J Mol Evol (2003) 56:673–688
DOI: 10.1007/s00239-002-2436-8

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary Patterns of Codon Usage in the Chloroplast Gene *rbc*L

Dennis P. Wall,[1] Joshua T. Herbeck[2,*]

[1] Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, USA
[2] Division of Insect Biology, University of California, Berkeley, Berkeley, CA 94720, USA

Received: 29 January 2002 / Accepted: 16 December 2002

Abstract. In this study we reconstruct the evolution of codon usage bias in the chloroplast gene *rbc*L using a phylogeny of 92 green-plant taxa. We employ a measure of codon usage bias that accounts for chloroplast genomic nucleotide content, as an attempt to limit plausible explanations for patterns of codon bias evolution to selection- or drift-based processes. This measure uses maximum likelihood-ratio tests to compare the performance of two models, one in which a single codon is overrepresented and one in which two codons are overrepresented. The measure allowed us to analyze both the extent of bias in each lineage and the evolution of codon choice across the phylogeny. Despite predictions based primarily on the low G + C content of the chloroplast and the high functional importance of *rbc*L, we found large differences in the extent of bias, suggesting differential molecular selection that is clade specific. The seed plants and simple leafy liverworts each independently derived a low level of bias in *rbc*L, perhaps indicating relaxed selectional constraint on molecular changes in the gene. Overrepresentation of a single codon was typically plesiomorphic, and transitions to overrepresentation of two codons occurred commonly across the phylogeny, possibly indicating biochemical selection. The total codon bias in each taxon, when regressed against the total bias of each amino acid, suggested that twofold amino acids play a strong role in inflating the level of codon usage bias in *rbc*L, despite the fact that twofolds compose a minority of residues in this gene. Those amino acids that contributed most to the total codon usage bias of each taxon are known through amino acid knockout and replacement to be of high functional importance. This suggests that codon usage bias may be constrained by particular amino acids and, thus, may serve as a good predictor of what residues are most important for protein fitness.

Key words: Codon bias — Codon usage bias — Green-plant phylogeny — *rbc*L — RuBisCo

Introduction

Codon usage bias is a pattern of unequal usage of codons within a single amino acid family, relative to codon frequencies predicted by the degeneracy of the genetic code. The phenomenon has been studied widely, as it is a link between patterns of genome organization and gene and protein evolution. Codon usage bias is often attributed to one of several potential causes: (1) overall genome nucleotide composition bias, as in mammalian isochores (Karlin and Mrazek 1996); (2) selection for translational accuracy and efficiency, as found in *E. coli* (Akashi 1994; Ikemura 1981a, 1982, 1985; Morton 1996; Xia 1996); or (3) a balance of mutational biases, natural selection, and genetic drift (Akashi 1997; Sharp et al. 1993). Most studies of codon bias have examined

*Present address: JBP Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA

Correspondence to: Dennis P. Wall, Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA; *email:* dpwall@stanford.edu

Table 1. Taxa used in this study, with major clade and GenBank accession number

Taxon	Plant group	Accession No.
Anacamptodon sp.	Musci: Bryopsida	
Andreaea rupestris	Musci: Bryopsida	L13473
Angiopteris evecta	Filicophyta	L11052
Anthoceros punctatus	Anthocerotophyta	U87063
Archidium sp.	Musci: Bryopsida	
Arthrocormus schimperi	Musci: Bryopsida	AF226797
Asterella tenella	Marchantiophyta: Marchantiales	U87064
Bazzania trilobata	Marchantiophyta: Jungermanniales	L11056
Brachythecium salebrosum	Musci: Bryopsida	AF158176
Brotherella recurvens	Musci: Bryopsida	L13475
Buxbaumia aphylla	Musci: Bryopsida	AF231909
Calocedrus decurrens	Coniferopsida	L12569
Calymperes afzelii sp.	Musci: Bryopsida	AF226789
Calypogeia muelleriana	Marchantiophyta: Jungermanniales	U87065
Chamaecyparis obtusa	Coniferopsida	L12570
Chara connivens	Charophyta	L13476
Chlamydomonas moewusii	Chlorophyta	M15842
Chlorella sp.	Chlorophyta	
Codium fragile	Chlorophyta: Ulvophyceae	M67453
Coleochaete orbicularis	Charophyta	L13477
Conocephalum conicum	Marchantiophyta	U87066
Cryptomeria japonica	Coniferopsida	L25751
Cycas circinalis	Cycadophyta	L12674
Dicranum bonjeani	Musci: Bryopsida	AF135068
Dumortiera hirsuta	Marchantiophyta	U87068
Encephalartos arenarius	Cycadophyta	L12676
Ephedra tweediana	Gnetophyta	L12677
Equisetum arvense	Equisetophyta	L1105
Exostratum blumei	Musci: Bryopsida	AF226800
Fissidens philonotulus	Musci: Bryopsida	AF226809
Funaria hygrometrica	Musci: Bryopsida	AF226818
Geothallus tuberosa	Marchantiophyta	U87070
Ginkgo biloba	Ginkgophyta	D10733
Gnetum parvifolium	Gnetophyta	D10734
Grimmia sp.	Musci: Bryopsida	
Haplomitrium hookeri	Marchantiophyta	U87072
Herbertus pensilis	Marchantiophyta	U87073
Hookeria acutifolia	Musci: Bryopsida	AF158170
Huperzia hippuridea	Lycopodiophyta	Y07931
Isoetes melanopoda	Lycopodiophyta	L11054
Jubula pennsylvanica	Marchantiophyta	U87074
Klebsormidium sp.	Charophyta	L13478
Lepidozia reptans	Marchantiophyta	U87075
Leucobryum antillarum	Musci: Bryopsida	AF226806
Leucodon temperatus	Musci: Bryopsida	AB019456
Leucophanes glaucum	Musci: Bryopsida	AF226799
Lilium superbum	Magnoliophyta	L12682
Linum perenne	Magnoliophyta	Z75681
Lophocolea heterophylla.	Marchantiophyta	U87076
Lunularia cruciata	Marchantiophyta	U87077
Lycopdium clavatum	Lycopodiopsida	Y07936
Magnolia tripetala	Magnoliophyta	AJ131927
Makinoa crispata	Marchantiophyta: Metzgeriales	U87078
Marchantia polymorpha	Marchantiophyta: Marchantiales	U87079
Marsilea quadrifolia	Filicophyta	L13480
Megaceros vincentianus	Anthocerotophyta	U87080
Metzgeria furcata	Marchantiophyta: Metzgeriales	U87081
Mitthyridium obtusifolium	Musci: Bryopsida	AF226777
Mnium cuspidatum	Musci: Bryopsida	U87082
Monoclea gottschei	Marchantiophyta: Monocleales	U87083
Nicotiana tabacum	Magnoliophyta	Z00044
Nitella translucens	Charophyta	L13482
Octoblepharum albidum	Musci: Bryopsida	AF226795
Ophioglossum engelmanii	Filicophyta	L11058

Table 1. Continued

Taxon	Plant group	Accession No.
Oryza sativa	Magnoliophyta	X15901
Pallavicinia lyellii	Marchantiophyta: Metzgeriales	U87084
Petalophyllum ralfsii	Marchantiophyta: Metzgeriales	U87086
Physcomitrella patens	Musci: Bryopsida	X74156
Picea sitchensis	Coniferopsida	X63660
Pinus radiata	Coniferopsida	X58134
Pleurozon sp.	Musci: Bryopsida	
Porella pinnata	Marchantiophyta: Jungermanniales	U87088
Pseudosymblepharis schimperiana	Musci: Bryopsida	AF226805
Pseudotsuga menziesii	Coniferopsida	X52937
Ricciocarpos natans	Marchantiophyta: Marchantiales	U87089
Securidaca diversifolia	Magnoliophyta	L01955
Selaginella sp.	Lycopodiophyta	L11280
Sequoia sempervirens	Coniferopsida	L25755
Sirogonium melanosporum	Charophyta	L13484
Smilax glauca	Magnoliophyta	Z77310
Sphaerocarpos texanus	Marchantiophyta: Sphaerocarpales	U87090
Sphagnum palustre	Musci: Bryopsida	L13485
Spirogyra maxima	Charophyta	L11057
Splachnum sphaericum	Musci: Bryopsida	AF005515
Stangeria	Cycadophyta	L12676
Calymperes tahitense	Musci: Bryopsida	AF226785
Tetraphis pellucida	Bryophyta: Polytrichopsida	U87091
Thuidium delicatulum	Musci: Bryopsida	AF158177
Tortula obtusissima	Musci: Bryopsida	AF226823
Tsuga heterophylla	Coniferopsida	X63659
Welwitschia mirabilis	Gnetophyta	AJ235814
Zamia inermis	Cycadophyta	L12683

