 1, along with the CAI values for both *Marchantia polymorpha* and *Nicotiana tabacum*. The genes are ranked in decreasing order of the *M. polymorpha* CAI values. The CAI values for *N. tabacum* genes were also calculated using *M. polymorpha* psbA codon use for fitness estimates as shown in table 1. The CAI value of *Chlamydomonas reinhardtii* genes for which sequence data were available are listed in table 1 as well. In all cases the CAI of a given gene was calculated using codon fitness estimates from the psbA gene of the same genome.

In figure 1 the CAI values from *M. polymorpha* and *N. tabacum* are plotted against one another. The CAI values from the two genomes are correlated \((r = 0.87)\). Because of the method of calculation, the psbA results are expected to influence this somewhat. However, there is still a correlation of CAI values without psbA \((r = 0.61)\). In figure 2 the CBI values for the same genes are plotted against one another and show no correlation.

In figure 3 the rate of nonsynonymous substitution between *M. polymorpha* and *C. reinhardtii* genes is shown plotted against the CAI values for both the *M. polymorpha* gene and the *C. reinhardtii* gene. Both CAI values are negatively correlated with the rate of nonsynonymous substitution \((r = 0.77\) for *C. reinhardtii*; \(r = 0.57\) for *M. polymorpha*). The genes used are those in table 1 that have a CAI value calculated for *C. reinhardtii*. The first 30 amino acids of the peta gene were excluded, as there are only two conserved amino acids.

**Table 1**

**CAIs of Chloroplast Genes**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Marchantia polymorpha</th>
<th>Nicotiana tabacum</th>
<th>N. tabacum</th>
<th>Chlamydomonas reinhardtii</th>
</tr>
</thead>
<tbody>
<tr>
<td>psbA</td>
<td>0.764</td>
<td>0.704</td>
<td>0.525</td>
<td>0.797</td>
</tr>
<tr>
<td>rbcL</td>
<td>0.572</td>
<td>0.578</td>
<td>0.425</td>
<td>0.725</td>
</tr>
<tr>
<td>atpA</td>
<td>0.563</td>
<td>0.532</td>
<td>0.355</td>
<td>...</td>
</tr>
<tr>
<td>psaA</td>
<td>0.529</td>
<td>0.492</td>
<td>0.333</td>
<td>...</td>
</tr>
<tr>
<td>psbC</td>
<td>0.506</td>
<td>0.524</td>
<td>0.348</td>
<td>0.640</td>
</tr>
<tr>
<td>psaB</td>
<td>0.499</td>
<td>0.518</td>
<td>0.372</td>
<td>...</td>
</tr>
<tr>
<td>atpB</td>
<td>0.497</td>
<td>0.500</td>
<td>0.345</td>
<td>0.606</td>
</tr>
<tr>
<td>psbD</td>
<td>0.493</td>
<td>0.548</td>
<td>0.388</td>
<td>0.645</td>
</tr>
<tr>
<td>psbB</td>
<td>0.489</td>
<td>0.476</td>
<td>0.336</td>
<td>0.577</td>
</tr>
<tr>
<td>rpoB</td>
<td>0.488</td>
<td>0.488</td>
<td>0.327</td>
<td>...</td>
</tr>
<tr>
<td>atpI</td>
<td>0.484</td>
<td>0.518</td>
<td>0.371</td>
<td>...</td>
</tr>
<tr>
<td>rps2</td>
<td>0.474</td>
<td>0.502</td>
<td>0.334</td>
<td>...</td>
</tr>
<tr>
<td>petA</td>
<td>0.472</td>
<td>0.480</td>
<td>0.322</td>
<td>0.560</td>
</tr>
<tr>
<td>rpoA</td>
<td>0.469</td>
<td>0.522</td>
<td>0.341</td>
<td>...</td>
</tr>
<tr>
<td>rpoC1</td>
<td>0.465</td>
<td>0.497</td>
<td>0.337</td>
<td>...</td>
</tr>
<tr>
<td>ndhA</td>
<td>0.461</td>
<td>0.540</td>
<td>0.375</td>
<td>...</td>
</tr>
<tr>
<td>ndhF</td>
<td>0.446</td>
<td>0.515</td>
<td>0.352</td>
<td>...</td>
</tr>
<tr>
<td>psbG</td>
<td>0.437</td>
<td>0.478</td>
<td>0.317</td>
<td>...</td>
</tr>
<tr>
<td>ndhB</td>
<td>0.428</td>
<td>0.521</td>
<td>0.336</td>
<td>...</td>
</tr>
<tr>
<td>ndhD</td>
<td>0.413</td>
<td>0.467</td>
<td>0.316</td>
<td>...</td>
</tr>
<tr>
<td>rps3</td>
<td>0.406</td>
<td>0.506</td>
<td>0.333</td>
<td>...</td>
</tr>
<tr>
<td>rpoF</td>
<td>0.374</td>
<td>0.446</td>
<td>0.282</td>
<td>...</td>
</tr>
<tr>
<td>Others</td>
<td>0.435 (0.053)</td>
<td>0.504 (0.038)</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

