Mutational Bias Affects Protein Evolution in Flowering Plants

Huai-chun Wang, Gregory A. C. Singer, and Donal A. Hickey

Department of Biology, University of Ottawa, Ottawa, Ontario, Canada

Amino acid sequences from several thousand homologous gene pairs were compared for two plant genomes, *Oryza sativa* and *Arabidopsis thaliana*. The *Arabidopsis* genes all have similar G+C (guanine plus cytosine) contents, whereas their homologs in rice span a wide range of G+C levels. The results show that those rice genes that display increased divergence in their nucleotide composition (specifically, increased G+C content) showed a corresponding, predictable change in the amino acid compositions of the encoded proteins relative to their *Arabidopsis* homologs. This trend was not seen in a "control" set of rice genes that had nucleotide contents closer to their *Arabidopsis* homologs. In addition to showing an overall difference in the amino acid composition of the homologous proteins, we were also able to investigate the biased patterns of amino acid substitution since the divergence of these two species. We found that the amino acid exchange matrix was highly asymmetric when comparing the High G+C rice genes with their *Arabidopsis* homologs. Finally, we investigated the possible causes of this biased pattern of sequence evolution. Our results indicate that the biased pattern of protein evolution is the consequence, rather than the cause, of the corresponding changes in nucleotide content. In fact, there is an even more marked asymmetry in the patterns of substitution at synonymous nucleotide sites. Surprisingly, there is a very strong negative correlation between the level of nucleotide bias and the length of the coding sequences within the rice genome. This difference in gene length may provide important clues about the underlying mechanisms.

Introduction

Differences in G+C content among genomes have been intensively studied and wide variations have been noted both among entire genomes and among genes within genomes (Li 1997; Karlin, Campbell, and Mrazek 1998; Gautier 2000). The differences in nucleotide content between genomes have been shown to cause concomitant changes in the amino acid compositions of the encoded proteins (Collins and Jukes 1993; Foster, Jermiin, and Hickey 1997; Lobry 1997; Wilquet and Van de Casteele 1999; Singer and Hickey 2000; Kreil and Ouzounis 2001). Most of these previous studies were based primarily on prokaryotic genomes because of the lack of large-scale genomic data for plants and animals. Such data are now becoming available, however. The recent availability of genomic data for multicellular plants and animals not only allows us to extend previous studies to the genomes of multicellular eukaryotes but also enables us to trace the patterns of nucleotide and amino acid substitution between lineages that have well-defined evolutionary relationships. Therefore, we not only see the end results of evolutionary changes between genomes but also trace the paths by which these changes took place.

In this study, we compared homologous gene pairs from two species of flowering plants, *Oryza sativa* (rice) and *Arabidopsis thaliana*. Because these two species diverged less than 200 MYA, many homologous sequences from the two genomes are unambiguously alignable. Moreover, the level of amino acid sequence divergence between homologous proteins is relatively low, allowing us to gauge the patterns of amino acid substitution. Finally, there is a wide variation in the nucleotide contents of the rice genes: some closely resemble their *Arabidopsis* homologs in G+C content, whereas others have significantly elevated levels of G+C relative to their homologs (Carels and Bernardi 2000). Because all of the genes diverged from their common ancestral sequences at the same point in evolutionary time, this provides us with a "controlled" evolutionary experiment, enabling us to do a comparative study of two sets of rice genes that are evolving under contrasting evolutionary constraints.

Key words: comparative genomics, angiosperm, nucleotide, amino acid.

E-mail: dhickey@uottawa.ca.

Mol. Biol. Evol. 21(1):90–96. 2004
DOI: 10.1093/molbev/msh003
Molecular Biology and Evolution vol. 21 no. 1

Materials and Methods

Sources of Sequence Data

Protein sequences from *O. sativa* were obtained from the Gramene database (Ware et al. 2002) (ftp://www.gramene.org/pub/gramene/protein/sequence/rice_sptrembl.fa). This database contained 8,985 sequences as of May 2002. From the protein sequence identifiers, we first got corresponding EMBL accession numbers by searching SwissProt (Bairoch and Apweiler 2000), then extracted corresponding EMBL sequence records (Stoesser et al. 2002). From the EMBL records we wrote a program to extract coding sequences and 9,916 coding sequences were obtained. A total of 443 sequences were shorter than 75 codons and were excluded from the analysis. The remaining sequences were subjected to a codon integrity check using CodonW (http://www.molbiol.ox.ac.uk/cu/), and we further cleaned the data by removing redundant sequences. The final data set of *O. sativa* coding sequences contains 7,886 nonredundant sequences. Using EMBOSS/transeq (Rice, Longden, and Bleasby 2000) to translate the file, we generated a corresponding nonredundant protein sequence file.

