Evolution of Base Composition in T-DNA Genes from *Agrobacterium*[1]

José M. Martínez-Zapater, Antonio Marín,† and José L. Oliver‡*

*Departamento de Protección Vegetal, CIT-INIA; †Departamento de Genética y Biotecnia, Universidad de Sevilla; and ‡Departamento de Genética e Instituto de Biotecnología, Universidad de Granada

T-DNA genes on Ti and Ri plasmids from *Agrobacterium* are replicated and repaired in bacteria but expressed in plant cells. Therefore, they can be useful tools to disclose the relative roles played by the two main mechanisms involved in the evolution of DNA base composition: (1) mutational bias along DNA replication/repair processes and (2) selective gene expression constraints. We compare the base-compositional features of 15 T-DNA genes with those of (1) other genes located on Ti or Ri plasmids but outside the T-DNA region (non-T-DNA genes) and (2) a sample of nuclear genes from a natural host plant species (tobacco). The similarity in G+C content found between T-DNA and plant genes at replacement sites, as well as the similar stronger avoidance of CpG at II–III codon positions, support an ancestral plant origin for T-DNA genes. When G+C content and codon usage are considered, T-DNA genes are more similar to non-T-DNA genes than to those of plants, indicating that the mutational bias along replication and repair processes in bacteria is the major factor driving the global compositional properties of T-DNA genes. However, when the reduction in the available CpG methylation targets and the distribution of these avoidances on the different codon positions are considered, T-DNA genes are more similar to those of plants than they are to the other plasmid genes. The requirements for expression of T-DNA genes in the plant cells would have modulated the compositional features of their sequences, mainly CpG avoidance.

Introduction

DNA base composition can be viewed as the result of a balance between mutation and selection (Bulmer 1988; Shields 1990; Eyre-Walker 1991). Directional mutation pressure (Sueoka 1962, 1992; Jukes and Bhushan 1986; Osawa et al. 1988; Wada et al. 1991) is recognized as the major factor driving DNA base composition, either increasing or decreasing GC content. Thus, either the differential mutational bias of DNA polymerases and mismatch repair mechanisms in germ-line cells (Filipski 1987) or the variation in mutation patterns in different chromosomal regions in the germ line (Wolfe et al. 1989) is considered to be responsible for the compositional differentiation among the different genome regions within the eukaryotic genome. On the other hand, selective constraints can also modulate overall DNA base composition. Thus, in *Escherichia coli* (Ikemura 1981; Gouy and Gautier 1982) and yeast (Bennetzen and Hall 1982; Ikemura 1982; Sharp and Li 1986), a relationship was found

1. Key words: base composition, mutational bias, selective gene expression constraints, T-DNA, *Agrobacterium*.

Address for correspondence and reprints: José L. Oliver, Departamento de Genética, Facultad de Ciencias, Universidad de Granada, E-18071 Granada, Spain.

Mol. Biol. Evol. 10(2):437–448. 1993.

0737-4038/93/1002-0014$02.00

between codon usage bias and gene expressivity, which would be the result of selection toward an optimization of gene expression. The codon usage of bacteriophage T7 (Sharp et al. 1985) and of the early genes of T4 (Cowe and Sharp 1991) has been found to be selectively adapted for translation by *E. coli* tRNAs, indicating that phage base composition is, to some extent, modulated by selective gene expression pressures. All these examples indicate that, although mutational bias may be the main factor, selective forces related to gene expression could also be contributing to the evolution of overall DNA base composition at coding sequences.