*Calculated using codon fitness estimates from the psbA gene of *M. polymorpha.*  
*Average CAI for genes <650 bp in length (no. in parentheses is variance).*
other than the initial methionine in this stretch. The synonymous rate also shows a significant negative correlation with the CAI values of both the *M. polymorpha* gene ($r = -0.76$) and the *C. reinhardtii* gene ($r = -0.54$), although the values are above saturation and have large stochastic errors (data not shown).

In figure 4, nonsynonymous rate estimates between homologous genes from *N. tabacum* and *M. polymorpha*, with the exception of *psbG* and *rpoC2*, are plotted against the CAI value of the *M. polymorpha* gene. The *psbG* and *rpoC2* genes were excluded because of alignment difficulties. As with the *M. polymorpha* and *C. reinhardtii* divergence, there is a negative correlation between the values ($r = -0.62$). In figure 5, synonymous rate estimates are shown plotted against the CAI values of the *M. polymorpha* gene and show negative correlation ($r = -0.54$).

### Discussion

Two lines of evidence, when considered together, support selection acting on synonymous codon use of *Marchantia polymorpha* chloroplast genes. The first is
that highly expressed genes have greater CAI values, on the basis of psbA codon fitness estimates. The second is that nucleotide substitution rates are inversely related to CAI. Both of these observations have been made in unicellular organisms (Sharp and Li 1987a, 1987b; Sharp 1991) and Drosophila (Shields et al. 1988). Selective constraints on synonymous substitutions, for increased translation efficiency, would explain these observations. There is no evidence that the mutation rate varies among genes in a single chloroplast genome, although the rate of substitution varies between genes in the inverted-repeat region and those in the single-copy regions (Wolfe and Sharp 1988).

An examination of the CAI values in table 1 suggests that selection is acting on the M. polymorpha chloroplast genes, with varying intensity, depending on translation rate. Rates of translation of chloroplast genes have not been quantified and would differ under many different developmental and physiological conditions. Available data on chloroplast protein levels, however, can be used as an approximation. In Hordeum vulgare chloroplasts grown in light and in chloroplasts recently exposed to light, the two most prominent proteins are the psbA and rbcL products (Klein and Mullet 1986, 1987). The other prominent proteins observed are (a) the α and β subunits of ATPase (atpA and atpB) and (b) major components of photosystems I and II (psaA, psaB, psbB, psbC, and psbD; Mullet 1988; Klein and Mullet 1986, 1987; Klein et al. 1988). These genes in M. polymorpha have the highest CAI values in table 1. If one assumes that relative protein levels are essentially the same in M. polymorpha as in higher plants, then CAI values are correlated with expression levels. Genes <650 bp in length, which are low-expression genes, have a low average CAI value in M. polymorpha. A correspondence between expression level and CAI in Nicotiana tabacum is not so apparent, although rbcL maintains a high CAI value, greater than that in all chloroplast genes except psbA. Moreover, there is less variation in CAI values of N. tabacum genes (table 1), and the shorter-length genes also have a much higher average CAI value in N. tabacum than in M. polymorpha.

Studies on Escherichia coli and Salmonella typhimurium (Sharp and Li 1987b; Sharp 1991) and Drosophila (Shields et al. 1988; Sharp and Li 1989) have demonstrated a correlation between synonymous substitution rates and codon bias, suggesting that selection is acting on codon use. Although sequence is available from only a small number of genes, the rates of both synonymous and nonsynonymous substitution between Chlamydomonas reinhardtii and M. polymorpha show a negative correlation with the CAI values of both the M. polymorpha gene and the C. reinhardtii gene. These relationships can be tested further as more gene sequences become available.

The codon use of the psbA gene from M. polymorpha, while different than that of other M. polymorpha chloroplast genes, is very similar to that of C. reinhardtii chloroplast genes (Morton 1993), suggesting that, if selection is responsible for the codon use pattern of the land plant psbA gene, then selection is acting on the codon use of C. reinhardtii chloroplast genes. This pattern of codon use is potentially confounding to estimates of rate divergence. The method used to estimate the number of synonymous substitutions does not take into account the context of each position. Sequence composition affects the statistical estimate of divergence, but composition is context dependent, and codon bias can affect estimates by altering third-position composition differently in different amino acids (Lewontin 1989). In the case discussed here, the rate difference may reflect to some degree the divergence in codon use, such that similarity in codon use leads to an underestimate of the actual number of events. Measurement of the CAI of M. polymorpha genes will also reflect similarity to psbA codon use and, consequently, to C. reinhardtii codon use, so that the correlation between synonymous rate and CAI may result from this confounding effect. For calculating divergence, Lewontin (1989) gave a method that takes into account the variance of equilibrium codon frequency, as opposed to the variance in equilibrium base frequency. This method, however, is inapplicable here, since, with the exception of psbA, codon use between the compared genes is very different and there is no acceptable way to estimate equilibrium frequencies.
When a codon-use distance between the two species was measured separately for each gene by the method of Long and Gillespie (1991), no correlation was noted between divergence and this distance (data not shown). This indicates that difference in codon use is not solely responsible for the correlation, but the full effect has not been determined.

If selection was acting on chloroplast ancestors in the manner proposed here, then a mutational bias would be degrading the ancestral codon use, with selection countering the mutational bias along the *M. polymorpha* lineage. The intensity of this countering selection would depend on translation level of the gene. Even if selection has not been acting along the *N. tabacum* lineage, the CAI values of *N. tabacum* genes may be correlated with the CAI of the *M. polymorpha*, because of the codon use of their common ancestor. The CAI values for *N. tabacum*, when the *M. polymorpha* *psbA* codon use is used to estimate codon fitness, suggest that selection on *psbA* codon use is less stringent in *N. tabacum* and that the shift in codon use by *psbA* occurs to a lesser extent in *N. tabacum* than in *M. polymorpha*. The overall deviation from uniform codon use as measured by the CBI is not correlated between *M. polymorpha* and *N. tabacum*. Therefore, if the CAI is a historical artifact, then it degrades much more slowly than strict codon bias; but this has not been examined.

As seen in figures 4 and 5, the *M. polymorpha* CAI is also inversely related to both nonsynonymous and synonymous rates of substitution between *M. polymorpha* and *N. tabacum*. A division of genes into two groups, high expression and low expression, shows a difference in average rate of synonymous substitution. When the genes discussed above are defined as high expression, *psbA* down to *psbB* in table 1, the average synonymous rate is 0.918 (standard error [SE] = 0.096), while for the remaining genes the average synonymous rate is 1.016 (SE = 0.112). The correlation between CAI and synonymous rate is weaker than the correlation reported for *E. coli* and *Salmonella typhimurium* (Sharp and Li 1987b). This is probably due to the fact that codon use is still dominated by a mutational bias in land plant chloroplast genes and that selection appears to be limited to a subset of amino acids and to only subtly affect codon use. For example, the amino acid phenylalanine, coded by TTY, has a preference for NNC in M. While this possibility must be explored as more data from close relatives of *M. polymorpha* become available, it holds no explanation for the correlation between synonymous and nonsynonymous rates and also fails to explain why the CAI correlation with expression level exists in *M. polymorpha* but not in *N. tabacum*.

Another explanation is that selection on codon use may affect rates of nonsynonymous substitution as well as synonymous substitution. There is no reason that highly expressed proteins should necessarily be under greater selection at the amino acid level or that selection at the amino acid level should affect codon use, since only relative synonymous codon use is measured. Silent and nonsynonymous replacements have been shown to interact in some vertebrate genes (Lipman and Wilbur 1985), and an effect of amino acid and second-position-nucleotide composition on the rate of synonymous substitution between human and rat sequences has been shown (Ticher and Graur 1989). Related factors may play a role in chloroplast genes, but it must also be considered that selection on codon use can affect amino acid substitutions (Lipman and Wilbur 1985; Sharp 1991). In the twofold-degenerate
amino acids, the psb$\Lambda$ gene of *M. polymorpha* has a bias toward NNC, as discussed above, while the bias of fourfold-degenerate amino acids is for NNT and NNA codons. Only 5 of 165 fourfold-degenerate-group codons are NNC in the *M. polymorpha* psb$\Lambda$ gene. Selection may be more stringent in this gene, for amino acid substitutions, since the substitution could frequently lead to a relatively uncommon codon and, presumably, to a decrease in codon fitness. Selection on codon use may affect nonsynonymous rates in this manner, by affecting the selective value or neutrality of an amino acid change. Such an effect would be dependent on the pattern of codon use and could contribute to the correlations seen in figures 3–5.

Wolfe et al. (1992) note that chloroplast gene codon use appears to reflect a mutational bias rather than selection. This is also suggested by the correlation between genome AT content and overall codon bias (Morton 1993). In the nonphotosynthetic plant *Epifagus virginiana* the chloroplast is vastly reduced in gene content (Wolfe et al. 1992). The chloroplast genome of *Epifagus* is also missing 13 tRNA genes relative to *N. tabacum*, but codon use does not differ from that in *N. tabacum* genes. The *Epifagus* genome, however, is missing all photosynthetic genes, those that have high CAI values. The genes still coded by *Epifagus* may be under very weak selection or no selection for codon use. The correspondence between *N. tabacum* CAI values and expression levels is much weaker than that in *M. polymorpha*, and selection may not be acting along the lineage leading to *N. tabacum* and other flowering plants. Moreover, as discussed above, selection has only a small effect on codon use by even the high-expression genes and, most likely, has no effect on codon use by low-expression genes, those still coded by the *Epifagus* genome.

Although codon use of chloroplast genes appears to be a result of a mutational bias toward a high AT content, the data presented here are consistent with some selection acting on codon use of chloroplast genes of *M. polymorpha*. The intensity of this selection is dependent on the expression level of the gene. Codon use is still dominated by the mutation bias, but this is countered to some degree by selection. The codon use pattern of *C. reinhardtii* chloroplast genes suggests that selection is also acting on this genome, with a stronger effect on all genes, and this is supported by the data shown in table 1; but this will be evaluated as more sequences become available. Since CAI values do not correspond well to expression levels and since the limited correspondence may be a historical artifact, the data set is inconclusive regarding selection on *N. tabacum* chloroplast genes.

Acknowledgments

I would like to thank Michael T. Clegg for generous support and encouragement, as well as Ken Wolfe and Wen-Hsiung Li for many useful suggestions. This work was supported in part by NIH grant GM 45144 to M. T. Clegg.

LITERATURE CITED


Wen-Hsiung Li, reviewing editor

Received June 3, 1993

Accepted October 14, 1993