A total of 26,178 protein-coding sequences from *A. thaliana* (from five chromosomes) were downloaded from National Center for Biotechnology Information (NCBI) FTP server (ftp://ftp.ncbi.nih.gov/genbank/genomes/A_thaliana/). After passing the sequences to CodonW for codon integrity check and removing genes shorter than 75 codons, a total of 25,625 *Arabidopsis* coding sequences remained for analysis. Protein sequences of *Arabidopsis*

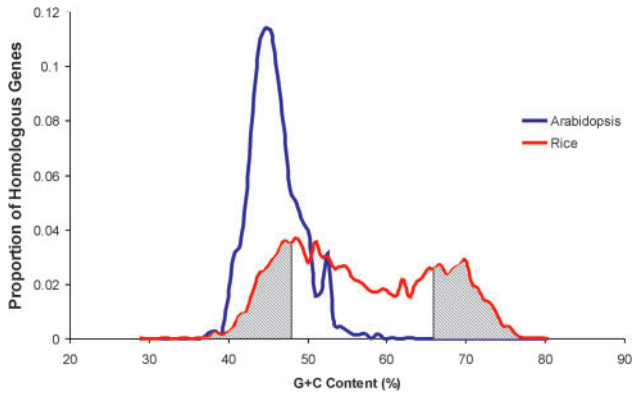

FIG. 1.—Distribution of G+C contents among rice and *Arabidopsis* genes. Homologous gene pairs only were used for this analysis. This data set included 8,894 genes—4,447 rice genes and 4,447 homologs from *Arabidopsis*.

were also obtained by translating the coding sequences using EMBOSS/transeq program.

Identification and Comparison of Homologous Sequences

Homologous protein pairs between *O. sativa* and *A. thaliana* were identified by performing BlastP searches (Altschul et al. 1990) of the rice protein sequences against *Arabidopsis* sequences with a cutoff expect score of 1e-20. When a rice protein had more than one *Arabidopsis* protein hit, the pair having the most significant expect score was retained. In all, 4,447 homologous pairs were identified.

After the homologous protein sequences had been identified, the corresponding nucleotide sequences were scored for nucleotide content. In this study, we ranked the rice homologs by their G+C content. We then compared the group of 1,000 rice genes with the highest G+C content (the "high G+C" class) to their homologs in the *Arabidopsis* genome. We also performed a parallel comparison between the group of 1,000 rice genes having the lowest G+C content (the "low G+C" class) and their homologs.

Identifying Amino Acids for GC-rich and AT-rich Codons

In the manner introduced by Foster, Jermiin, and Hickey (1997), we partitioned the codon table into three groups: codons that were GC-rich at the first two codon positions, codons that were AT-rich at the first two codon positions; and unbiased codons. The GC-rich codons encode glycine, alanine, arginine, and proline (GARP). The AT-rich codons encode phenylalanine, tyrosine, methionine, isoleucine, asparagines, and lysine (FYMINK). The unbiased codons fill two quadrants of the rearranged codon table, and they encoded serine (S), threonine (T), cysteine (C), tryptophan (W) and valine (V), leucine (L), glutamic acid (E), aspartic acid (D), histidine (H), and glutamine (Q).

Results

Compositional Distribution of Rice and *Arabidopsis* Homologous Genes

First, we confirmed previous reports that the genomes of monocotyledonous plants, including rice, have elevated G+C contents (Carels et al. 1998; Sasaki et al. 2002; Wong et al. 2002) and that there is a wide variation in G+C content among rice genes (Carels and Bernardi 2000; Yu et al. 2002). *Arabidopsis* genes have relatively low G+C contents, and they form a unimodal distribution, with a mean G+C content of about 44%. Rice genes, on the other hand show a much broader, multimodal distribution. One possible interpretation of these results is that the rice genome contains a unique set of genes that are characterized by a higher G+C content than the set of genes that is shared between the two genomes. To investigate this possibility, we repeated the same analysis based on 4,447 pairs of homologous genes that we identified using Blast searches (see *Materials and Methods*). As can be seen in figure 1, the same trends are seen in this subset of homologous genes. This indicates that the differences in nucleotide content are not simply the result of differences in gene content between the two genomes.