multiple genes from a single genome but focused on a relatively narrow taxonomic sample (Comeron and Kreitman 1998; Duret and Mouchiroud 1999; Moriyama and Powell 1997; Morton 1993, 1998; Pan et al. 1998). To our knowledge only one study (Morton and Levin 1997) has investigated the evolution of codon usage bias within a relatively expansive phylogeny, in this case one of several angiosperms and outgroups. This kind of comparative phylogenetic research can focus the research attention more on the adaptive processes that direct molecular evolution. Thus it is surprising that such studies are so few, considering that they may lead to predictions that could direct both biochemical and phylogenetic research. For these reasons, we investigate here the evolution of codon usage bias in a single chloroplast gene, *rbc*L, within a broad phylogenetic framework of 92 taxa that includes all the major clades of green plants.

The gene *rbc*L encodes the large subunit of RuBisCO (ribulose 1,5-bisphosphate carboxlyase), a crucial enzyme in photosynthesis, and is perhaps the most abundant protein on the planet. Because *rbc*L plays a critical role in photosynthesis, its biochemistry has been studied extensively (Kellogg and Juliano 1997). Because of its functional importance, the gene evolves slowly and so suits phylogenetic studies that care to resolve deep splits in the evolution of green plants and their relatives. Both fields, biochemistry and phylogenetics, have fostered extensive sequencing of *rbc*L and have built a valuable database that includes the protein behavior of the gene.

The high functional significance and low rate of sequence divergence in *rbc*L have led authors to argue that change in codon usage bias is unlikely and should be historically infrequent (Albert et al. 1994; Morton 1993, 1996). This claim may be tested using a phylogeny provided that the taxonomic sampling and the measure of codon usage bias are sufficiently sensitive.

In our study, we optimize a new measure of codon usage bias to our green-plant phylogeny to test this claim and search for evolutionary trends in codon choice and the level of bias. While the use of the comparative phylogenetic methods is crucial to detect evolutionary patterns in codon bias, comparisons across a wide and taxonomically diverse set of taxa will be influenced strongly by fluctuations in genomic G + C content. The influence of G + C content fluctuations will confound explanations for evolutionary trends in levels of codon usage bias, making it impossible to discriminate fixed historical mutational biases (G + C content) from selection pressure (e.g., translational accuracy), the two general causes for codon usage bias mentioned above. To abate this problem we develop a maximum likelihood measure of codon usage bias that accounts for nucleotide composition. By factoring out potential effects of nucleotide composition on the measure of codon

bias, we attempt to limit explanations of phylogenetic trends in codon usage bias to selection and/or genetic drift caused by population demographics.

Methods

Taxa and Sequences

Ninety-two green-plant taxa were selected for analysis, spanning all major lineages of plants including green algae, liverworts, mosses, hornworts, pteridophytes, and seed plants (Table 1). Taxa were chosen according to simple criteria—an approximately even sampling across phylogenetic diversity and availability of a nearly complete *rbc*L sequence (minimum length of 1330 nucleotide bases). The taxa were organized into a composite tree (Fig. 1), using information and phylogenetic hypotheses culled from the current literature (Chapman et al. 1998; Hoot et al. 1999; Kenrick and Crane 1997; Pryer et al. 1995; Soltis and Soltis 2000; Soltis et al. 1999; Wolf et al. 1998).

Codon Frequencies

Codon frequencies for the 92 sequences were calculated using the Codon Usage Database (Nakamura et al. 2000) and arranged into a table for importation to the codon usage bias analysis program described in the appendix to this paper, in this issue (Slatkin and Novembre 2003).

Maximum Likelihood Analysis of Codon Usage

Given the phylogenetic scope of our analysis, it was also necessary to use a measure of codon bias that did not rely on a previously identified set of preferred codons. Also, to deduce the mechanisms influencing evolution of codon usage bias, it is necessary to account for multiple variables, including mutational dynamics (e.g., Morton 2001), and nucleotide composition. In our study, we quantify the extent of bias not attributable to G + C content by removing nucleotide composition from the calculation of codon bias using a new method described here briefly and in more detail in the appendix to this paper (Slatkin and Novembre 2003).

This measure assumes a null distribution of codon use that is determined solely by the individual taxon's G + C content (f). In this study, the f values were taken from all three positions of *rbc*L separately for each taxon. A section on different sources of G + C content follows.

For each amino acid in each taxon, the data consisted of the numbers of codons, n_i ($i = 1,...,k$). The total number of codons for that amino acid in that taxon is $n = \sum_{i=1}^{k} n_i$. The first question addressed by the method is whether there is a significant deviation from the expectation derived from the G + C content (f). A χ^2 test with $k-1$ degrees of freedom and a 5% significance threshold was used to determine whether the n_i, the usage data for an amino acid in one taxon, differed significantly from the expected usage, ne_i. If the null hypothesis that there is no significant difference was not rejected, then there was no significant bias for that amino acid in that taxon, and no overrepresented codons, and no further analysis was carried out.

When the null hypothesis was rejected, the usage data were fit first to a model that assumes a single overrepresented codon, hereafter referred to as the "single-preference codon model." This sample, n_i, derived from the single-preference codon model, was assumed to be a multinomial sample of size n, and from this distribution the bias, b, was estimated by maximizing the likelihood. The single-preference codon model was fit to all amino acid data whenever the null hypothesis was rejected.

For amino acids with a redundancy greater than two, the codon usage data were also fit to a model that assumes two preferred codons, hereafter referred to as "the double-preference codon model." Again, the sample, n_i, was assumed to be a multinomial sample of size, n, and the bias, b, was estimated by maximizing the likelihood.

Finally, a likelihood-ratio test was used to determine whether there was a significant preference for the single-preference codon model over the double-preference codon model, or vice versa. The ratio $R = -2\ln(L_2/L_1)$ was assumed to have a χ^2 distribution with 1 degree of freedom, where L_2 is the maximum likelihood for the double-preference codon model and L_1 is the maximum likelihood for the single-preference codon model. A 5% significance threshold was used to favor one of the two models of codon use.

G + C Content

Because of the lack of entire chloroplast genome sequences for all of the taxa included in this study, we were unable to use an exact f value (i.e., each organism's true chloroplast genome base composition). However, given the central importance of f to the results of our method, we checked the sensitivity of the estimated codon usage bias to four other derivations of f, beyond that derived from the three codon positions in *rbc*L.

1. The total G + C content of the chloroplast was used when available for pertinent taxa, of which there were six: *Mesostigma viride* (NC002186), *Chlorella vulgaris* (NC001865), *Marchantia polymorpha* (NC001319), *Pinus thunbergii* (NC001865), *Nicotiana tabacum* (NC001879), and *Oryza sativa* (NC001320). Although *Mesostigma* was not included in the 92-taxon phylogeny, we chose this taxon to represent phylogenetically the base of the algae (Lemieux et al. 2000; Qiu and Lee 2000). We also gathered GC content information for all other plant chloroplast genomes currently available in GenBank. These additional taxa were *Arabidopsis thaliana* (NC000932), *Spinacia oleracea* (NC002202), *Lotus japonicus* (NC002694), *Nephroselmis olivacea* (NC000927), *Triticum aestivum* (NC002762), and *Epifagus virginiana* (NC001568). These were used to explore the correlation between the entire-genome GC content and the GC content of *rbc*L.
2. The noncoding chloroplast sequence *trn*L was used when available for pertinent taxa. These were *Anacamptodon*, *Andreaea*, *Asterella*, *Buxbaumia*, *Chamaecyparis*, *Codium*, *Conocephalum*, *Dicranum*, *Fissidens*, *Haplomitrium*, *Hookeria*, *Huperzia*, *Leucobryum*, *Leucodon*, *Lophocholea*, *Marchantia*, *Mnium*, *Monoclea*, *Nicotiana*, *Octoblepharum*, *Ophioglossum*, *Picea*, *Ricciocarpos*, *Sequoia*, *Syrrhopodon*, *Tetraphis*, *Thuidium*, and *Tortula* (Table 2). This chloroplast region, called "*trn*L," spans the intergenic spacer bewteen *trn*A (L) and *trn*A (F).
3. The value of f was set to 0.5, thereby creating a completely neutral base composition such that all codons in a synonymous family are equally likely to be used.
4. G + C contents of third positions in *rbc*L were calculated separately for each of the 92 taxa.

After analyses of codon usage bias using different derivations of f, we settled on the one that maximized taxon diversity but did not cause obvious, spurious error in the calculations. We used this for all subsequent analyses.

Comparison with the "Effective Number of Codons"

We compared our codon usage bias measures with the measure "effective number of codons" (ENC) (Wright 1990). We chose this measure because it is widely used and considered more robust than many other measures of codon bias (Comeron and Aguade 1998). All comparisons were done with linear regression.

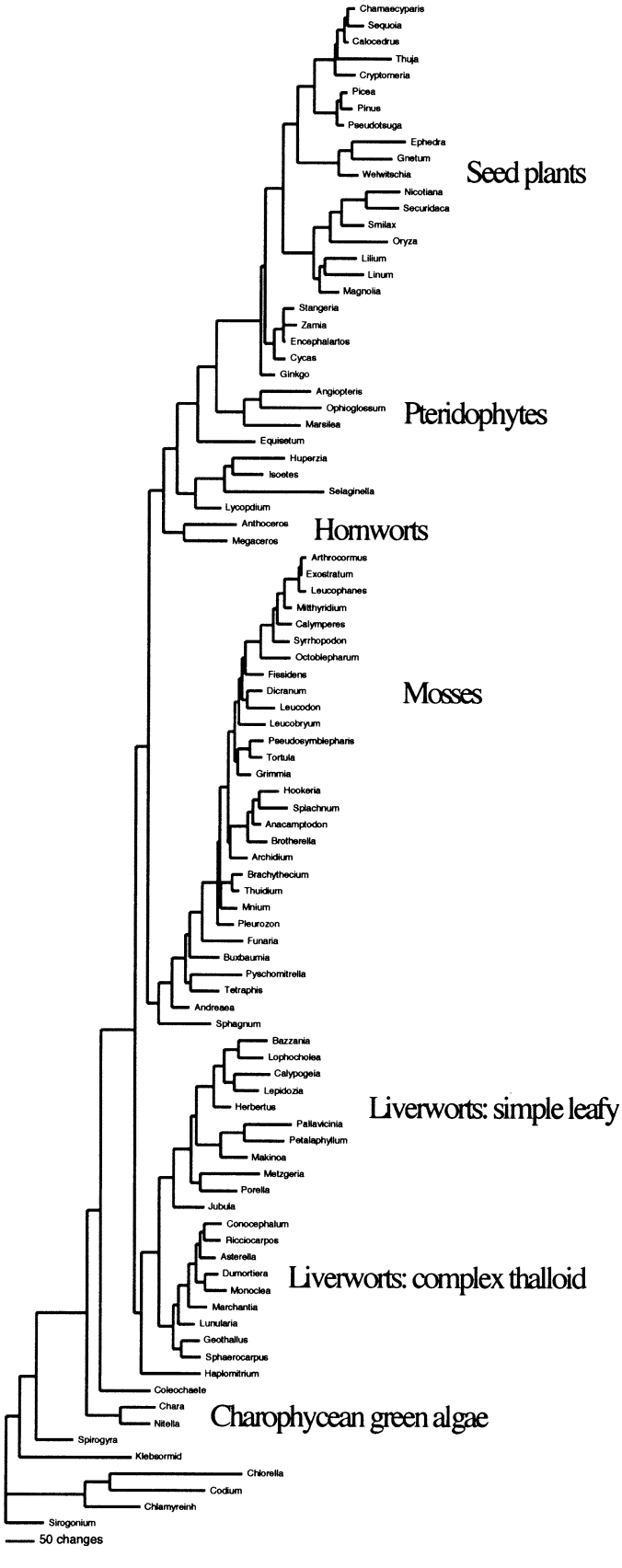

Fig. 1. Composite green-plant phylogeny of 92 taxa. Taxa were chosen from GenBank to attain an approximately even sampling across the major green-plant clades. The composite tree was built from published papers treating clades within the green plants.

Table 2. GenBank accession numbers and taxon names for the 28 sequences of *trn*L[a]

Taxon	Accession No.
Andreaea rothii	AF231905
Asterella californica	F228785
Buxbaumia aphylla	AF231909
Chamaecyparis obtusa	AB030061
Codium fragile	CFU12468
Conocephalum conicum	AF071834
Dicranum bonjeanil	AF135068
Fissidens subbasilaris	AF229913
Haplomitrium mnioides	AF231900
Hookeria acutifolia	AF161164
Huperzia phyllantha	AJ224555
Leucobryum glaucum	AF136090
Leucodon brachypus	AF161097
Lophocolea heterophylla	AF231899
Magnolia macrophylla	AF040680
Marchantia polymorpha	AF228786
Mnium hornum	AF023767
Monoclea gottschei	AF228788
Octoblepharum albidum	AF231156
Ophioglossum petiolatum	AF071849
Picea koyamae	AB045083
Lycopodium clavatum	LCL133265
Ricciocarpos natans	AF227672
Sequoia sempervirens	AB030051
Syrrhopodon gardneri	AF231146
Tetraphis pellucida	AF231908
Thuidium delicatulum	AF161132
Tortula ruralis	AF023722

[a] See Table 1 for the major clades to which these species belong.

Total Bias Calculations

To obtain a single value representative of the overall codon usage bias, referred to hereafter as total bias (TB), we summed the maximum likelihood estimates of bias, *b,* derived from the favorite of either the single-preference or the double-preference codon model, across two-, four-, and sixfold synonymous families. We found this sum to be strongly correlated with the summed χ^2/sample size ($r^2 = 0.96$).

Character Evolution

Of Total Bias. Evolutionary trends in codon usage bias and codon representation were studied on the phylogeny with MacClade version 4.0 (Maddison and Maddison 2000). The TB values were coded into discrete states by rounding.

Of Model Preference. The preference models for each taxon were provided as a separate output file by the program (0 for no preference, 1 when usage data best fit the single-preference codon model, and 2 when usage data best fit the double-preference codon model). These were optimized directly onto the composite phylogeny (Fig. 1). All optimizations were done by accelerated transformation (ACCTRAN).

Of Codon Preference. We used binary arithmetic to code preferences within each amino acid. For example, the binary number for a 1000 usage pattern (i.e., if only NNA codons were used in a fourfold amino acid) is 1. A fourfold amino acid using only NNT codons, corresponding to a 0001 number sequence, would received a binary code of 8 (2^3). If the NNA and NNG codons were preferred (1010), the binary number assigned was 5. In the sixfold amino acids, serine, leucine, and arginine, the sequences were alphabetical by third codon position for the fourfold component followed by the twofold component, also listed alphabetically by third position base. Thus in the sixfold serine that preferentially used the first NNT of the fourfold component, corresponding to a 000100 sequence, the binary number assigned was 8; if the second NNT codon was used preferentially—a 000001 sequence—the binary number assigned was 64.

This binary number output is contingent upon the order of codons within each synonymous family in the original codon frequency matrix. Therefore, the descriptions above hold for other data sets only if the original order of codons in the frequency matrix follows the order described above. The program (*see* the appendix to this paper [Slatkin and Novembre 2003]) gives the binary coded preferences in a simple table that can be imported into MacClade 4.0 after minor alteration. We did this to analyze the evolutionary pattern of codon preferences in the 92-taxon tree.

MajorAmino Acid Contributors to Total Bias

Finally, to infer those amino acids that contributed most to the overall TB for each taxon, we plotted individual bias measures per amino acid against the TB and searched for a significant correlation.

Results

Phylogenetic Analysis of Codon Bias

Figure 2 shows the phylogenetic trends in codon usage bias across an abbreviated phylogeny for four measures of *f* (G + C content). Marked differences in levels of total bias (TB) occurred across the land-plant lineages and algal outgroups using all measures of *f* (Fig. 2). The most significant evolutionary transition in codon usage bias found was the evolution of a low bias in the seed plants from a high bias in the basal-diverging land plants and algae. This particular trend was found in all of the different calculations of TB, except where *f* was set to the G + C content of the third codon position of *rbc*L. The TB values derived from this GC3 *f* produced the weakest pattern of transition from high to low bias in the seed plants and simple leafy liverwort clades, but the pattern was found using the the same *f* in the corresponding sums of χ^2 (Fig. 2). This was the only instance in which summed χ^2 and TB were not highly correlated (in all others $r^2 > 0.8$). One of the clearest patterns in the evolution of TB was when *f* was set to the G + C content of *trn*L. The total bias values in this case ranged from 6.0 and 7.1 in *Codium* and *Conocephalum* to 0.91 and 1.7 in *Picea* and *Nicotiana*. The TB values derived from the true (i.e., chloroplast genome) calculation of *f* showed a similarly stark evolutionary transition from TB values of 5.0 and 5.38 in *Chlorella* and *Marchantia* to 1.899

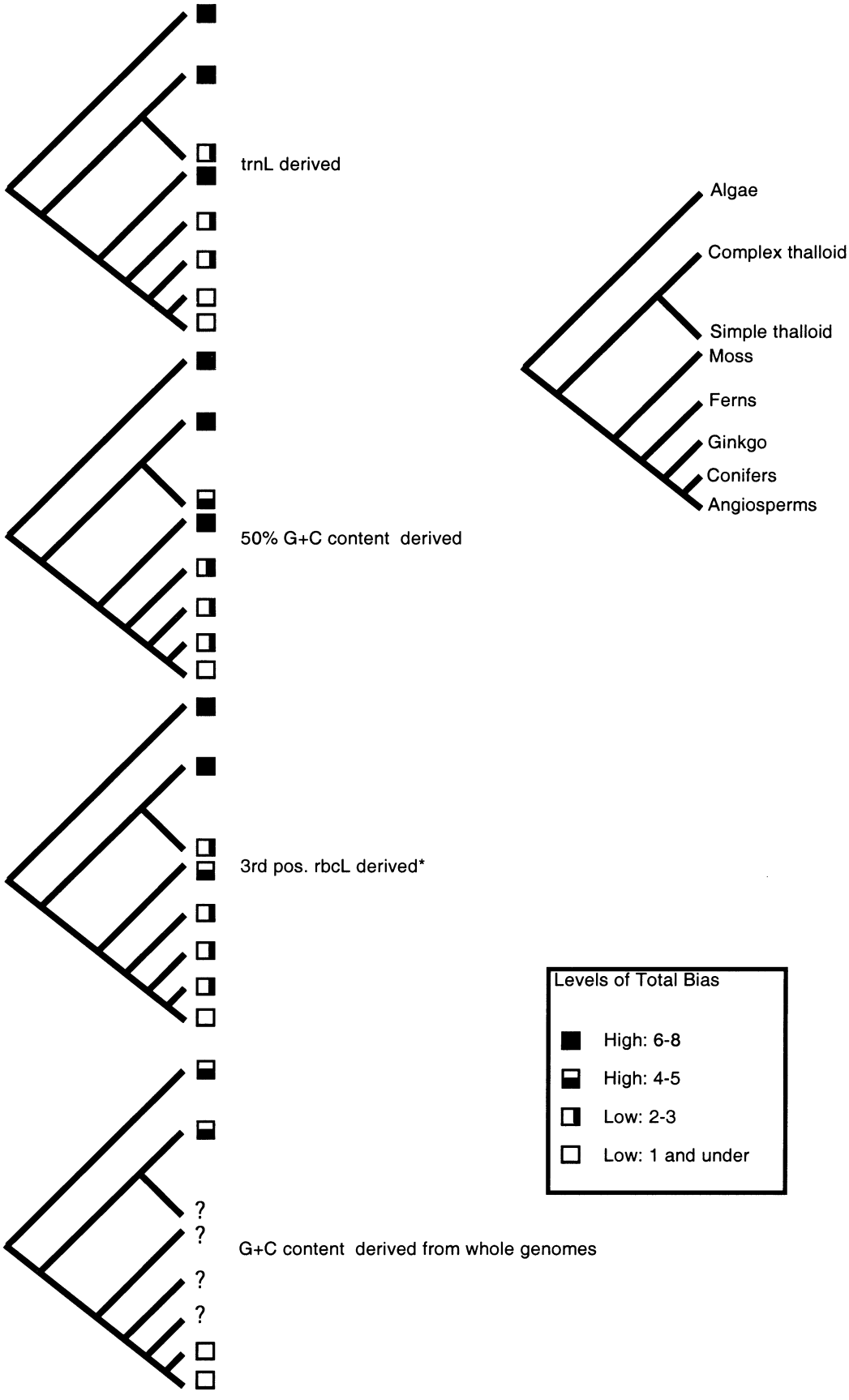

Fig. 2. The evolution of total codon usage bias across an abbreviated phylogeny of major green-plant clades, except hornworts (a phylogeny legend is given at the *upper right*). Each phylogeny is labeled by the data source for G + C content (f) that was used to estimate codon usage bias using the analysis described in the text and the appendix to this paper (Slatkin and Novembre 2003). The individual amino acid estimates of codon bias were summed to give the total codon usage bias. *Question marks* indicate where data were not available. The *asterisk* indicates that the pattern shown is based on sums of χ^2 rather than the total codon usage bias.

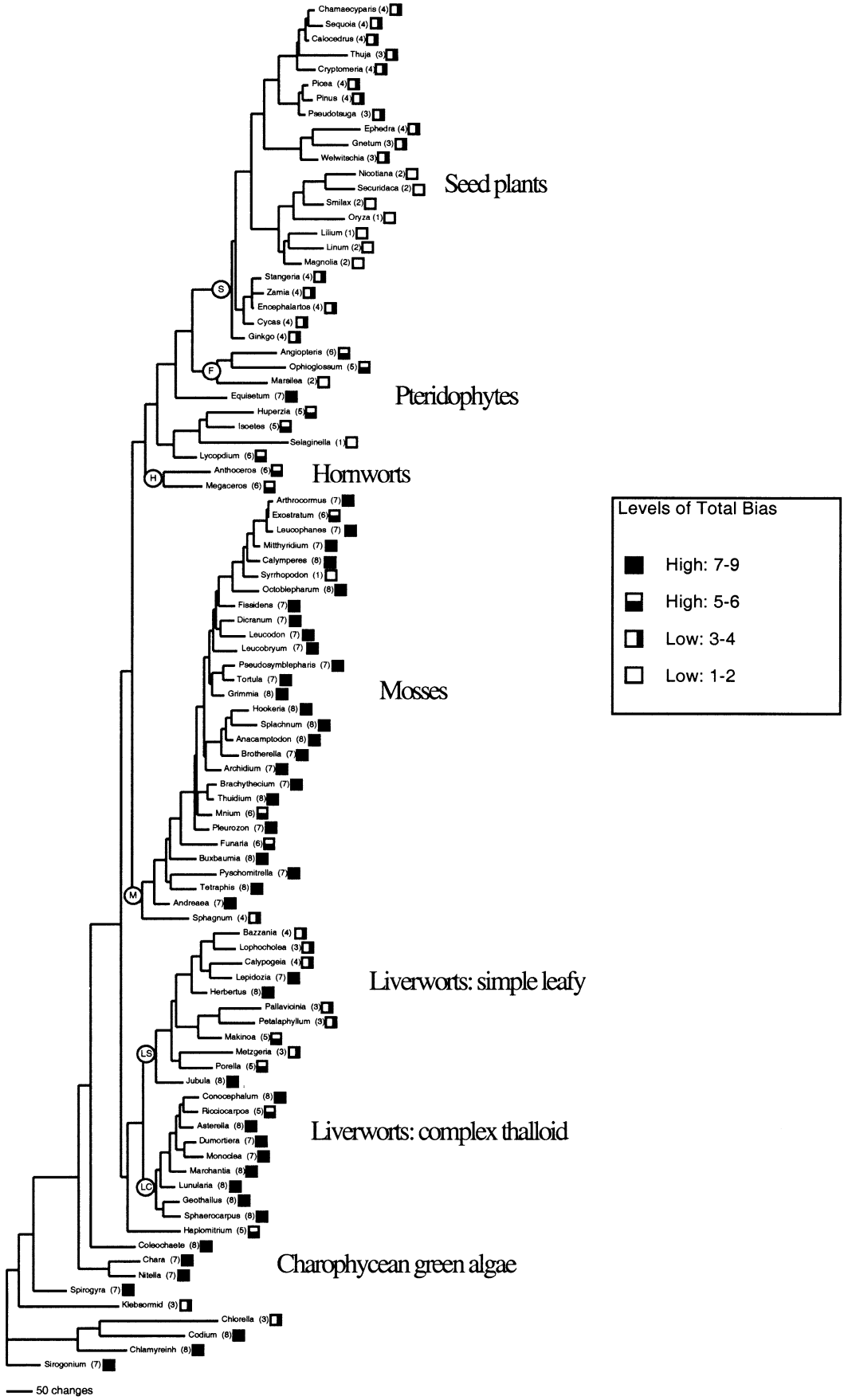

Fig. 3. The evolution of total codon usage bias across the green plants. Total codon usage is the sum of individual amino acid bias estimates. Circles at the base of clades identify major clades: LS, simple leafy liverworts; LC, complex thalloid liverworts; M, moss; H, hornworts; F, ferns; S, seed plants.

in *Oryza*. Similarly, the TB values derived from $f = 0.5$ shifted from 8.1 in *Codium* to 2.1 in *Magnolia*. In all cases where data were available (thus excluding TB values where f = genome G+C content), a similar but independent transition from high to low bias occurred at the base of the simple leafy liverwort clade (Fig. 2).

A comparison between our measure of codon usage bias and the "effective number of codons" (ENC) showed that the two indices are highly consistent: $r^2 = 0.84$ when TB values were calculated using f = genomic GC, $r^2 = 0.30$ when f = *trn*L GC, $r^2 = 0.03$ when f = third positions of *rbc*L, $r^2 = 0.94$ when $f = 0.5$, and $r^2 = 0.817$ when f = all three positions of *rbc*L. Only the third-position G+C content of *rbc*L was found not to be significantly associated with ENC.

The use of G+C content from the three codon positions of *rbc*L for a measure of f produced results highly consistent with those described above (Fig. 3). Again, two independent reductions in the level of codon usage bias were found at the base of the simple leafy liverwort and seed-plant clades (Fig. 3). We found the chloroplast genome G+C content to be significantly correlated with the G+C content at all three positions in *rbc*L codons ($r^2 = 0.18$) based on comparison between the plant chloroplast genomes and their corresponding *rbc*L sequence. All subsequent results are based on analyses using that derivation of f.

Evolution of Codon Preferences

A summary of the best fit for the single-preference codon model or the double-preference codon model in all four- and sixfold amino acids indicated that the usage patterns of one-half of these amino acids were better explained by the double-preference codon model (Table 3). Among those amino acids, only one, leucine, is a sixfold. Taxa whose usage data best fit the double-preference over the single-preferrence codon model >50% of the time (*Anthoceros, Brachythecium, Brotherella, Chlorella, Cycas, Encephalartos, Equisetum, Gnetum, Herbertus, Huperzia, Jubula, Leucodon, Lophocholea, Marsilea, Ophioglossum, Pleurozon, Porella, Securidaca, Selaginella, Sphagnum, Stangeria, Thuja, Thuidium*) were spread across the phylogeny, indicating a high degree of homoplasy in this character state.

Figure 4 provides an example, typical of all two-, four-, and sixfold amino acids, of the evolution of codon use across the green-plant phylogeny. The complete absence of codon preference was a homologous state found in the seed plants and simple leafy liverworts in most amino acid families, corresponding to the comparatively low level of biases detected in these two clades (Fig. 3). Elsewhere in the phylogeny, such shifts to complete lack of bias were infrequent and always homoplasious.

Table 3. The ratio of the double-preference codon model to the single-preference codon model for all four- and sixfold amino acids, ranked in increasing order

Amino acid	Double-:Single-preference codon model
Thr	0.37
Arg	0.39
Ser	0.40
Ala	0.65
Leu	0.98
Pro	1.13
Gly	1.27
Val	3.62

Four- and Sixfold Amino Acids

The usage data that best fit the single-preference codon model among the four- and sixfold amino acids (Table 3), when optimized to the green-plant phylogeny (Fig. 3), showed that the single-preference codon model was plesiomorphic and changes from the single- to the double-preference codon model were largely homoplasious. The only exception was alanine, which has two overrepresented codons ancestrally, and in which changes from the double-preference codon model to the single-preference codon model occurred as homologous transitions at the base of the mosses and traditional vascular plants (pteridophytes and seed plants; Fig. 3). The four- and sixfold amino acids that best fit the double-preference codon model (Table 3) were from taxa deeper in the phylogeny. One exception was the usage pattern found in proline, which was homologous in mosses and best fit the double-preference codon model but was otherwise homoplasious across the liverworts, hornworts, pteridophytes, and cycads, having converged from a plesiomorphic single-preference codon model (all other seed plants showed no preference in this amino acid). In almost all cases, when the plesiomorphic usage condition best fit the double-preference codon model, the codons overrepresented were NNA and NNT. Changes from the double-preference codon model of NNA/NNT to a single-preference codon model occurred, with rare exception, as a loss of preference for NNA. In one exception, valine, where the double-codon representation NNA/NNT was plesiomorphic and found in the basal lineages of green plants, losses of overrepresentation were of NNT and were largely homoplasious. The other exceptions were when both codons became unpreferred in favor of two completely different codons. This occurred twice in alanine, once in glycine, and eight times in leucine. All shifts from one double-preference codon model to another composed of entirely differently codons