The assessment of the relative roles played by both mutational and selective pressures in the compositional evolution of a given piece of DNA has been often complicated by the fact that in most systems a gene is replicated, repaired, and expressed in the same genomic environment. An appropriate test would preferably use a system in which replication/repair and expression environments were physically separated. Bacterial genes responsible for the crown gall and hairy-root diseases in dicotyledoneous plants could represent such a system. These plant tumors are induced by oncogenic strains of *Agrobacterium tumefaciens* and *A. rhizogenes*, respectively (for recent reviews, see Zambryski et al. 1989; Winans 1992). Virulent bacteria harbor a large plasmid (pTi or pRi), a specific segment of which, known as "T-DNA," is copied and transferred to the plant cell. Expression of T-DNA genes integrated in the plant chromosomes produces altered hormone levels or hormone sensitivity, responsible for the tumorous growth and the synthesis of enzymes involved in the production of opines by the plant cells. These opines not only serve as a carbon and nitrogen source for the bacteria carrying the pTi or pRi plasmid that bears the opine-catabolizing genes, but they also seem to promote conjugal transfer of Ti/Ri plasmids to other *Agrobacterium* cells in the saprophytic soil population. Thus, successful expression of T-DNA genes in the plant cell ensures the survival of the plasmids carrying that T-DNA in the bacterial population. In this way, the evolution of base composition in plasmid T-DNA genes is subjected both to the mutational pressure operating in bacteria (replication/repair environment) and to the selection pressure coming from the plant cells (expression environment). The analysis of T-DNA base composition offers the possibility to identify the relative contributions of mutational and selective forces to the final base composition of a given piece of DNA. A compositional analysis of T-DNA genes can also help to understand the molecular basis of compositional adaptation to eukaryotic cellular environments.

We report here the compositional analysis we have performed on a set of T-DNA genes from different Ti and Ri plasmids of *Agrobacterium*, comparing them to (i) other genes located on pTi or pRi but outside the T-DNA region (in the following, non-T-DNA genes) and (ii) a sample of nuclear genes from a host-plant species (tobacco). The results are discussed in terms of evaluating the role of the different evolutionary mechanisms acting on T-DNA base composition.

Data and Methods

Table 1 shows the accession numbers of the plasmid genes analyzed. Release 29 (December 1991) of the EMBL Nucleotide Sequence Data Library on CD-ROM (Stoehr and Cameron 1991) was searched and yielded 15 T-DNA genes, including three open reading frames whose transcripts have been identified in transformed plant cells but whose functions are unknown. Many other genes on different pTi or pRi but not included in the T-DNA region are available in data banks; 20 of these non-T-DNA genes were randomly selected for this study. Both T-DNA and non-T-DNA

Table 1
Bacterial Genes Used in Present Study

Accession Number	Gene	Plasmid	Protein or Transcript	%G+C Total	%G+C RS	%G+C SS
T-DNA genes:						
K03313	*rolA*	pRiA4	rolA protein	55.1	54.0	57.0
X03433	*rolB*	pRiA4	rolB protein	49.6	47.8	52.9
X03433	*rolC*	pRiA4	rolC protein	52.3	48.2	60.1
K02000	*tmr*	pTiA6NC	tmr protein	50.1	49.4	51.4
K02553	*tms1*	pTiA6NC	tms1 protein	50.7	53.0	46.8
K02554	*tms2*	pTiA6NC	tms2 protein	49.6	49.7	49.4
X00493	*tml*	pTi15955	tml protein	49.0	48.0	50.8
X00493	*ocs*	pTi15955	Octopine synthase	51.4	50.9	52.1
X00493	*orf1*	pTi15955	Transcript 5	42.8	44.4	39.7
X00493	*orf3*	pTi15955	Transcript 7	43.0	42.0	44.9
X00493	*orf24*	pTi15955	1.6-kb RNA	61.0	50.7	51.5
J01541	*nos*	pTit37	Nopaline synthase	49.0	48.2	49.9
M16877	*6a*	pTit37	6a protein	48.8	47.6	51.0
M16877	*6b*	pTit37	6b protein	48.7	49.5	47.3
X56185	*iaaM*	Tm4	Tryptophane monooxygenase	49.9	50.3	49.1
Mean				50.1	48.9	50.3
Standard error				1.1	0.8	1.2
Non-T-DNA genes:						
X04833	*repA*	pRiA4b	Replicator region	53.8	50.3	60.6
X04833	*repB*	pRiA4b	Replicator region	57.0	55.2	60.2
X04833	*repC*	pRiA4b	Replicator region	56.9	54.5	61.4
M15814	*virEa*	pTiC58	9-kDa virulence protein	56.4	61.9	46.7
M15814	*virEb*	pTiC58	7.1-kDa virulence protein	45.8	40.0	56.7
M15814	*virEc*	pTiC58	63.5-kDa virulence protein	50.3	49.4	52.0
M11311	*hdv13*	pTiC58	13-kDa hdv virulence protein	53.0	51.9	55.1
M11311	*hdv15*	pTiC58	15-kDa hdv virulence protein	55.1	52.9	59.2
M11311	*hdv28*	pTiC58	28-kDa hdv virulence protein	60.0	60.4	58.4
M11311	*hdv29*	pTiC58	29-kDa hdv virulence protein	56.1	57.4	53.9
X05240	*virA*	pTiA6	wide host-range virA	49.8	50.2	49.2
Y00535	*virB*	pTiC58	virB virulence protein	55.2	52.8	59.3
Y00535	*virG*	pTiC58	virG virulence protein	53.8	52.3	56.4
M16397	*virC1*	pTiC58	virC1 virulence protein	54.0	51.7	58.4
M16397	*virC2*	pTiC58	virC2 virulence protein	55.8	54.6	58.0
M14762	*virD1*	pTiA6NC	virD1 virulence protein	53.2	51.4	56.1
M14762	*virD2*	pTiA6NC	virD2 virulence protein	55.0	56.6	52.1
X15884	*argin*	pTiC58	Arginase	60.7	57.8	66.1
X13981	*virF*	pTi15955	virF virulence protein	48.8	51.1	44.9
X02423	*tzs*	pTit37	trans-zeatin synthase	54.0	54.7	52.5
Mean				54.2	53.4	55.9
Standard error				0.8	1.0	1.2

gene sets include genes from Ti (nopaline and octopine) and Ri plasmids. The base compositions of 32 nuclear genes from the solanaceous *Nicotiana tabacum* (tobacco) were also used for comparison (table 2). Solanaceous plants are often infected by *Agrobacterium* in nature.

The 5′ and 3′ flanking regions from all the genes, as well as the introns from the plant genes, were removed before any calculations were made. Nucleotide sites subject

Table 2
Tobacco Genes Used in Present Study

Accession Number	Protein or mRNA	%G+C		
		Total	RS	SS
X07644	Acetolactate synthase A	48.1	51.5	42.0
X07645	Acetolactate synthase B	48.9	51.8	43.8
X02868	β ATPase	48.4	53.4	39.0
M21397	a/b-binding protein C	50.1	54.1	42.2
M21398	a/b-binding protein E	50.8	54.8	42.9
M14417	Chloroplast GAPDH A subunit	50.0	51.1	47.8
M14418	Chloroplast GAPDH B subunit	45.3	49.2	38.0
M14419	Cytosolic GAPDH	46.1	48.3	41.8
M23120	β-glucanase mRNA	42.3	47.1	33.6
M24600	Glycoprotein 2	40.2	40.6	39.3
M19700	Glycine-rich protein	48.8	55.4	36.6
M29279	Osmotin	47.0	53.3	35.3
M29274	par mRNA	40.6	42.6	36.9
X05959	Pathogenesis PR-1a protein	45.0	51.3	32.6
X03465	Pathogenesis PR protein	44.6	49.4	34.9
J01308	RuBisCo small subunit	48.6	47.9	50.0
X03913	TMV-induced protein homologous to thaumatin	47.0	50.7	39.9
J02979	Lignin-forming peroxidase mRNA	37.3	48.4	17.1
X53011	Ubiquitin carboxyl extension protein	58.9	50.7	74.7
X58527	Thioredoxin	47.0	48.1	44.8
X16938	Chitinase	51.0	56.3	40.7
X52283	Anther-specific gene TA-29	46.9	55.2	30.9
X60060	Luminal binding protein	46.7	46.9	46.4
M61904	5-Enolpyruvylshikimate-3-phosphate synthetase	45.0	50.1	35.6
X13885	Extensin	55.1	71.7	24.8
X14058	Nitrate reductase	45.3	46.9	42.4
X55354	Photosystem II 23-kDa polypeptide	44.9	49.5	36.4
X59016	Phosphoenolpyruvate carboxylase	45.6	49.5	38.5
X61102	Pectate lyase	43.6	45.2	40.6
X54855	Root-specific membrane channel protein	46.6	53.9	33.0
X56268	Auxin-induced protein (pGNT1)	41.4	44.9	34.9
M64261	3-Deoxy-D-arabino-heptulosonate-7-phosphate synthase	46.4	50.5	38.8
Mean		46.7	50.6	39.3
Standard error		0.7	0.9	1.6

to silent changes [silent sites (SS)] were considered as the following: A in third positions of all codons plus A in first positions of AGR codons; C in third positions of all codons plus C in first positions of CTR and CGR codons; G in third positions of all codons minus G in third positions of ATG and TGG codons; and T in third positions of all codons plus T in first positions of TTR codons [N = A, C, G, or T(U); R = A or G; and Y = C or T(U)]. All other sites are replacement sites (RS; Jukes and Bhushan 1986).

The codon-usage frequencies were determined according to a method described elsewhere (Oliver et al. 1990). We define as codon groups the sets of synonymous codons differing only in the third nucleotide. We exclude from the analysis termination codons and single-codon groups (methionine and tryptophan), which leaves 21 codon groups comprising 59 codons. Then, we count codon appearances in each gene and

compute the relative frequency of each codon in each of the codon groups (the count of that codon divided by the total of codons in its group); this method draws attention to the specific choices among different options (the synonymous codons), made by the species regardless of the frequencies of the different amino acids in their proteins.

Correspondence analysis is a statistical tool introduced by Benzécri (1976, pp. 1–616), and general presentations of it can be found in papers by Hill (1974) and Lébart et al. (1984, pp. 222–256). This method has been introduced into the molecular biology field by Grantham et al. (1980) and is commonly used to analyze global codon-usage differentiation among and within gene sets (Shields et al. 1988; Oliver et al. 1990; Cowe and Sharp 1991). The FORTRAN program of Lébart and Fenelon (1975, pp. 305–307) was used to carry out a correspondence analysis on the codon-usage frequencies in the three gene sets analyzed here. Only the $n - 1$ independent frequencies in each codon group of each gene were used for the analysis. The program BMDP3S from the BMDP statistical package (Dixon and Brown 1979) was used to carry out several Mann-Whitney nonparametric tests among the different gene sets.

Results

To disclose the factors involved in the evolution of T-DNA base composition, we need to compare T-DNA genes to two other gene sets: (1) genes subjected to the same replication/repair conditions and (2) genes expressed in the same environment. Plant genes are suitable as the second set. For the first comparison, the choice could be between either other plasmid genes or the chromosomal *Agrobacterium* genes. We have analyzed the 12 *Agrobacterium* chromosomal genes contained in release 29 of the EMBL nucleotide data base and have found an average G+C content of 70.6% at SS, a figure considerably different from that found in plasmid genes, in which it is 50%–55% (see table 1). Such difference between plasmid and chromosomal base compositions, which is known to occur also in other species (Data 1985, pp. 3–16), led us to use the non-T-DNA plasmid genes as the proper counterpart for the first comparison.

Nucleotide Composition

The G+C contents in the entire sequence as well as at replacement and silent codon positions in the samples of T-DNA, non-T-DNA, and tobacco genes are given in tables 1 and 2. Despite the inclusion of genes from different Ti (nopaline and octopine) and Ri plasmids, variation in G+C contents within T-DNA and non-T-DNA gene sets was small, compared with the wider differences between the two gene sets (see below). Both T-DNA and non-T-DNA genes show slightly higher G+C contents at SS than at RS; however, in tobacco the G+C contents at SS are much lower than those at RS. Both T-DNA and non-T-DNA have higher G+C contents at SS than do plant genes.

By using the Mann-Whitney nonparametric test, we have made all pairwise comparisons of G+C contents, at both replacement and silent codon positions, among the three gene sets. