J Mol Evol (1998) 46:449–459

JOURNAL OF MOLECULAR EVOLUTION

Selection on the Codon Bias of Chloroplast and Cyanelle Genes in Different Plant and Algal Lineages

Brian R. Morton

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

Received: 6 June 1997 / Accepted: 24 July 1997

Abstract. In the plant chloroplast genome the codon usage of the highly expressed *psbA* gene is unique and is adapted to the tRNA population, probably due to selection for translation efficiency. In this study the role of selection on codon usage in each of the fully sequenced chloroplast genomes, in addition to *Chlamydomonas reinhardtii,* is investigated by measuring adaptation to this pattern of codon usage. A method is developed which tests selection on each gene individually by constructing sequences with the same amino acid composition as the gene and randomly assigning codons based on the nucleotide composition of noncoding regions of that genome. The codon bias of the actual gene is then compared to a distribution of random sequences. The data indicate that within the algae selection is strong in *Cyanophora paradoxa,* affecting a majority of genes, of intermediate intensity in *Odontella sinensis,* and weaker in *Porphyra purpurea* and *Euglena gracilis.* In the plants, selection is found to be quite weak in *Pinus thunbergii* and the angiosperms but there is evidence that an intermediate level of selection exists in the liverwort *Marchantia polymorpha.* The role of selection is then further investigated in two comparative studies. It is shown that average relative codon bias is correlated with expression level and that, despite saturation levels of substitution, there is a strong correlation among the algae genomes in the degree of codon bias of homologous genes. All of these data indicate that selection for translation efficiency plays a significant role in determining the codon bias of chloroplast genes but that it acts with different intensities in different lineages. In general it is stronger in the algae than the higher plants, but within the algae *Euglena* is found to have several unusual features which are noted. The factors that might be responsible for this variation in intensity among the various genomes are discussed.

Key words: Chloroplast DNA — Codon usage — Natural selection

Introduction

Two underlying factors, with different relative importances in different species, contribute to codon bias. The first is the genome composition bias which generates a bias in degenerate positions of coding sequences (Bernardi and Bernardi 1986) and the second is selection on coding sequences for specific codons, most likely to increase translation efficiency (Ikemura 1985; Sharp and Li 1987a). Evidence for this type of selection comes predominantly from *Escherichia coli* where codon usage is adapted to the tRNA population in the cell (Ikemura 1985) and the degree of codon bias varies among genes, correlated with expression level (Sharp and Li 1987a).

In the chloroplast genome of plants, composition bias appears to be the predominant factor influencing codon bias. Degenerate positions are strongly biased toward a high A+T content, which matches the composition bias of noncoding regions (Morton 1993). The exception that has been noted is the *psbA* gene, which has a high C content at the third position of specific synonymous groups (Morton 1996). The *psbA* codon usage more

e-mail: bmorton@barnard.columbia.edu

closely matches the tRNA population of the chloroplast, an observation which, together with the fact that *psbA* is the most highly translated chloroplast gene, suggests that selection for translational efficiency is responsible (Morton 1993, 1996). The pattern of codon usage of the plant *psbA* is also observed in the chloroplast genes of the green alga *Chlamydomonas reinhardtii,* suggesting that selection is much more prevalent in this organism (Morton 1996). Therefore, among chloroplast genes studied to date, two basic patterns of codon usage have been observed which will be referred to here as the composition pattern and the selection pattern.

Although most plant chloroplast genes appear to have a composition pattern of codon usage, there is some evidence for weak selection. The codon usages of the highly expressed photosynthetic genes, most noticeably *rbcL,* are more similar to *psbA* than are other chloroplast genes, particularly in *Marchantia polymorpha.* Although degenerate positions in the highly expressed genes have a very high A+T content, there is actually a gradient in similarity to the selection pattern, an observation which is consistent with the model of selection based on translation efficiency (Morton 1994). This suggests that even though composition bias is predominant, selection on codon usage might not be completely absent. Therefore, selection is thought to act strongly on the codon usage of *psbA* such that it has a noticeably unique codon usage pattern, and at a very weak intensity on the codon usage of some other highly expressed genes of the plant chloroplast genome (Morton 1994).

Recent observations have forced a reconsideration of this model. Within the flowering plants, a significantly increased rate of silent substitution is observed at the *psbA* locus specifically in those synonymous groups in which it has an atypical codon usage, an observation that is not consistent with constraining selection (Morton 1997). This has raised the possibility that the unusual codon usage of the flowering plant *psbA* gene may actually be the remnant of an ancestral bias that is currently being degraded as a result of a recent loss in selective constraints. This new model is supported by a comparative analysis of codon usage (Morton and Levin 1997), and if it is correct, the codon bias gradient observed among plant chloroplast genes would be an even fainter remnant. What this model implies is a recent change in selective pressure on codon usage of plant chloroplast genes, raising general questions about how and why selection on codon usage has changed during the evolution of chloroplast DNA.

Given the possible recent shift in the role of selection on plant chloroplast genes, we are interested in understanding the evolutionary dynamics of codon usage in chloroplast genes across a wide taxonomic range. In the current study, this is examined using the fully sequenced chloroplast genomes: *C. paradoxa* (Stirewalt et al. 1995), the red alga *P. purpurea* (Reith and Munholland 1995), the diatom *O. sinensis* (Kowallik et al. 1995), the green alga *E. gracilis* (Hallick et al. 1993), and the gymnosperm *P. thunbergii* (Wakasugi et al. 1994), in addition to the plants used in previous analyses (Morton 1994). Because of its close relationship to rice, which is used in this study, the *Zea mays* genome was excluded. Chloroplast genes from *C. reinhardtii* were also added because of previous work on this species (Morton 1996). Ignoring secondary endosymbiotic events, evidence supports a monophyletic origin for all of these plastid genomes (Reith 1995). Most importantly, the tRNA gene content is essentially identical in each of these genomes and every genome has a high A+T content. Therefore, composition bias is very similar and adaptation to the tRNA population should result in similar codon biases among all of the genomes.

The patterns of codon usage, as well as the degree of bias that exist in the different lineages, are examined to test whether or not there is evidence for selection and to compare relative intensities. The results indicate that selection favors the same pattern of codon usage in each lineage and appears to have at least some role in all chloroplasts, but that the intensity varies widely. In general, it appears to be strongest in *Chlamydomonas* as well as in the cyanelle of *Cyanophora,* of intermediate strength in *Odontella, Porphyra,* and *Marchatia* but very weak in the flowering plants and *Pinus,* essentially being limited in these species to *psbA* and *rbcL.* The last observation is consistent with a recent loss of selection that has been proposed (Morton 1997) since the tests used here cannot exclude remnants of an ancestral bias. The other species examined, *Euglena,* is found to be a very interesting case as there is evidence for selection on at least some genes from all tests. However, there are no genes encoded in its genome that have a codon usage with an apparent selection pattern. The case of *Euglena,* as well as factors that could affect selection intensity, such as effective population size, genome copy number, and reliance on photosynthesis, are discussed with regard to the evolution of codon bias across chloroplast lineages.

Materials and Methods

All complete chloroplast genome sequences and the cyanelle genome of *Cyanophora* were used in this study and are listed in Table 1. To avoid constant reiteration, all references to chloroplast genomes from here on will include the *Cyanophora* cyanelle. Using the information in the GenBank file, all protein coding, ORF, and ycf (conserved open reading frames, Hallick and Bairoch 1994) sequences greater than 350 nucleotides in length were extracted directly from the genome sequence. All available genes greater than 350 nucleotides in length from the chloroplast of *Chlamydomonas* were extracted directly from GenBank. The codon usage patterns of the *psbA* genes from each genome and the *rbcL* genes from higher plants were compared by a UPGMA cluster analysis using distances calculated by the method of Long and Gillespie (1991). This compares genes solely on the basis of similarity in codon usage.

Table 1. Complete genome sequences used in this study

Organism	Classification	GenBank accession	A + T content[a]
Cyanophora paradoxa	Glaucocystophyceae	U30821	80.9
Odontella sinensis	Chromophyte	Z67753	80.7
Porphyra purpurea	Rhodophyte	U38804	77.9
Euglena gracilis	Chlorophyte	X70810	82.4
Marchantia polymorpha	Bryophyte	X04465	84.3
Pinus thunbergii	Gymnosperm	D17510	66.2
Nicotiana tabacum	Angiosperm	Z00044	69.6
Oryza sativa	Angiosperm	X15901	67.8

[a]A + T content of noncoding regions

To compare codon bias, the Codon Adaptation Index (CAI, Sharp and Li 1987b) was used. The CAI is a measure of adaptation to a specified pattern of codon bias, and in this case we are interested in adaptation to the selection pattern. Therefore, two rounds of calculation were performed to ensure that all genes were measured against the same reference and that this reference was strongly biased to the selection pattern. Initially, since *psbA* is known to be the primary translation product of the chloroplast (Mullet and Klein 1987), the CAI value for every gene was calculated relative to the *psbA* gene of the same genome. The codon usages of genes that had a CAI value of 0.75 or greater were then combined. This cumulative codon usage was then used as a reference to calculate a new CAI for every gene. This second set of CAI values was used in all subsequent work.

A simulation approach was used to test selection on individual genes. For every species the noncoding regions were combined and then used to calculate nucleotide frequencies. Following this, for each gene within the genome 100 random codon usage tables were generated. Each table was generated by setting the number of occurrences of each twofold, threefold, and fourfold synonymous group (with sixfold degenerate amino acids divided into a twofold and a fourfold degenerate group) equal to the number observed in the coding sequence. For each occurrence, one of the synonymous codons was assigned at random based on the relative frequencies of the alternative nucleotides at the third position of that synonymous group. Once all codons were assigned, the CAI value of that codon usage table was calculated, generating a distribution of expected CAI from the 100 replicates for that gene. Significant deviation from the expectation could then be determined by comparing the CAI value of the actual gene to the random distribution. Given the observations of the influence of a genomic signature on codon choice (Karlin and Mrazek 1996) the procedure was repeated using dinucleotide compositions from noncoding regions and then assigning codons randomly as a function of the composition of the second codon position.

For each gene a composition statistic, the C content at the third position of the synonymous groups AAY, CAY, GAY, TAY, and TTY, was also calculated. This statistic is based on the atypical codon bias of the plant *psbA* gene and the *Chlamydomonas* chloroplast genes. In these genes, the NNY synonymous groups, those listed above, have a very high C content at the third codon position despite a strong bias toward A and T in noncoding regions and in other synonymous groups (Morton 1993).

Results and Discussion

Codon Usage Patterns in the Different Chloroplast Genomes

Among flowering plant chloroplast genes, two basic patterns of codon usage are observed. One of these is unique to the *psbA* gene and is the pattern that is thought to result from selection (Morton 1993, 1996). The most distinctive difference between the two patterns of codon usage is that selection favors C at the third position of NNY synonymous groups while the composition bias is toward T. A second, less noticeable, difference between the two patterns is that selection favors T at the third position of fourfold degenerate groups so that, even though the composition bias results in a high T content, it is increased further by selection (Morton 1996). Both of these features appear to be an adaptation to match the 31 tRNA genes of the chloroplast that are available for translation (Morton 1996).

In order to study selection over a broader taxonomic range we have to determine if the same selection pattern exists in the chloroplast genomes of other species. The composition bias and the tRNA content of the different chloroplast genomes in the current study are both quite consistent. All genomes have a high A+T content in noncoding regions, although in the algae it tends to be slightly greater (see Table 1). The same 31 tRNA genes are encoded by each of the genomes except *Nicotiana,* which is lacking two tRNAs, complementary to the CCC and CGG codons, that are present in the other genomes. This similar composition bias and tRNA content leads to the expectation that selection on codon usage should favor the same set of codons in each genome, but this has to be established. Since *psbA* is the prominent translation product of the chloroplast genome, if there is any selection at all for translation efficiency, the codon usage of *psbA* will show the pattern favored by selection in each genome. Therefore, we want to determine if the *psbA* genes from the different genomes have the same pattern of codon usage.