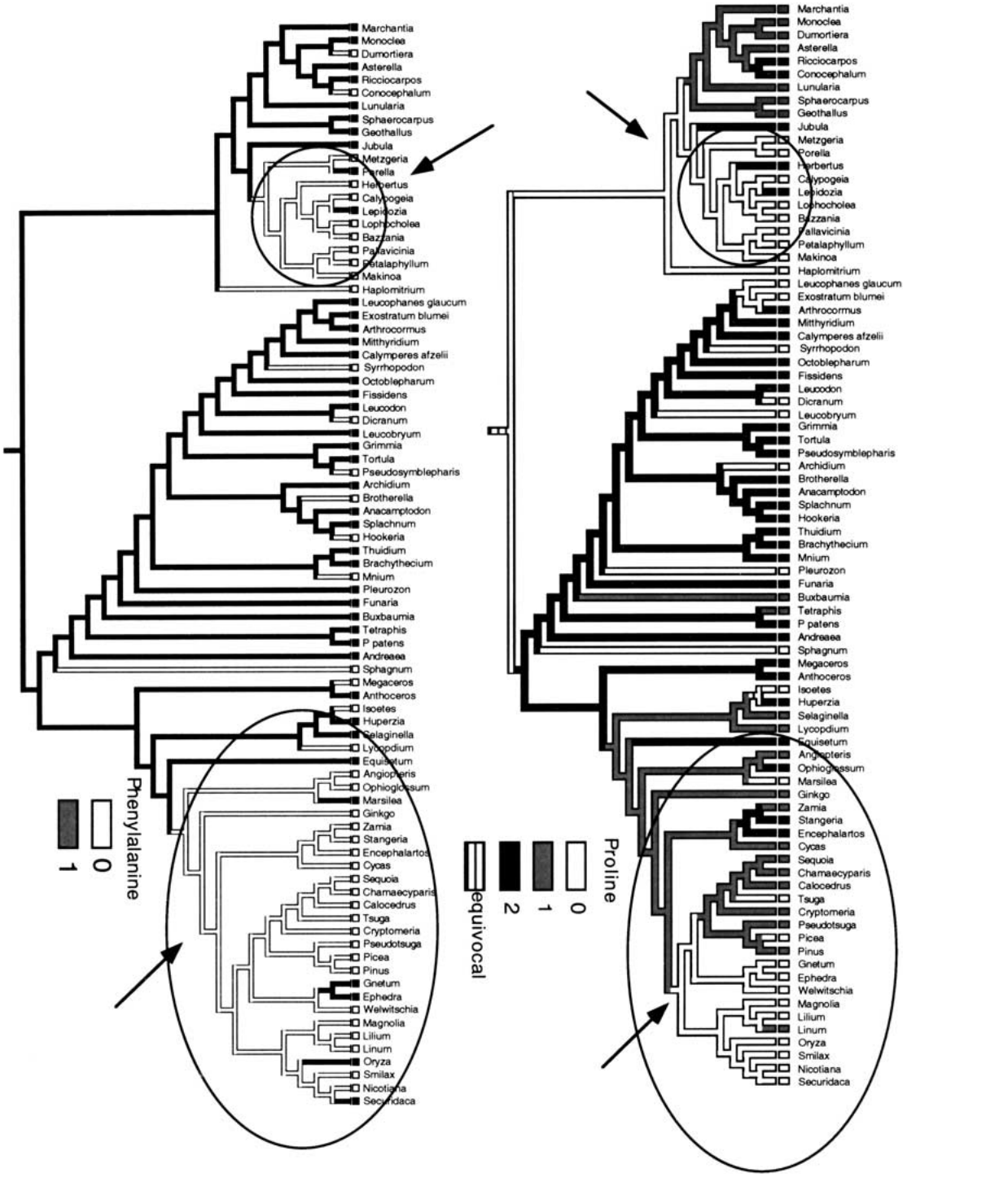

Fig. 4. An example typical of two-, four-, and sixfold amino acids of the evolution in codon preference: 0 = no preference, 1 = single-preference codon model, and 2 = double-preference codon model across an abbreviated version of the full phylogeny shown in Fig. 1 of 92 green plants. The examples provided are of phenylalanine and proline, two- and fourfold amino acids, respectively.

occurred independently and, thus, were always homoplasious.

There was a nearly equal number of gains and losses of codons used by taxa in four- and sixfold amino acids (Table 4). The vast majority of evolutionary gains of an additional preferred codon were "add-ons," where an NNT was joined with another codon, either NNA or NNC. Still a large number of shifts represented novel preference gains, where an NNT overrepresentation was supplanted by an NNA and NNG codon choice. In all the latter cases, the evolutionary transitions were homplasious. The former cases were a mix of homologous and homoplasious changes.

Table 4. (a) Summary of phylogenetic shifts in the preferred codon for four- and sixfold amino acids across all 92 green plant taxa;[a] (b) Summary of phylogenetic shifts in the preferred codon for twofold amino acids across all 92 land-plant taxa[b]

A

	To									
From	0	T	C	A	G	C&A	G&T	C&T	G&A	T&A
0	—	2	5	0	0	0	0	0	1	2
T	13	—	3	2	1	0	0	12	18	21
C	0	0	—	0	0	0	0	0	1	0
A	0	0	0	—	0	0	0	0	0	0
G	0	0	0	0	—	0	0	0	0	0
C&A	0	0	0	0	0	—	0	0	0	0
G&T	0	0	0	0	0	0	—	0	0	0
C&T	0	0	0	0	0	0	0	—	0	0
G&A	1	0	0	0	0	0	0	0	—	1
T&A	13	22	0	4	1	0	0	0	3	—

B

	To		
From	0 preferred	NNA or T	NNC or G
0 preferred	—	22	3
NNA or T	60	—	5
NNC or G	0	0	—

[a] Letters denote third-position nucleotides, under either the single- for the double-preference codon model.
[b] All twofold usage patterns can be explained by either the null model or the single-preference codon model.

Twofold Amino Acids

In the nine twofold synonymous families, NNT or NNA were always plesiomorphic and overrepresented when the null model was rejected. The taxa *Chlorella, Syhrropodon, Funaria, Marsilea,* and *Selaginella* were the only taxa that showed overrepresentation of NNC or NNG (all have peculiar, "long-branch" *rbc*L sequences). While the total loss of codon overrepresentation was common and often convergent among the twofold amino acids (Table 4) in most groups, such losses were synapomorphic in the seed-plant and simple leary liverwort clades for the twofold families phenylalanine, cystine, histine, and aspartic acid.

Isoleucine

The pattern of codon usage bias was identical to that shown in Fig. 3. Bias is high in the basal algal lineages, the complex liverwort and moss clades. Independent transitions to a significantly lower bias occurred in the simple leafy liverwort and seed-plant lineages. Only two taxa, *Syrrhopodon* and *Selaginella,* displayed no codon usage bias.

ATT was plesiomorphic and predominant among basal lineages of the green-plant phylogeny. The fern *Marsilea* was the only taxon that showed preference for ATC. Shifts from the single-preference codon model (preference for ATT) were to the double-preference codon model for ATT and ATA. These transitions occurred independently in *Sphagnum, Petalaphyllum, Ricciocarpos,* and *Sirogonium* and, also, as a single homologous transition marking a synapomorphy at the base of the ferns (*Angiopteris, Ophioglossum*) and seed plants. Reversions to the single-preference codon model for ATT occurred once in cycads, once in conifers, and once in the gnetophyte clade.

Major Amino Acid Contributors to the Overall Total Bias

Glutamic acid, lysine, cystine, and phenylalanine, all twofold amino acids (comprising only 35% of the total codons examined), were most strongly correlated with the TB (Fig. 5).Other twofolds, such as Gln and Tyr, also showed a significant correlation with TB ($r^2 = 0.562$ and 0.561, respectively). Leucine was the only sixfold amino acid that showed a relatively high correlation with TB; arginine and serine were weakly correlated ($r^2 = 0.311$ and 0.310). The fourfold amino acids showed the weakest correlations with TB (Fig. 5).

Discussion

The primary goal of this study was to analyze codon usage bias within a large comparative phylogenetic framework, an approach we believe can offer considerable insight to understanding molecular evolution. The second primary objective of this study was to introduce a new measure of codon usage bias that

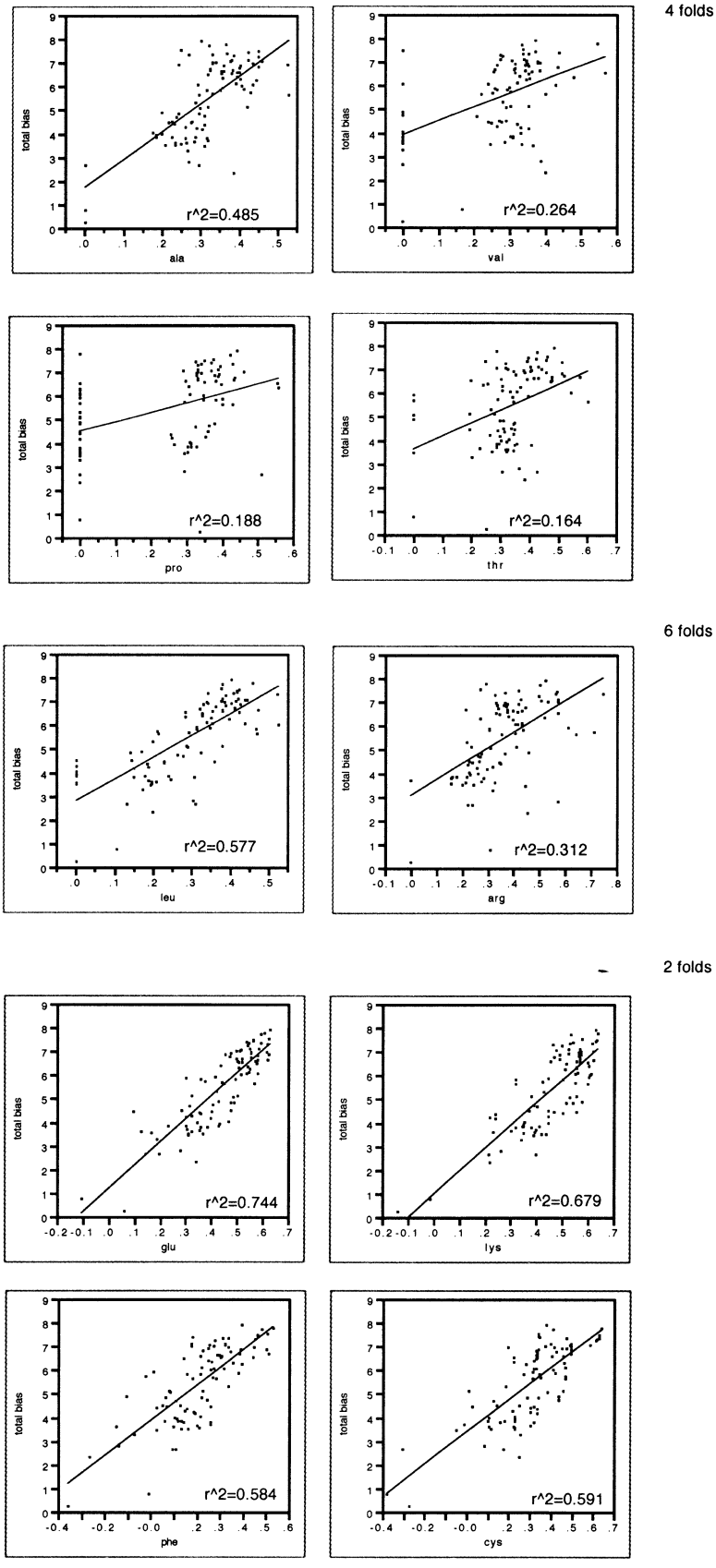

Fig. 5. Linear regressions of individual amino acid codon usage bias against total bias estimates to search for major amino acid contributors to the level of total bias reconstructed in *rbc*L across 92 green plants. The remaining amino acid r^2 values were as follows: Gln, 0.563; Tyr, 0.561; Asn, 0.377; Asp, 0.376; His, 0.357; and Gly, 0.22.

accounts for genomic G+C content (*f*) (*see* the appendix to this paper [Slatkin and Novembre 2003]). Many other commonly used measures of codon usage bias do not account for G+C content (but see Urrutia and Hurst 2001) and thus may conflate nucleotide composition with natural selection or

mutational dynamics (Morton 2001) as causes for codon usage bias. To meet our objectives, we were forced to use an approximate value of the chloroplast G+C content (f) since there are too few whole chloroplasts sequenced to build a sufficiently diverse phylogeny. Yet, because the value of f is critical to the value of codon usage bias estimated from the new measure, we checked for effects of variation in f on our estimates of bias. All but one of five sources of f yielded the same pattern of total bias (TB) evolution—the independent reduction in bias seen at the bases of both the simple leafy liverwort and the seed-plant clades. The exception was when f was derived from the third codon position of *rbc*L to calculate TB values, which showed virtually no pattern (interestingly, the summed χ^2 produced a pattern entirely consistent with the phylogenetic trends in TB derived from the other sources of f [the only instance where TB and summed χ^2 were not highly, positively correlated]). The reason for this different result is straightforward: codon usage bias is highly correlated with third-position G+C content. Third positions often have been used as an indirect measure of the extent of bias in preference for synonymous codons (Akashi et al. 1998; Comeron and Aguade 1998; Shields et al. 1988). Factoring out the content of nucleotides at the third position of a gene generally precludes the possibility of observing codon usage bias in that gene, except in cases of extreme bias. While third positions may be a plausible source for an f most similar to the overall genome nucleotide composition when whole-genomic data are unavailable, the third positions should come from genes other than the one(s) under investigation. We believe that our measure will become especially useful as more genomes are sequenced in entirety, thereby allowing use of the true G+C content rather than approximate values, as was necessary here. Nevertheless, our measure was robust to fluctuations in f and was found to yield information consistent with ENC, "the effective number of codons" (Wright 1990), a measure often used in studies of codon bias.

Because the evolutionary pattern was robust to changes in f, we chose to use all positions in *rbc*L since it permitted the use of broad phylogenetic diversity. That diversity allowed for a more in-depth analysis of both the evolution in level of bias and the evolution in codon choice.

Previous studies have shown that some codon bias exists for most amino acids and that individual taxa "choose" different codons within specific synonymous families, indicating differences in organismal "dialect" (Bennetzen and Hall 1982; Grantham et al. 1980; Ikemura 1981a, 1982, 1985). However, those changes in dialect are predicted not to be accompanied by increases in the level of bias only rarely, especially in genes of high functional importance (Kreitman and Antezana 1999). That is, for functionally important proteins, the level of codon usage bias should be the same (and high relative to that of less important genes) in all taxa that use the protein. Similarly, authors studying codon usage in green plants have shown that *rbc*L does not show a bias pattern reflective of selection but, rather, one that reflects the low G+C content characteristic of the chloroplast genome (Albert et al. 1994; Morton 1994; Morton and Levin 1997). Therefore, the codon usage bias differences across the green plants found here were unexpected given both the high level of conservation and the high functional importance of *rbc*L.

The prediction by Grantham et al. (1980) and others of dialect changes in codon use among organisms for the same gene proved to be true in our study. Overrepresented codons among the green plants are taxonomically distinct and consequently highly homoplasious. However, the causes for such changes in dialect remain unclear. The comparative phylogenetic method used here will allow for tests of adaptations at the interface between codons and amino acids and may provide an ideal framework for understanding processes that drive codon evolution.

The cases where an evolutionary transition occurred from a double-preference to a single-preference codon model occurred almost strictly as a loss of one codon from the set of two overrepresented codons, rather than loss of both and replacement by a uniquely overrepresented codon. This limited variety of change may reflect background mutational biases. Furthermore, the double-preference condition NNA/NNT was most common and most often found among the deeper branches of the green-plant phylogeny. The codon lost most frequently was NNA rather than NNT. This evolutionary pattern demonstrates that early in green-plant evolution, codons ending in A and T were both overrepresented, but later selection drove usage preference either toward a single-codon choice of NNT or toward a double-preference model consisting of two codons. The evolution of the single-codon overrepresentation to a double-codon overrepresentation was most commonly additive, where a new codon was added to the previous, rather than the addition of two new codons and the loss of the former single overrepresented codon.

Shifts from the single- to the double-preference codon model were more common than transitions in the other direction. A vast majority of the former evolutionary shifts were homoplasious, evolving independently numerous times across green plants, possibly reflecting a common selection regime. This theory is corroborated by the large number of evolutionary changes from a single-preference codon model favoring NNT to a double-preference for both

NNA/NNG, where each of the two newly preferred codons is unique.

Given that our pattern of codon bias and usage in *rbc*L cannot be explained by G+C content, an alternative cause—selection or genetic drift—may be responsible. One possible explanation has been selection via adaptation to the *trn*A pool, presumably for enhanced translational efficiency (Bagnoli and Lio 1995; Ikemura 1985; Miyasaka 1999). Studies of codon usage in the chloroplast gene *psb*A, which codes for the central protein of photosystem II, have found adaptation to the *trn*A pool of the chloroplast a likely explanation (Morton 1993, 1994, 1996, 1998; Morton and Levin 1997). These studies have shown that for the twofold amino acids and the threefold isoleucine, the only *trn*A encoded by the chloroplast genome is complementary to the NNC codon—and that this codon is used preferentially in *psb*A for these amino acids (Morton 1993; Umesono et al. 1988). In our study, preference for NNC in these amino acids was rare, suggesting that adaptation to *trn*A abundance may play less of a role than previously thought.

Further questioning of the strength of the hypothesis favoring adaptation to *trn*A abundance arises from the fact that the double codon preference model is commonly the best explanation for the usage pattern, especially in the amino acids leucine, proline, glycine, and valine, where this model fits the usage data of all 92 taxa $>50\%$ of the time. As far as is known, most chloroplast genomes of green plants (excepting *Nicotiana* and *Epifagus*, for example) carry 31 *trn*A's (Gillham 1994). In theory, this number is enough that *trn*A's from other genomes need not be employed (Gillham 1994). However, the high percentage of two overrepresented codons suggests that without invoking *trn*A importation, the typical chloroplast does not have enough *trn*A's to explain such a usage pattern. Although importation may occur with concomitant adaptation to a *trn*A pool that extends beyond the boundaries of the chloroplast, another explanation may prove more parsimonious. One possible explanation for two overrepresented codons is the combined effects of genetic drift and mutational biases allowing less optimal codons to become overrepresented.

Our data suggest that certain amino acids constrain the overall level of codon bias in *rbc*L, especially the twofolds glutamic acid, lysine, cystine, and phenylalanine. Previous biochemical knockout experiments, reviewed earlier (Kellogg and Juliano 1997), have examined the significance of particular amino acids to the function of the photosynthetic emzyme, RuBisCO. These studies have shown that lysine and glutamic acid are among the most functionally significant, since all attempts to knock out these amino acid residues at various positions within the gene (particularly at or around the active site) caused complete loss of protein function (Kellogg and Juliano 1997). In our study, glutamic acid and lysine contribute most to overall codon bias levels in *rbc*L. In contrast, threonine, which contributed least to overall bias levels, does not appear to effect the function of RuBisCO when transmutated to another amino acid. These results are too cursory to be conclusive, but more research on the predictive power of codon bias (as an indicator of which amino acids are of more importance to the function of the protein) is warranted.

While the processes driving the codon usage bias and shifts in codon preference remain unknown, the fact remains that such dramatic drops in codon usage bias—bordering no preference at all in many amino acids—occurred twice and, surprisingly, in seed plants, the most diverse group in our study. The evolution of such a decreased codon usage bias may suggest relaxed selectional constraints, as found in another chloroplast gene, *psb*A (Morton and Levin 1997). We predict that additional phylogenetic comparisons between the simple liverwort and the seed-plant clades, and of *rbc*L to other chloroplast genes, may improve our understanding of the processes responsible for the decreases in codon usage bias.

Further research on amino acid location in the secondary and tertiary structure of proteins may help clarify the relationship between codon usage bias and structural and/or functional importance. This perhaps will provide a detailed framework from which to build more robust models to improve our understanding of not only molecular evolution in general, but also how we interpret molecular data for reconstructing phylogenies.

In summary, this comparative analysis of synonymous codon usage in the AT-rich chloroplast highlights patterns of codon bias and preference. We used an estimate of codon bias that accounts for one factor influencing observed differences in the use of synonymous codons in *rbc*L—genome nucleotide composition. The estimate leaves three other explanations for observed changes in codon usage patterns: (1) selection, (2) mutational dynamics such as those suggested by Morton (2001), and (3) genetic drift. The presence of NNC and NNG preferred codons in several taxa is consistent with a selection hypothesis, given the extreme AT richeness of all chloroplasts in this study. However, it is clear that codon bias is heavily affected by background mutational biases and genetic drift. It follows that codon bias in *rbc*L is under weak selection, at best, and would most likely be only of intermediate (not extreme) bias, given equal mild composition bias in the chloroplast. Future work is needed to understand the relative importance of mutational dynamics, drift, and selection on the evolution of codon choice in *rbc*L and, indeed, in any other system.

Acknowledgments. We thank Montgomery Slatkin, Rosalyn Sayman, John Novembre, and the NSF PEET program for its support (NSF DEB-9712347 to Brent D. Mishler [D.P.W] and NSF DEB-9521835 to Jerry Powell and Felix Sperling [J.H]). Our manuscript was greatly improved by comments of Martin Krietman and two anonymous reviewers.

References

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

Akashi H (l997) Codon bias evolution in Drosophila. Population genetics of mutation-selection drift. Gene (Amsterdam) 205:269–278

Akashi H, Kliman RM, Eyre-Walker A (1998) Mutation pressure, natural selection, and the evolution of base composition in Drosophila. Genetica (Dordrecht) 102–103:49–60

Albert VA, Backlund A, Bremer K, et al. (1994) Functional constraints and rbcL evidence for land plant phylogeny. Ann Ma Bot Garden 81:534–568

Bagnoli F, Lio P (1995) Selection, mutations and codon usage in a bacterial model. J Theor Biol 173:271–281

Bennetzen JL, Hall BD (1982) Codon usage in yeast. J Biol Chem 257:3026–3031

Chapman RL, Buchheim MA, Delwiche CF, et al. (1998) Molecular systematics of the green algae. In: Solt is DES, Doyle PS (eds) *Molecular systematics of plants, II: DNA sequencing*. Kluwer Academic, Norwell, MA, Dordrecht, pp 508–540

Comeron JM, Aguade M (1998) An evaluation of measures of synonymous codon usage bias. J Mol Evol 47:268–274

Comeron JM, Kreitman M (1998) The correlation between synonymous and nonsynonymous substitutions in Drosophila: Mutation, selection or relaxed constraints? Genetics 150:767–775

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

Gillham N (1994) *Organelle genes and genomes*. Oxford University Press, New York

Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:43–74

Hoot SB, Magallon S, Crane PR (1999) Phylogeny of basal eudicots based on three molecular data sets: atpB, rbcL, and 18s nuclear ribosomal DNA sequences. Ann Ma Bot Garden 86: 1–32

Ikemura T (1981a) Correlation between the abundance of Escherichia coli transfer in RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146: 1–21

Ikemura T (1981b) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol 151:389–409

Ikemura T (1982) Correlation between the abundance of yeast tRNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 158:573–579

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

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

Kellogg EA, Juliano ND (1997) The structure and function of RuBisCO and their implications for systematic studies. Am J Bot 84:413–428

Kenrick P, Crane PR (1997) The origin and early evolution of plants on land. Nature (London) 389:33–39

Kreitman M, Antezana M (1999) The population and evolutionary genetics of codon bias. In: Singh RS, Krimbas CB (eds) Evolutionary genetics: From molecules to morphology. Cambridge University Press, Cambridge, Vol 1, pp 83–101

Lemieux C, Otis C, Turmel M (2000) Ancestral chloroplast genome in Mesostigma viride reveals an early branch of green plant evolution. Nature (London) 403:649–652

Maddison DR, Maddison WP (2000) MacClade 4: Analysis of phylogeny and character evolution. Singuer Associates, Sunderland MG

Miyasaka H (1999) The positive relationship between codon usage bias and translation initiation AUG context in Saccharomyces cerevisiae. Yeast 15:633–637

Moriyama EN, Powell JR (1997) Codon usage bias and tRNA abundance in Drosophila. J Mol Evol 45:514–523

Morton BR (1993) Chloroplast DNA codon use: Evidence for selection at the psb A 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 (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 (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, Levin JA (1997) The atypical codon usage of the plant psbA gene may be the remnant of an ancestral bias. Proc Nat Acad Sci USA 94:11434–11438

Nakamura Y, Gojobori T, Ikemura T (2000) Codon usage tabulated from international DNA sequence databases: Status for the year 2000. Nucleic Acids Res 28:292

Pan A, Dutta C, Das J (1998) Codon usage in highly expressed genes of Haemophilus influenzae and Mycobacterium tuberculosis: Translational selection versus mutational bias. Gene (Amsterdam) 215:405–413

Pryer KM, Smith AR, Skog JE (1995) Phylogenetic relationships of extant ferns based on evidence from morphology and rbcL sequences. Am Fern J 85:205–282

Qiu Y-L, Lee J (2000) Transition to a land flora: A molecular phylogenetic perspective. J. Phycol 36:799–802

Sharp PM, Stenico M, Peden JF, Lloyd AT (1993) Codon usage: Mutational bias, translational selection, or both? Biochem Soc Trans 21:835–841

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in Drosophila genes are not neutral: Evidence of selection among synonymous codons. Mol Biol Evol 6:704–716

Slatkin M, Novembre J (2003) Appendix to paper by Wall and Herbeck. J Mol Evol 56:689–690

Soltis ED, Soltis PS (2000) Contributions of plant molecular systematics to studies of molecular evolution. Plant Mol Biol 42:45–75

Soltis PS, Soltis DE, Wolf PG, et al. (1999) The phylogeny of land plants inferred from 18S rDNA sequences: Pushing the limits of rDNA signal? Mol Biol Evol 16:1774–1784

Umesono K, Inokuchi Y, Shiki M, et al. (1988) Structure and organization of Marchantia polymorpha chloroplast genome II. Gene organization of the large single copy region from rps′ 12 to atpB. J Mol Biol 203:299–331

Urrutia AO, Hurst LD (2001) Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics 159: 1191–1199

Wolf PG, Pryer KM, Smith AR, Hasebe M (1998) Phylogenetic studies of extant pteridophytes. In: Soltis DES, Doyle PS (eds) *Molecular systematics of plants. II: DNA sequencing*. Kluwer Academic, Norwell, Ma, Dordrecht, pp 541–556

Wright F (1990) The ''effective number of codons'' used in a gene. Gene 87:23–29

Xia X (1996) Maximizing transcription efficiency causes codon usage bias. Genetics 144:1309–1320