For our subsequent analyses, we compared the rice genes from the two ends of the distribution shown in figure 1 with their *Arabidopsis* homologs. Specifically, we chose 1,000 rice genes with the lowest G+C contents as one set ("low G+C" rice genes) and those with the highest G+C content as the other set ("high G+C" rice genes). Each of these sets represents approximately 25% of the total set of homologous sequence pairs (fig. 1). The low G+C rice genes have nucleotide contents that lie within the distribution of their *Arabidopsis* homologs, whereas the high G+C genes lie well outside the *Arabidopsis* range. The average nucleotide contents of these gene sets are shown in table 1. The average value for the low G+C rice genes is very close to that of their *Arabidopsis* homologs (table 1), whereas the high G+C rice genes have diverged greatly from their homologs. This is especially evident at the third positions of codons where the rice genes have a G+C content that is almost twice as large as that of their homologs (91.8% versus 46.9% [table 1]). It is interesting to note that, despite the large differences in average G+C content between the two groups of rice genes, the *Arabidopsis* homologs of all of these groups have relatively constant G+C contents (table 1). This is consistent with the unimodal distribution of nucleotide contents among all *Arabidopsis* genes (fig. 1).

Amino Acid Substitutions Between Rice and *Arabidopsis* Homologs

The main focus of our study was to investigate the degree to which changes in nucleotide composition could influence the evolution of the encoded proteins. In particular, we asked if different rice proteins might follow different evolutionary trajectories, depending on their evolving nucleotide content since the divergence of rice and *Arabidopsis*. A previous analysis of many prokaryotic

Table 1
Average Nucleotide Contents of Homologous Genes in Rice and *Arabidopsis* (Expressed As Percentages of G+C)

	Codon Position			
	First	Second	Third	Average
All homologous pairs (n = 4,447)				
Rice	58.1	44.7	66.4	56.4
Arabidopsis	51.5	41.0	43.9	45.5
High G+C genes[a] (n = 1,000)				
Rice	65.4	51.6	91.8	69.6
Arabidopsis	51.7	43.5	46.9	47.3
Low G+C genes[a] (n = 1,000)				
Rice	51.6	39.1	43.3	44.7
Arabidopsis	50.8	38.9	41.0	43.5

NOTE.—This table shows the G+C content of each codon position, along with the average value for all three codon positions.

[a] "High G+C" and "low G+C" refers to the nucleotide content of the rice genes only, not to their *Arabidopsis* homologs.

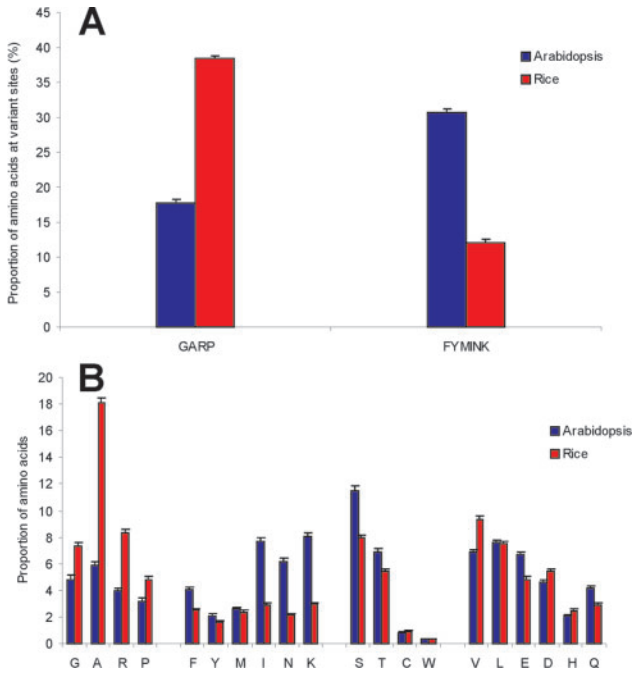

FIG. 2.—Amino acid content of homologous rice and *Arabidopsis* protein sequences. (*A*) The content of G, A, R, and P and F, Y, M, I, N, and K amino acids (expressed as percentages) for high G+C rice genes and their *Arabidopsis* homologs (1,000 genes each). These data are based on variant sites only in the aligned homologous sequences (see *Materials and Methods*). (*B*) Proportions of individual amino acids at variant sites (expressed as numbers per 10,000 variant sites) plotted for the high G+C rice genes and their homologs from *Arabidopsis*. The values for the rice genes are shown in red, and the values for the *Arabidopsis* homologs are shown in blue. The error bars represent the 99% confidence intervals.

genomes (Singer and Hickey 2000) showed that proteins encoded by GC-rich sequences are characterized by increased levels of the amino acids G, A, R, and P. These proteins show a corresponding decrease of amino acids encoded by AT-rich codons—namely, F, Y, M, I, N, and K. In this study, we compared the amino acid contents of proteins encoded by high G+C rice genes with their *Arabidopsis* homologs. We found that these rice genes do indeed show a highly significant increase in the level of GARP amino acids and a corresponding decrease in FYMINK amino acids. In contrast to this, the control set of low G+C rice genes encode proteins that have amino acid contents very similar to their *Arabidopsis* homologs.

In addition to showing simple differences in amino acid compositions between the homologous protein sequences, we wanted to investigate the patterns of amino acid substitution during the course of their evolutionary divergence. To do this, we aligned the homologous sequence pairs, and we then concatenated these alignments. The aligned sites can be classified as invariant (where the same amino acid appears in the rice and *Arabidopsis* sequences) or variant (where there is a difference between the two sequences). Because it is only these latter sites that contain information about sequence divergence, we recalculated the amino acid frequencies for these sites only. The results for the high G+C rice genes and their *Arabidopsis* homologs are shown in figure 2*A*. In this case, there is a twofold increase in the proportion of GARP amino acids in the rice sequences and an even greater proportional decrease in FYMINK amino acids. Not does a large average difference exist between the two concatenated sequences but also a consistent trend is seen among individual homologous gene pairs. For instance, 971 out of the 1,000 high G+C rice genes have higher GARP levels than their *Arabidopsis* homologs, and this trend is highly significant ($P < 0.00001$ in a one-tailed, paired-sample *t*-test). There are also consistent differences for all of the individual amino acids, within both the GARP and the FYMINK groups of amino acids (fig. 2*B*). Some of these frequency changes for individual amino acids are quite dramatic. For instance, the rice genes have a threefold increase in the proportion of alanine (A) at the variant sites and a twofold increase in arginine (R). They show a correspondingly large (more than twofold) decreases in isoleucine (I), asparagine (N), and lysine (K). The differences in amino acid composition are highly significant ($P < 0.001$) for all but three of the 20 pairwise comparisons. The exceptions are cysteine (C), tryptophan (W), and leucine (L).

The fact that there are large differences in the proportions of certain amino acids at the variable sites in aligned amino acid sequences indicates that the pattern of amino acid substitution between the high G+C rice genes and their *Arabidopsis* homologs is highly asymmetric. This can be seen more clearly when we construct an amino acid exchange matrix for the aligned sequences. Such a matrix is shown in figure 3. In this matrix, the rows represent the rice sequence data, the columns represent the *Arabidopsis* data, and the diagonal represents invariant sites. For example, there are 127 sites per 10,000 (shown in green on the matrix) where a lysine (K) in the *Arabidopsis* sequence is aligned with an arginine (R) in the rice sequence. In contrast, we see only 32 sites (also shown in green) where an arginine (R) in the *Arabidopsis* sequence has been aligned with a lysine (K) in the rice sequence. In other words, the great majority of arginine-lysine mismatches between the aligned sequences contain an arginine in the rice sequence and a lysine in the *Arabidopsis* sequence. Because arginine and lysine are

ARABIDOPSIS

RICE	G	A	R	P	F	Y	M	I	N	K	S	T	C	W	V	L	E	D	H	Q
G	**581**	39	15	14	4	3	2	4	40	22	59	17	3	1	9	7	26	31	6	11
A	71	**476**	26	39	17	6	15	34	32	43	159	81	11	2	81	46	49	27	10	26
R	15	13	**330**	9	4	5	5	6	23	127	29	17	3	2	7	14	24	10	13	29
P	12	23	10	**384**	3	2	3	5	10	16	38	18	1	0	10	13	14	10	6	11
F	2	4	1	1	**272**	30	3	11	2	0	5	3	2	2	9	28	1	1	2	1
Y	1	1	1	1	32	**190**	1	3	3	1	4	1	1	2	3	6	1	1	6	1
M	1	5	2	2	6	1	**101**	16	2	4	5	5	0	0	13	34	2	0	1	3
I	1	4	1	0	8	1	10	**181**	1	2	3	6	0	0	45	40	1	1	0	1
N	8	4	6	2	1	2	0	1	**167**	11	17	8	0	0	2	2	6	14	5	4
K	5	6	32	4	1	1	2	2	11	**213**	14	8	0	0	3	3	14	5	3	13
S	31	48	14	18	7	5	5	7	39	23	**339**	56	7	1	12	11	19	19	7	11
T	8	24	10	11	5	2	7	12	16	16	58	**231**	2	0	20	13	9	8	3	8
C	3	4	1	1	3	2	0	1	1	1	9	2	**139**	1	3	3	0	1	1	0
W	1	1	0	0	3	2	0	0	0	1	2	1	0	**123**	1	3	0	0	0	0
V	7	33	6	7	19	4	18	144	5	8	15	27	3	1	**382**	81	9	4	2	5
L	4	14	8	5	51	8	34	76	5	9	12	12	3	3	57	**584**	5	2	4	8
E	10	11	11	6	1	1	2	3	14	24	20	11	0	0	9	5	**287**	52	5	22
D	16	9	5	8	2	2	1	2	38	13	24	11	0	0	5	3	76	**290**	6	12
H	4	3	10	3	5	12	1	1	13	7	9	4	0	0	2	5	7	6	**120**	12
Q	4	6	12	6	1	2	2	1	9	17	12	7	0	0	3	7	22	6	8	**131**

FIG. 3.—Amino acid exchange matrix. This figure shows the pattern of sequence divergence between the high G+C rice genes and their *Arabidopsis* homologs. Homologous sequences were aligned, and the numbers of sequence mismatches were scored. Values in the matrix were scaled to represent the number of amino acid mismatches per 10,000 sites. Highlighted areas are discussed in the text.

biochemically similar amino acids, this allows the rice genes to increase their G+C content while maintaining their biochemical function.

The trend illustrated above by the arginine and lysine sequence mismatches extends to the entire group of GARP and FYMINK amino acids between the two sets of homologs (see yellow-shaded quadrants in fig. 3). For instance, out of a total of 526 mismatches between GARP and FYMINK amino acids, the rice sequence is overwhelmingly more likely (431 out of 526 times) to possess an amino acid from the GARP group. This represents a more than fourfold asymmetry in the pattern of amino acid substitution. Because we have a large data set of aligned homologous sequences, we are able to evaluate the rate of exchange between all 400 amino acid combinations. As expected, the rates of exchange between biochemically similar amino acids, such as isoleucine and valine, are high. What is noteworthy, however, is that there is a strong asymmetry in these exchanges. Isoleucine is found in the *Arabidopsis* sequence and valine is found in the rice sequence 144 times out of a total of 189 mismatches (fig. 3). This asymmetry that can be seen throughout the exchange matrix is indicative of significant, nonrandom evolutionary changes in the rice proteins as a result of the mutational bias at the DNA level.

It should be noted that it takes two nucleotide substitutions (at both the first and second codon positions) to achieve an amino acid exchange between the GARP and FYMINK amino acids. This means that, during a period of increasing G+C content in the rice sequences, many of the substitutions involve "intermediate" amino acids; that is, amino acids that are encoded by codons with intermediate nucleotide content. For instance, as the rice genes become more G+C rich, they are expected to gain GARP amino acids by single nucleotide substitutions from the pool of codons with intermediate nucleotide content. At the same time, they will lose FYMINK amino acids because of mutations that change the latter codons into codons of intermediate nucleotide content. Thus, these intermediate codons act as the "flow-through" from one extreme to the other. This effect is also illustrated in figure 3. For instance, we see that the source of the huge increase in alanine (A) in the rice sequences is not primarily the result of direct substitution from the FYMINK group, but substitution from other amino acids such as serine (S) (shown in blue). Likewise, the greatest single loss of isoleucine (I) in the rice sequences is to the G+C-intermediate valine (V, shown in red). More generally, we can see that exchanges between alanine in the rice sequences and the intermediate group of V, L, E, D, H, and Q (shown in orange in fig. 3) results in a net increase of 163 alanines (239 – 76 = 163). We used Fisher's exact test to calculate the significance of these asymmetries. All of the differences mentioned above are highly statistically significant ($P < 0.001$) and, in fact, the gain of alanine from all other amino acids is significant ($P < 0.05$) except for tyrosine (Y), cysteine (C), and tryptophan (W). In other words, alanine becomes a "sink" in the high G+C rice genes.

In contrast to these findings, when we constructed a parallel exchange matrix (not shown) for the low G+C rice genes and their *Arabidopsis* homologs, we found no evidence of asymmetry in the patterns of amino acid

substitution. Thus the asymmetric pattern of protein evolution is correlated with the changes in nucleotide content among the rice genes.

Possible Sources of Compositional Bias in Rice Genes and Their Encoded Proteins

Although our primary purpose was to explore the effects of mutational bias on the patterns of protein evolution, we also wished to infer the causes of this variation in nucleotide content between rice genes. First, we wished to reconcile the reports of Carels and Bernardi (2000), who state that there are two classes of genes in plants (one class being G+C-rich), and of Wong et al. (2002), who find there is a gradient of G+C content along individual rice genes. For instance, it might be possible that the high G+C genes had increased levels of these nucleotides at their 5′ ends only, resulting in an amino acid bias that is concentrated at the amino-terminal of the encoded proteins. We show the results for the patterns of amino acid composition in the high G+C and low G+C rice genes (fig. 4*A*). These results at the protein level reflect the underlying patterns of nucleotide composition. They illustrate that all rice genes tend to have especially elevated levels of G+C-rich codons (encoding the G, A, R, and P amino acids) at their 5′ ends but that the high G+C class is characterized by a tendency to have this elevated level extend over the entire coding sequence. In summary, we found that the differences in nucleotide composition between rice genes are caused by a combination of a gradient along the gene length (as noted by Wong et al. [2002]) and an overall average difference between the genes (as noted by Carels and Bernardi [2000]). Neither the compositional gradient along the coding sequence length, nor the bimodal distribution of nucleotide composition among genes is seen in the *Arabidopsis* genome.

The existence of a compositional gradient along the coding sequence of individual rice genes suggests that the forces acting to increase the G+C level of rice genes since their divergence from their angiosperm ancestors is somehow linked to the orientation of gene transcription. This led us to wonder whether there might be some strand asymmetry in the pattern of bias, as seen in some prokaryotic and organelle genomes (Lobry 1996; Morton 1999; Tillier and Collins 2000). We investigated this by calculating the frequencies of individual nucleotides, rather than the sum of G plus C. The result (not shown) revealed no indication of strand asymmetry. The levels of both G and C were equally elevated in the G+C-rich genes and the levels of A and T were equally reduced. To our surprise, however, we did find a very strong link between gene length and G+C content (fig. 4*B*). Specifically, the high G+C rice genes are much shorter, on average, than the low G+C genes. One's first thought is that the increased nucleotide bias might somehow act to decrease gene length. This unlikely scenario can quickly be discounted, however, by noting that the difference in gene length is equally impressive for the *Arabidopsis* homologs, all of which have virtually identical nucleotide contents. The most parsimonious explanation is that shorter genes

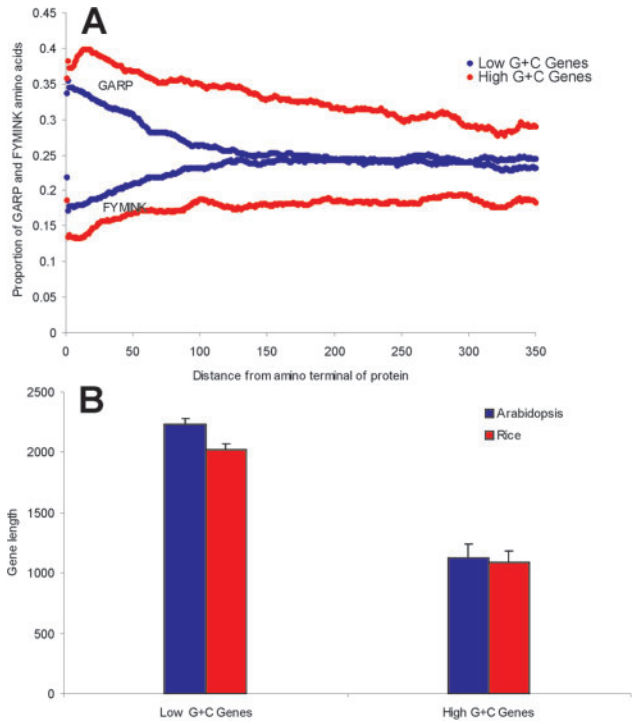

FIG. 4.—The degree of mutational bias correlates with position within the gene and with coding sequence length. (*A*) There is a gradient in the frequency of GARP (top) and FYMINK (bottom) amino acids along the encoded proteins. A sliding window of 17 amino acids was used to score the frequency of GARP and FYMINK along the first 350 amino acid positions. (*B*) The relationship between coding sequence length and G+C content. On the left, the average lengths and 99% confidence intervals are shown for the low G+C rice genes and their *Arabidopsis* homologs. The corresponding data for the high G+C rice genes are shown on the right.

are more susceptible to whatever mutational forces are causing some of the rice genes to become G+C rich.

Discussion

Our results show a clear correlation between the variations in nucleotide composition of different rice genes and the evolutionary changes in the amino acid composition of their encoded proteins. Such a correlation could reflect either a primary effect at the level of nucleotide bias that produces a secondary effect at the protein level or, alternatively, selection for amino acid content at the protein level. The first indication that mutational bias at the nucleotide level is, indeed, the primary cause comes from the observation that the differences in G+C content are greatest at the third codon position (table 1). If the changes in average nucleotide content were a primary effect at the amino acid level, we would expect that the greatest change would be at the first and second codon positions. A related method for distinguishing between nucleotide-level and protein-level effects is to compare the calculated rates of synonymous and nonsynonymous nucleotide substitutions. We used the method of Yang and Nielsen (2000) to calculate these rates for the two groups of rice genes (high G+C and low G+C) compared with their *Arabidopsis* homologs. If the nucleotide composition of the high G+C

Table 2
Exon-Intron Structure of Rice Genes and Their *Arabidopsis* Homologs

	Number of Exons			
	1	2	3	4+
High G+C genes[a] (n = 1000)				
Rice	434	294	165	107
Arabidopsis	308	224	182	286
Low G+C genes[a] (n = 1000)				
Rice	260	72	165	503
Arabidopsis	159	89	57	695

NOTE.—Two general trends can be seen: rice genes tend to have fewer exons, on average, than their *Arabidopsis* homologs, and the high G+C rice genes—and their *Arabidopsis* homologs—have fewer exons than the low G+C genes. These differences are highly significant ($P < 0.0001$ in a chi-square test).

[a] "High G+C" and "low G+C" refers to the nucleotide content of the rice genes only, not to their *Arabidopsis* homologs. The numbers in the table refer to the number of genes that fall into each exon class.

rice genes is affected primarily by selection at the protein level, we should see elevated rates of nonsynonymous changes. If, on the other hand, the primary effect is at the nucleotide level, we should see an elevation in the synonymous substitution rate. The results show very clearly that the increase in substitution rate happens at the synonymous sites, where there is a twofold increase relative to the rate for the low G+C genes (the average values of d_S are 7.7 ± 0.6 and 3.6 ± 0.1 for the high G+C and low G+C genes, respectively). The nonsynonymous substitution rate remains relatively constant between the two sets of genes (average value of $d_N = 0.5$ for both groups). This points to mutational bias at the nucleotide level, rather than functional selection at the protein level.

Our finding that shorter coding sequences have a greater tendency to increase in G+C content confirms the findings of Carels and Bernardi (2000), and it is reminiscent of the finding of Duret et al. (1995) who showed that the G+C content of many vertebrate genes was negatively correlated with coding sequence length. More recently, a similar trend has been noted in a survey of single-exon coding sequences (Xia et al. 2003). It is intriguing to observe the same length correlations in both vertebrates and plants. One possible explanation for this trend is that the increased G+C content is linked to the absence of introns, if one assumes that longer genes are more likely to have multiple exons. However, we tested this hypothesis by confining our analysis to single-exon genes only, and we found that the presence of introns was not the primary determining factor of nucleotide content. For instance, with this more restricted data set (single exon genes only), the difference in gene length was just as great as that shown in figure 4*B* for all genes. Thus, the length difference is maintained even in the absence of introns. Interestingly, however, we found that the high G+C rice genes included relatively few multiple-exon genes, especially genes with three or more exons (table 2). This suggests that the presence of multiple introns may prevent even short genes from becoming G+C rich. This is supported by the observation that among the low G+C rice genes, the average length of three-exon and four-exon genes is only 60% of the length of one-exon and two-exon genes. In other words, even though the coding sequences of these genes are relatively short, the presence of multiple introns may prevent them from becoming G+C rich. This implies that there are selective constraints related to RNA splicing that counter the effects of mutational bias in these genes. Such a constraint would not, however, explain the fact that long, single-exon coding sequences also remain relatively immune to mutational bias. The answer may lie in the fact that RNA splicing is only one form of RNA processing. In general, it may be that longer genes encode more complex transcripts and proteins that have a greater chance of being functionally disrupted by biased mutational changes. Shorter genes are also at risk, but they provide a smaller target for these mutations and, consequently, they are subject to lesser selective constraint.