A comparison shows that the selection pattern from flowering plants is observed in the *psbA* gene of all other genomes except *Euglena.* When the C content of the NNY groups is calculated, it is found to be very high except in *Euglena* (Table 2). In contrast, the C content at the third codon position of fourfold degenerate groups is extremely low; these codon groups have a strong bias toward T. The *rbcL* genes from plants, which are examples of genes which have a codon usage dominated by composition bias, have a low C content in all synonymous groups (Table 2). In general, for each synonymous group the favored codon or codons are the same in all *psbA* genes, excluding *psbA* from *Euglena* (data not shown). In *Euglena,* all codons with A or T at the third position are predominant. A cluster analysis of the genes in Table 2 also shows the similarity of all *psbA* genes, except *Euglena,* in terms of codon usage (Fig. 1). The existence of the same selection pattern in each genome means that we can compare adaptation to a specific codon usage pattern in all of the different lineages. The exception of *Euglena psbA* suggests that selection is very weak or completely absent from this genome. The codon

Table 2. Third position composition of *psbA* and plant *rbcL* genes

Organism and gene	Four fold %C[a]	Four fold %C[b]	NNY %C[c]
All *psbA* genes			
Cyanophora psbA	1.6	2.2	73.4
Chlamydomonas psbA	0.9	1.3	93.8
Odontella psbA	1.6	2.4	87.7
Porphyra psbA	3.9	6.2	69.7
Marchantia psbA	2.4	3.0	66.2
Pinus psbA	17.4	23.3	64.9
Nicotiana psbA	10.7	14.9	62.3
Oryza psbA	9.1	12.5	57.1
Euglena psbA	0	0	18.8
Plant *rbcL* genes			
Marchantia rbcL	1.2	1.8	25.0
Pinus rbcL	10.9	17.6	30.5
Nicotiana rbcL	10.1	17.0	33.7
Oryza rbcL	11.4	19.8	33.3

[a] Percent C at third position of four fold degenerate codon groups
[b] C/(C + T) at the third position of four fold degenerate groups
[c] Percent C at the third positionof the CAY, GAY, TAY, TTY, and AAY synonymous groups

usage of all genes from *Euglena* is dominated by a very strong bias toward A and T at degenerate positions, indicative of a simple composition bias.

Testing Selection on Codon Bias by Random Gene Construction

As a first step, selection on each individual gene was tested by generating an expected distribution of CAI based on the genome nucleotide composition and the amino acid composition of that gene as described in Materials and Methods. A CAI value for the actual gene that is 2 or more standard deviations above the mean of the distribution indicates that the codon bias is significantly greater than expected based on composition bias alone and, therefore, is considered evidence for selection. The advantage of this method is that it tests significant deviation toward the selection pattern described above, not simply significant deviation from random codon usage.

Representative results of the stimulation are given in Table 3. A number of the genes that are coded by most of the genomes are listed and those that have a CAI that is at least 2 standard deviations above the mean of the randomly generated distribution are indicated. A summary for each genome is shown in Table 4. The simulation which took into account dinucleotide frequencies from noncoding regions gave essentially identical results in terms of Table 4 (data not shown).

In the simulation study, 75% of all *Cyanophora* genes were found to have a significant CAI value, and the deviations from expectation tend to be very large (Table 3). Therefore, it appears that selection is a significant factor affecting codon usage in this genome. In *Odontella* and *Porphyra,* there is evidence for selection on a number of genes, suggesting that it is an important factor but weaker in general than in *Cyanophora.* Conversely, few genes in *Euglena* have a CAI significantly higher than the expectation, and those that do tend to be much closer to the mean than in the other algae. Therefore, selection on codon usage appears to be much weaker in *Euglena* than the other algae, which is consistent with the observation made previously based on the third position composition (Table 2). However, it is interesting that a few genes from *Euglena* do show evidence for selection despite the strong A+T bias at the third position.

The data from plant genes in Table 3 indicate a lower intensity of selection in general. In *Pinus* and the angiosperm, very few genes show a significantly higher CAI than expected under the given compositions and those that do tend to be barely significant unlike in the algae. Half of all *Marchantia* genes, however, have significant CAI values (Table 4), suggesting that selection is more important in *Marchantia* than other higher plant genomes.

Although this method gives strong evidence for selection on specific genes, a drawback is that it cannot give absolute answers concerning relative selective pressure in different lineages. A significant result for a gene in one genome, but a nonsignificant result for the homologous gene in another genome, does not necessarily indicate a difference in selection intensity. Even if selective requirements are the same, a different composition bias in the two genomes could mean that an adequate level codon bias is generated by composition bias alone in one genome but not the other. This is likely to be a minor problem in our analysis, since composition bias is so similar in the different genomes, but it does mean that we have to test variation in selection intensity more rigorously.

Levels of Codon Bias in the Different Chloroplast Genomes

Variation in the degree to which different genomes are adapted to the selection pattern provides evidence concerning relative selection intensity in the different species. In this case we cannot generate direct evidence for selection, as in the first analysis, but relative CAI values can be used to examine variation in intensity across lineages more thoroughly than in the simulation analysis.

The CAI values for various genes from each species, including *rbcL* and *psbA* from *Chlamydomonas,* are given in Table 5. In general, the algae have a greater bias in codon usage than do the plants, indicating stronger selection, which supports what the simulation analysis indicated. Within the algae, *Cyanophora, Chlamydomonas,* and *Odontella* tend to have a stronger codon bias than either *Porphyra* or *Euglena.* In addition, variation within both *Cyanophora* and *Odontella* is greater and they have numerous genes with a CAI value greater than

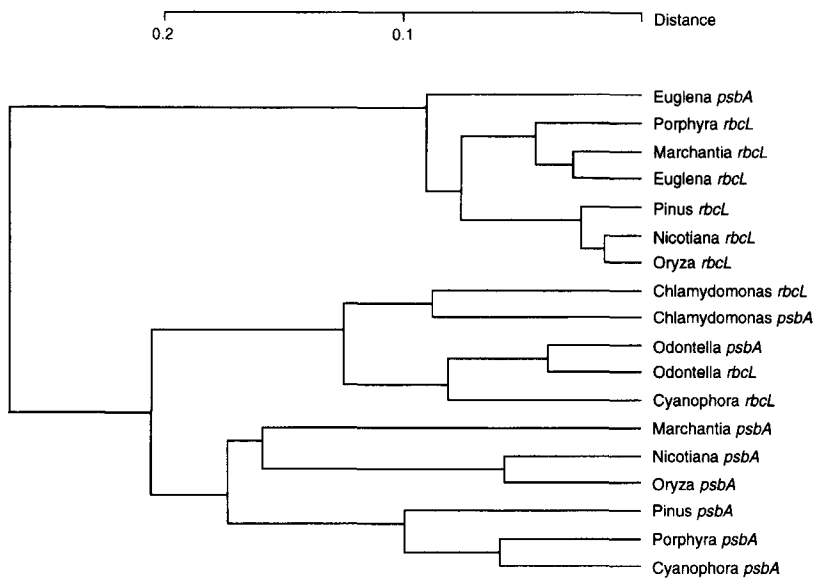

Fig. 1. Cluster of genes based on similarity of codon usage as measured by the method of Long and Gillespie (1991). The two major groups are those genes with the composition pattern (top) and those with the selection pattern (bottom).

0.5, a value exceeded in *Euglena* and the higher plants only by the *psbA* and *rbcL* genes of *Marchantia.*

The results in Table 5, supported by Tables 3 and 4, suggest a basic ranking by selection intensity within the algae of *Cyanophora* followed by *Odontella, Porphyra,* and then *Euglena.* Based on the high CAI values from available genes (Table 7), *Chlamydomonas* appears to be under strong selection, similar to *Cyanophora.* Within the higher plants, which in general have lower selection intensity, *Marchantia* has the strongest selection, followed by *Pinus* and then the flowering plants. This ranking of genomes by codon bias and the greater CAI values for algae is not changed if we use only the plant *psbA* genes as the basis for CAI calculation (data not shown), demonstrating that algae genes are in general more strongly adapted to the selection pattern.

The variation in adaptation to the selection pattern is also apparent from a comparison of C content at the third position of the NNY synonymous groups in different genes and species (Table 6). Despite the strong A+T bias in noncoding regions of each genome (Table 1), certain genes have a strong bias to C in the NNY groups. In particular, *psbA* and *rbcL* tend to have high C content relative to other genes in the same genome. Also, the exception of *Euglena* in terms of this wide variation in C content is noticeable. This wide variation in C content is very strong evidence for selection on codon usage since we otherwise have to invoke a locus-specific mutation bias that is limited to specific codon groups (Morton 1996).

Codon Bias and Gene Expression

The simulation analysis provides very strong evidence, supported by the relative codon biases of different genomes, that selection acts on the codon bias of many chloroplast genes. One of the fundamental observations concerning selection on codon usage in *E. coli* is that codon bias is correlated with expression level (Sharp and Li 1987a). Since the selection pattern of codon usage in chloroplast genes is an adaptation to the tRNA population and is thought to be based on translation efficiency, we should observe a gradient among chloroplast genes in terms of degree of bias toward this pattern, and this should be correlated with expression level.

The ranking in Table 5, in which the highly expressed *psbA* and *rbcL* genes consistently have high CAI values, is suggestive that such a correlation exists, but a better test is required. To accomplish this, the 85 genes that are coded by at least two genomes were ranked by average relative CAI. The coding sequences from each genome were ranked by CAI from greatest to smallest, and the placement for each gene was then calculated by dividing the ranking by the number of protein coding genes in that genome. The overall ranking of each gene was then found by determining the average placement in the genomes which encode that gene (Table 8) with a low average ranking indicating that a gene tends to have a high relative CAI. All of the genomes were included in this study despite the evidence that selection does not act with equal intensity. The assumption is that any ranking arising from selection will not be masked by any random placement of the genes from genomes under weak selection. When this was tested by excluding the flowering plants, the ranking remained essentially identical.

When the average ranking by CAI is compared to what we know about expression level of chloroplast genes, the expected relationship is apparent. Various studies of chloroplast proteins have all indicated the same general relative protein levels, which are likely to reflect to a large degree the relative translation levels. The major photosynthetic proteins are present in the largest amounts in the chloroplast and are found in much higher concentrations than other proteins. The *psbA* gene

Table 3. Results of CAI simulation analysis[a]

Gene	Organism[b]							
	Cpa	Osi	Ppu	Egr	Mpo	Pth	Nta	Osa
psbA	16.1	25.7	17.8	5.8	17.9	9.0	10.9	10.6
—B	11.9	8.2	3.9	3.7	3.8	3.1	***	***
—C	14.4	9.7	4.9	4.4	5.8	2.3	***	***
—D	9.5	8.4	4.9	***	5.0	3.8	2.2	***
rbcL	22.1	21.5	9.4	7.4	9.9	4.5	4.9	4.5
—S	4.7	9.2	3.5					
tufA	12.7	6.4	4.8	***				
psaA	13.6	9.2	6.3	2.7	6.3	3.0	***	2.0
—B	16.3	9.4	4.4	***	4.8	2.3	***	***
atpA	14.0	8.3	3.6	***	8.9	***	***	***
—B	9.5	8.9	5.6	***	3.6	2.7	***	2.0
—D	***	***	2.0					
—E	3.9	4.0	2.2			***	***	***
—F	3.3	2.4	***	***	***		***	***
—I		***	***	***	2.3	2.2	2.1	***
petA	7.2	3.7	***		2.7	***	***	***
—B	4.9	3.4	4.3	***		2.6	***	***
—D	6.8	3.3	3.3			***	***	***
rpoA	2.4	***	***		***	***	***	***
—B	5.7	***	2.3	***	2.3	***	***	***
—C1	4.1	3.4	***	***	2.8	***		***
—C2	5.6	2.3	3.1	***	***	***		***
rps2	4.3	***	2.0	***	2.4	***	***	***
—3	5.0	***	***	***	***	***	***	***
—4	3.7	***	***	***	***	***	***	***
—7	3.1	2.2	2.9	***	***	***		***
—8	2.8	***	***	***		***	***	***
—9	3.4	2.4	2.3	***				
—11	2.5	***	***	***	***	***	***	2.1
—12	3.0	***	2.8	***		***		
rp12	4.1	2.2	***	***		***		***
—5	5.1	***	***	***				
—14	***	***	***	***		***	***	***
—16	3.3	***	***	***	***		***	***

[a] Number of standard deviations above the mean of the random distribution is given for the CAI value of each gene. Those genes that are less than 2 are designated by ***. If a gene is not coded by a particular genome it is left blank

[b] Organisms are abbreviated by first letter of genus and first two letters of species (see Table 1)

product is known to be the most prominent translation product due to a rapid turnover resulting from oxidative damage (Mullet and Klein 1987). Other prominent proteins in plants are coded by the *rbcL, psaA* and *psaB, atpA* and *atpB* genes. Additionally, the PSII proteins besides the *psbA* product, which are coded by *psbB, psbC,* and *psbD,* are also highly expressed (Klein and Mullet 1986; Mullet and Klein 1987). All of these genes rank very highly in terms of codon bias. Another high-ranking gene is *rbcS,* coded only in some algae, and this ranking is consistent with the proposed correlation since this gene codes for the small subunit of Rubisco, the most abundant chloroplast protein. Genes that code for subunits of the cytochrome b6/f complex (designated *pet*) are also highly ranked and, again, these are prominent chloroplast proteins (Klein and Mullet 1986). The *tufA* gene, which codes for the protein translation elongation factor EFTu, also has a high ranking. This is notable since the homolog in *E. coli* has the third highest CAI value in that genome (Lobry and Gautier 1994). Finally, the genes coding for the prominent allophycocyanin (*apc*) and phycocyanin (*cpc*) proteins of *Cyanophora* and *Porphyra* are also very highly ranked.

At the opposite end of the ranking is a variety of genes, primarily ribosomal protein genes, subunits of the chloroplast RNA polymerase, and conserved open reading frames (ycf) of unknown function. None of these low-ranking genes codes for a protein that is known to be present in large amounts in the chloroplast (Klein and Mullet 1986; Mullet and Klein 1987). Therefore, the ranking clearly shows that the highly expressed photosynthetic genes generally have larger CAI values. Further, in the algae and *Marchantia,* the difference in CAI values between high and low expression genes is quite pronounced (Table 5).

The ranking of genes by relative codon bias has two interesting features worth noting. First, as apparent from Table 8, is the general rule that ribosomal protein genes are not highly ranked. In *E. coli* these genes tend to have the high rankings (Lobry and Gautier 1994), probably due to their role in translation. In the chloroplast, though, they are not as highly expressed as the major photosynthetic genes (Klein and Mullet 1986), so the lower ranking is expected. However, the *rpl12* gene, which codes the ribosomal protein L12, is a noticeable exception to this rule. This gene is coded only in *Euglena, Odontella,* and *Porphyra* and in each case has a very high relative CAI value (data not shown). If the ranking seen in Table 8 is in fact indicative of variation in selection on codon usage as a function of translation rate, then *rpl12* should be highly expressed. This remains to be tested, but it is interesting that L12 is one of two proteins of the large subunit that is known to be present in multiple copies in the functional *E. coli* ribosome (Lake 1985)—the other being L7, which is not coded by any of the sequenced chloroplast genomes. Therefore, L12 is potentially a highly expressed gene. The second point concerns the

Table 4. Summary of the simulation analysis

Organism	Total number of genes[b]	CAI > 2 SD over mean[a]	
		Number of genes	Proportion of total
C. paradoxa	93	70	0.75
O. sinensis	75	36	0.48
P. purpurea	136	50	0.37
E. gracilis	36	6	0.17
M. polymorpha	42	21	0.50
P. thunbergii	42	10	0.24
N. tabacum	37	4	0.11
O. sativa	48	4	0.08

[a] Mean is taken from the random gene distribution (see text)

[b] Number of protein coding and ycf genes more than 350 nucleotides in length

Table 5. CAI values of selected genes or groups of genes

Gene(s)[b]	Organism[a]								
	Cpa	Cre	Osi	Ppu	Egr	Mpo	Pth	Nta	Osa
psbA	0.729	0.782	0.830	0.611	0.436	0.634	0.407	0.455	0.443
rbcL	0.704	0.754	0.739	0.478	0.484	0.514	0.334	0.363	0.356
rbcS	0.613	n/a[c]	0.643	0.437	n/a	n/a	n/a	n/a	n/a
tufA	0.609	n/a	0.517	0.432	0.338	n/a	n/a	n/a	n/a
psaA/B	0.537	n/a	0.451	0.360	0.356	0.403	0.281	0.264	0.269
psbB/C/D	0.564	n/a	0.498	0.363	0.390	0.420	0.296	0.285	0.267
atp	0.509	n/a	0.443	0.380	0.361	n/a	0.268	0.273	0.259
petA/B/D	0.544	n/a	0.453	0.392	n/a	0.424	0.291	0.275	0.276
rps	0.437	n/a	0.375	0.329	0.275	0.377	0.269	0.268	0.286
rpl	0.442	n/a	0.385	0.355	0.372	0.326	0.210	0.247	0.236
rpo	0.419	n/a	0.363	0.339	0.324	0.380	0.261	0.260	0.262

[a] Organisms are designated by first letter of genus and first two letters of species

[b] When multiple genes are designated, such as *psaA/B*, the average of CAI of the genes is given. In cases where subunits are not designated, such as *atp*, the average CAI of all genes with the same three letter designation is given

[c] Gene, or more than half of that group, is not coded in the chloroplast of this organism or is not available for *Chlamydomonas* (Cre)

open reading frame ORF180 in *Cyanophora,* which has a CAI of 0.664, one of the highest of all chloroplast genes (Table 5), and much greater than any other ORF; all other *Cyanophora* ORFs are within the range 0.328 to 0.402. The model proposed here predicts that if this ORF codes a functional product, it should be relatively highly expressed.

Testing Selection by the Conservation of Codon Bias Across Lineages

If, as the data strongly suggest, selection is maintaining codon bias as a function of expression level and genes are expressed at roughly the same relative levels in each species, we would expect the relative codon bias of homologous genes to be conserved across lineages. Given the fairly consistent gene content of all chloroplast genomes, it is reasonable to expect that relative expression levels are roughly equal, at least to the degree that the major photosynthetic genes are the most highly expressed in each genome. Therefore, it is possible to test selection by comparing CAI across lineages.

Since *Cyanophora* tends to have the strongest codon bias (Table 5) of all species with a complete genome sequence available, indicating that it is under the strongest selection, it was selected as the basis for the comparison. For the other species, the CAI value of each gene was compared to the CAI of the homologous gene from *Cyanophora,* and in each case a correlation is observed (Fig. 2). Testing the correlations by both Fisher's z transformation and the Kendall Rank test (Sokal and Rohlf 1981) yields the same result, which is that all correlations with the exception of the comparison with

Table 6. C content of NNY synonymous groups for selected genes

Gene(s)[b]	Organism[a]								
	Cpa	Cre	Osi	Ppu	Egr	Mpo	Pth	Nta	Osa
psbA	0.734	0.938	0.877	0.697	0.188	0.662	0.649	0.623	0.571
rbcL	0.638	0.884	0.600	0.327	0.185	0.250	0.305	0.337	0.333
rbcS	0.579	n/a[c]	0.581	0.265	n/a	n/a	n/a	n/a	n/a
tufA	0.310	n/a	0.303	0.200	0.219	n/a	n/a	n/a	n/a
psaA/B	0.354	n/a	0.300	0.274	0.089	0.113	0.281	0.251	0.297
psbB/C/D	0.375	n/a	0.386	0.320	0.141	0.117	0.313	0.293	0.269
atp	0.175	n/a	0.231	0.178	0.139	0.097	0.201	0.253	0.270
petA/B/D	0.223	n/a	0.253	0.318	n/a	n/a	0.229	0.198	0.230
rps	0.071	n/a	0.150	0.170	0.235	0.095	0.209	0.241	0.280
rpl	0.093	n/a	0.157	0.142	0.177	0.089	0.267	0.227	0.346
rpo	0.115	n/a	0.139	0.169	0.106	0.098	0.255	0.186	0.240

[a]Organisms are designated as in Table 5

[b]When multiple genes are designated, such as *psaA/B,* the average C content is given. In cases where subunits are not designated, such as *atp,* the average C content of all genes with the same three-letter designation is given

[c]Gene, or more than half of that group, is not coded in the chloroplast of this organism or is not available for *Chlamydomonas* (Cre)

Table 7. CAI values of *Chlamydomonas* genes

Gene	CAI
psbA	0.782
rbcL	0.754
psbC	0.711
psbD	0.712
psbE	0.678
psbH	0.692
atpA	0.725
atpB	0.723
atpE	0.661
petA	0.725
petB	0.703
petD	0.804
rps7	0.471
rps12	0.618
rpl14	0.578
chlB	0.384
ycf5	0.410
hbpX	0.385

Oryza are significant at the 5% level (Table 9). These correlations in codon bias exist despite saturation levels of divergence at degenerate positions (data not shown). In addition, within the algae, with the exception of *Euglena,* the relative C content of NNY synonymous groups from homologous genes is also significantly correlated (data not shown)—again, despite saturation levels of divergence.

When the comparisons to *Pinus* and the angiosperm are examined it seems clear that the *rbcL* and *psbA* genes, indicated in Fig. 2, stand apart from the other genes, and there appears to be little, if any, correlation among the other genes. Since *rbcL* and *psbA* are the two most highly expressed genes and show highly significant deviations in all species in the simulation analysis (Table 3), it is possible that any correlations to *Pinus* and the angiosperm are generated primarily by these genes. Therefore, the analysis was repeated excluding both genes (Table 9). In this case, for the comparisons to both *Oryza* and with *Pinus* no significant correlation exists, and in *Nicotiana* the correlation is just barely significant at the 5% level. This is consistent with the results from the simulation analysis, which indicate that selection on codon bias is limited to just a few genes in *Pinus* and the flowering plants. The strong correlations in codon bias within the algae, except *Euglena,* and in the comparison to *Marchantia* are also consistent with the simulation results and the relative codon biases (Table 5).

Selection on the Codon Bias in Different Lineages

All of these data can now be considered in an examination of the role of selection in the different species. The correlation of CAI with expression level makes it likely that the selection that was detected by the simulation analysis is to increase translation efficiency. All of the evidence indicates that, in general, selection is much stronger in the algae than in plants. In the algae, selection acts most strongly in *Cyanophora* and, based on Table 7, probably *Chlamydomonas,* with lower intensity in *Odontella* and *Porphyra* and lower still in *Euglena.*

Table 8. Average ranking of chloroplast genes by CAI[a]

Genes 1–22		Genes 23–44		Genes 45–66		Genes 67–85	
Gene	Avg. rank	Gene	Avg. rank	Gene	Avg. rank	Gene	Avg. rank
psbA	0.035	*psbC*	0.306	*rps9*	0.517	*rpoB*	0.708
cpcB	0.054	*rps2*	0.308	*ycf24*	0.522	*rps8*	0.712
rpl12	0.055	*dnaK*	0.320	*ycf37*	0.547	*ycf21*	0.724
cpcA	0.056	*rpl1*	0.355	*rps7*	0.553	*rps11*	0.729
apcB	0.058	*rpl18*	0.366	*rpl3*	0.558	*rpoA*	0.731
rbcL	0.063	*rpl13*	0.368	*rpl19*	0.566	*psaD*	0.773
rbcS	0.066	*ycf35*	0.382	*rpl5*	0.569	*dnaB*	0.784
apcA	0.078	*rpl11*	0.390	*rps4*	0.581	*trpG*	0.799
tufA	0.098	*groEL*	0.397	*secA*	0.581	*chlB*	0.810
petD	0.105	*rps12*	0.400	*atpF*	0.584	*rpl16*	0.814
atpA	0.145	*apcE*	0.403	*rpl6*	0.596	*secY*	0.842
atpB	0.157	*thiG*	0.416	*rps5*	0.611	*psaL*	0.851
apcF	0.198	*atpD*	0.416	*rpoC2*	0.637	*ycf5*	0.853
psbD	0.236	*ycf25*	0.417	*ycf23*	0.639	*rpl4*	0.876
psbB	0.237	*apcD*	0.437	*rpoC1*	0.640	*rpl14*	0.880
petB	0.266	*preA*	0.450	*rps3*	0.641	*rpl2*	0.899
psbV	0.272	*chlN*	0.456	*ycf36*	0.650	*ycf38*	0.900
psaA	0.274	*clpC*	0.463	*ycf3*	0.657	*rps13*	0.958
atpE	0.293	*ycf39*	0.474	*ycf16*	0.664	*ycf4*	0.968
psaB	0.295	*chlI*	0.483	*ycf30*	0.683		
chlL	0.296	*atpG*	0.486	*ycf29*	0.691		
petA	0.297	*psaF*	0.510	*atpI*	0.700		

[a]Average ranking by CAI in the genomes coding for each gene as described in the text

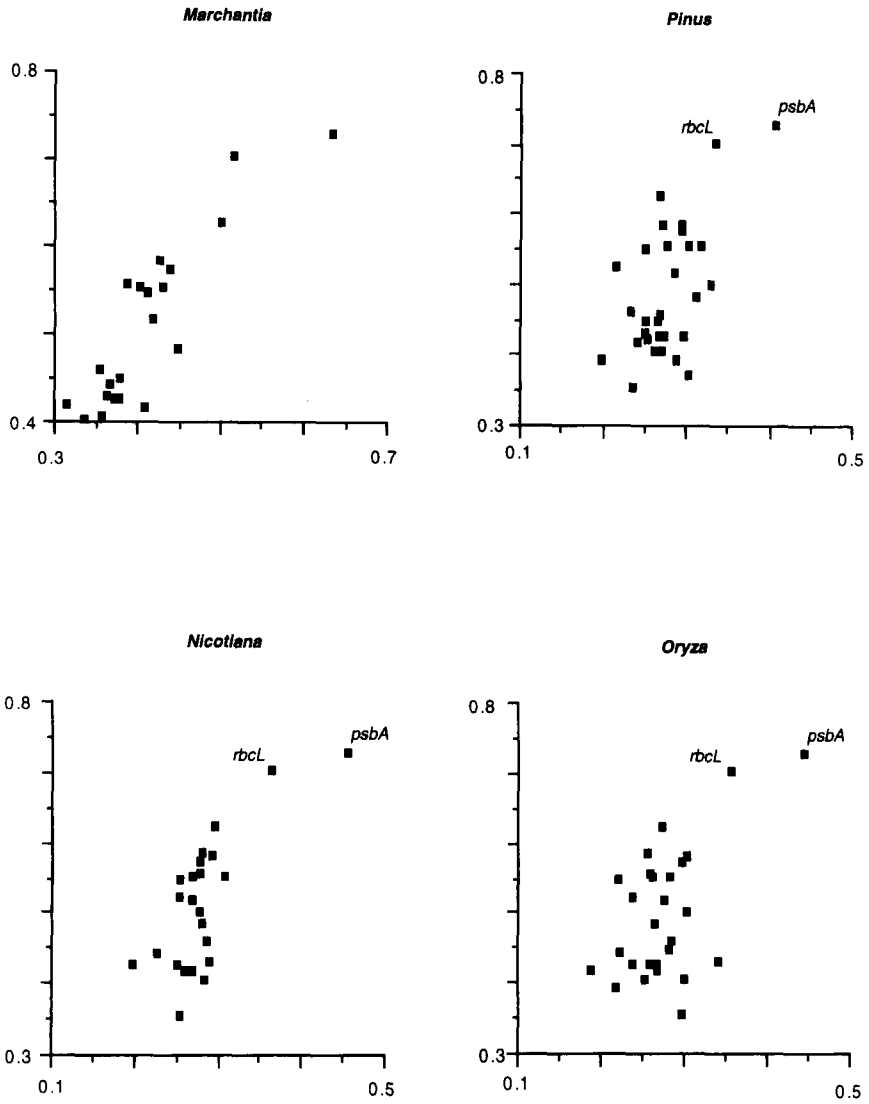

Fig. 2. Comparison of the CAI values of genes from *Cyanophora* with the CAI values of homologous genes from other species. Each plot shows a comparison to one of the other genomes and significance values for the correlations are given in Table 9. In each case, *Cyanophora* is plotted on the Y axis. The *rbcL* and *psbA* genes are indicated for the comparisons to *Pinus*, *Nicotiana*, and *Oryza* (see text).

The case of *Euglena* is quite interesting and deserves special mention. Based on both the cluster diagram and the third position composition (Table 2) there is no evidence at all for a selection pattern of codon usage at any locus, including *psbA*, of this species. Despite this, the highly expressed *rbcL* and *psbA* genes show a relatively high CAI value compared to other genes from *Euglena* (Table 5). Further, comparisons to the randomly generated distributions clearly indicate selection on some of the highly expressed genes (Table 3). Therefore, despite the apparent lack of a selection pattern in any gene, there is evidence that selection does play a role although it appears to be much weaker in *Euglena* than the other algae.

Table 9. Significance of correlations of CAI values from comparisons of *Cyanophora* to other species

To:	All genes	Excluding *psbA* and *rbcL*
Chlamydomonas	$P < 0.01$	$P < 0.01$
Odontella	$P < 0.01$	$P < 0.01$
Porphyra	$P < 0.01$	$P < 0.01$
Euglena	$P < 0.01$	$P < 0.05$
Marchantia	$P < 0.01$	$P < 0.01$
Pinus	$P < 0.05$	N.S.
Nicotiana	$P < 0.01$	$P < 0.05$
Oryza	N.S.	N.S.

For the plants, the results for *Nicotiana* and *Oryza* presented here are consistent with previous results which suggested that selection is essentially absent from the vast majority of angiosperm chloroplast genes (Morton 1993, 1996). The data indicate that *Pinus* is very similar to the flowering plants with regard to selection on codon usage. Therefore, *Pinus* and the flowering plants appear to have very weak selection on codon usage such that only a few genes, in particular *rbcL* and *psbA*, show any evidence for selection. In the case of *Marchantia*, the CAI values tend to be more comparable to the algae (Table 5), indicating that selection is stronger than in the other higher plants. There is also evidence from both the comparisons to *Cyanophora* and from the gene construction test that selection plays a significant role in this species. It is interesting that the three higher plant species that have weak selection also have lower genome A+T

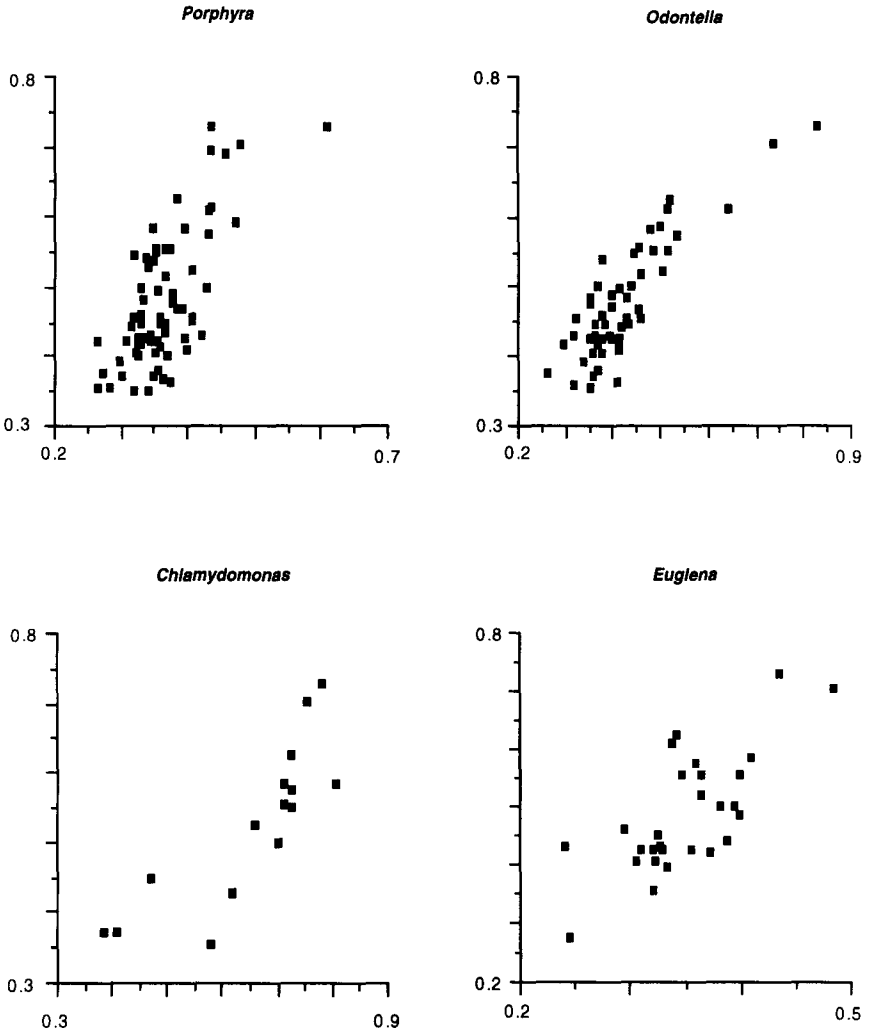

Fig. 2. Continued.

contents (Table 1). Whether there is any significance to this is unclear, but the possibility that selection on codon bias might be coupled to a form of selection on genome composition, perhaps for increased transcription rate, is intriguing.

It should be noted that the findings here are consistent with the recent evidence which indicates that the selection pattern of codon usage at the *psbA* locus of flowering plants is a remnant of an ancestral codon bias (Morton 1997; Morton and Levin 1997). The simulation method used in this study that demonstrates selection on the *psbA* gene of both *Nicotiana* and *Oryza* (Table 3) cannot distinguish between the codon bias resulting from current selection or being a remnant, as has been suggested.

It is likely that several factors are responsible for the variation in selection intensity across the different lineages. One that is likely to be responsible for the some of the difference observed between higher plants and algae is population size. Selection on codon bias is expected to exist only in species with fairly large effective population sizes (Li 1987), and it is possible that the generally larger CAI values in algae are a reflection, in part, of different effective population sizes.

The weaker selection in *Euglena* and *Porphyra* indicates that other factors are also likely to be important. A potential factor that must be considered is the reliance by a species on photosynthesis. *Euglena,* which shows such an unusual pattern of codon bias, is unique among the species in this study in that it can survive as a heterotroph if its chloroplasts are removed (van den Hoek 1995). This lack of reliance on photosynthesis could mean that selection on the translation efficiency of the major photosynthetic genes of the chloroplast is significantly reduced. Variation in photosynthetic rate between the green, red, and brown could contribute to the differences in selection intensity among the algae in this study.

A third possible factor is genome copy number. In organisms with large genome copy numbers, transcript availability could potentially decrease the need for high translation efficiency of individual transcripts (Morton 1997). This may account for a significant part of the variation between algae, such as *Chlamydomonas,* which has a single chloroplast per cell (ven den Hoek 1995), and flowering plants, which have large numbers, on the order of hundreds, of copies of the genome in most cells. In addition, the difference between *Cyanophora* or *Chlamydomonas* and *Euglena* may be partially due to the

fact that *Euglena* has multiple chloroplasts per cell and multiple genomes (van den Hoek 1995).

Conclusions

It is clear that the evolution of codon bias over all of the chloroplast lineages is a complex matter. Several factors are likely to be involved in determining the selective constraints on codon bias, and recent work has indicated that it is a dynamic process (Morton and Levin 1997). The variation in selective constraints among the different lineages also makes it likely that substitution dynamics are substantially different in different lineages which might be related to the debate concerning how composition bias influences the phylogenetic reconstruction of chloroplast origins (Lockhart et al. 1992). Even the use of amino acid data for phylogenetic analysis may be affected since it is quite possible that if selection on codon bias is strong, amino acid replacements that are normally neutral may be influenced by selection if the codons of the two amino acids have very different fitness values with regard to translation. Further work should help us understand the factors that have influenced the evolution of selective constraints on codon bias during chloroplast DNA evolution and how this affects the molecular evolution of chloroplast genes.

Acknowledgments. I would like to thank Brandon Gaut for some very helpful discussions concerning this manuscript and Michael Clegg for generous assistance. In addition, Walter Fitch kindly provided space and computer resources for some of the work described here.

References

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Hallick RB, Bairoch A (1994) Proposals for the naming of chloroplast genes. III. Nomenclature for open reading frames encoded in chloroplast genomes. Plant Mol Biol Reptr 12:S29–30

Hallick RB, Hong L, Drager RG, Favreau MR, Montfort A, Orsat B, Spielman A, Stutz E (1993) Complete sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acids Res 21:3537–3544

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

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

Klein RR, Mullet JE (1986) Regulation of chloroplast-encoded chlorophyll-binding protein translation during higher plant chloroplast biogenesis. J Biol Chem 261:11138–11145

Kowallik KV, Stoebe B, Schaffran I, Freier U (1995) The chloroplast genome of a chlorophyll a+c-containing alga, *Odontella sinensis*. Plant Mol Biol Reptr 13:336–342

Lake JA (1985) Evolving ribosome structure: domains in archaebacteria, eubacteria, eocytes and eukaryotes. Annu Rev Biochem 54: 507–530

Li WH (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol 24:337–345

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

Lockhart PJ, Penny D, Hendy MD, Howe CJ, Beanland TJ, Larkum AWD (1992) Controversy on chloroplast origins. FEBS Lett 301: 127–131

Long M, Gillespie JH (1991) Codon usage divergence of homologous vertebrate genes and codon usage clock. J Mol Evol 32:6–15

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 (1994) Codon use and the rate of divergence of land plant chloroplast genes. Mol Biol Evol 11:231–238

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

Morton BR (1997) Rates of synonymous substitution do not indicate selective constraints on the codon use of the plant *psbA* gene. Mol Biol Evol 14:412–419

Morton BR, Levin JA (1997) The atypical codon use of the plant *psbA* gene may be the remnant of an ancestral bias. Proc Natl Acad Sci USA 94:11434–11438

Mullet JE, Klein RR (1987) Transcription and RNA stability are important determinants of higher plant chloroplast RNA levels. EMBO J 6:1571–1579

Reith M (1995) Molecular biology of rhodophyte and chromophyte plastids. Annu Rev Plant Physiol Plant Mol Biol 46:549–575

Reith M, Munholland J (1995) Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. Plant Mol Biol Reptr 13: 333–335

Sharp PM, Li W-H (1987a) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230

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

Sokal RR, Rohlf FJ (1981) Biometry: the principles and practice of statistics in biological research. WH Freeman, San Francisco

Stirewalt VL, Michalowski CB, Loffelhardt W, Bohnert HJ, Bryant DA (1995) Nucleotide sequence of the cyanelle genome from *Cyanophora paradoxa*. Plant Mol Biol Reptr 13:327–332

van den Hoek C (1995) Algae: an introduction to phycology. Cambridge University Press, Cambridge

Wakasugi T, Tsudzuki J, Ito S, Shibata M, Sugiura M (1994) A physical map and clone bank of the black pine (*Pinus thunbergii*) chloroplast genome. Plant Mol Biol Reptr 12:227–241