Differences in coding sequence length might explain why some, but not all, rice gene are subject to mutational bias. However, they cannot explain the differences in nucleotide content between the two plant genomes. A possible explanation for the intergenomic difference is the fact that the rice genome is much larger than the *Arabidopsis* genome and that it contains more genes (Sasaki et al. 2002). This would lead one to predict that gene families in rice may contain more members than in *Arabidopsis*. In fact, Sasaki et al. (2002) report that a significant number of the genes found on rice chromosome 1 are duplicated and arrayed in tandem. This difference in the size of gene families could affect the nucleotide content of the coding sequences because it has been shown that gene conversion between members of gene families can lead to increasing G+C content of the converted sequences (Hickey et al 1991; Galtier 2003). Another difference between the two genomes is that rice genes appear to have a lower average number of introns per coding sequence than do their *Arabidopsis* homologs (Carels and Bernardi 2000 and see table 2).

In summary, we have shown that mutational bias can have profound effects on the patterns of evolutionary divergence between homologous plant protein sequences. This indicates that mutational bias can be a major determinant of the patterns of protein evolution in eukaryotes. The rice genome does not, however, have a uniformly elevated G+C content among its coding sequences. The result of this heterogeneity in the nucleotide content among the coding sequences is reflected in the very different amino acid compositions among the encoded proteins.

Acknowledgments

This work was supported by a research grant from NSERC Canada to D.A.H. and by scholarships from the University of Ottawa (H-c.W.) and NSERC Canada (G.A.C.S.).

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Bairoch, A., and R. Apweiler. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28**:45–48.

Carels, N., and G. Bernardi. 2000. Two classes of genes in plants. Genetics **154**:1819–1825.

Carels, N., P. Hatey, P. Jabbari, and G. Bernardi. 1998. Compositional properties of homologous coding sequences from plants. J. Mol. Evol. **46**:45–53.

Collins, D. W., and T. H. Jukes. 1993. Relationship between G + C in silent sites of codons and amino acid composition of human proteins. J. Mol. Evol. **36**:201–213.

Duret, L., D. Mouchiroud, and C. Gautier. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol. **40**:308–317.

Foster, P. G., L. S. Jermiin, and D. A. Hickey. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J. Mol. Evol. **44**:282–288.

Galtier, N. 2003. Gene conversion drives GC content evolution in mammalian histones. Trends Genet. **19**:65–68.

Gautier, C. 2000. Compositional bias in DNA. Curr. Opin. Genet. Dev. **10**:656–661.

Hickey, D. A., L. Bally-Cuif, S. Abukashawa, V. Payant, and B. F. Benkel. 1991. Concerted evolution of duplicated protein-coding genes in *Drosophila*. Proc. Natl. Acad. Sci. USA **88**:1611–1615.

Karlin, S., A. M. Campbell, and J. Mrazek. 1998. Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. **32**:185–225.

Kreil, D. P., and C. A. Ouzounis. 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res. **29**:1608–1615.

Li, W. H. 1997. Molecular Evolution. Sinauer, Sunderland, Mass.

Lobry, J. R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13**:660–665.

———. 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene **205**:309–316.

Morton, B. R. 1999. Strand asymmetry and codon usage bias in the chloroplast genome of *Euglena gracilis*. Proc. Natl. Acad. Sci. USA **96**:5123–5128.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: The European molecular biology open software suite. Trends Genet. **16**:276–277.

Sasaki, T., T. Matsumoto, K. Yamamoto et al. (80 co-authors). 2002. The genome sequence and structure of rice chromosome 1. Nature **420**:312–316.

Singer, G. A. C., and D. A. Hickey. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol. Biol. Evol. **17**:1581–1588.

Stoesser, G., W. Baker, A. van den Broek et al. (16 co-authors). 2002. The EMBL nucleotide sequence database. Nucleic Acids Res. **30**:21–26.

Tillier, E. R., and R. A. Collins. 2000. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J. Mol. Evol. **50**:249–257.

Ware, D., P. Jaiswal, J. Ni et al. (12 co-authors). 2002. Gramene: a resource for comparative grass genomics. Nucleic Acids Res. **30**:103–105.

Wilquet, V., and M. Van de Casteele. 1999. The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. Res. Microbiol. **150**:21–32.

Wong, G. K., J. Wang, L. Tao, J. Tan, J. Zhang, D. A. Passey, and J. Yu. 2002. Compositional gradients in Gramineae genes. Genome Res. **12**:851–856.

Xia, X., Z. Xie, and W. H. Li. 2003. Effects of GC content and mutational pressure on the lengths of exons and coding sequences. J. Mol. Evol. **56**:362–370.

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

Yu, J., S. Hu, J. Wang et al. (100 co-authors). 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296**:79–92.

David Irwin, Associate Editor

Accepted August 15 2003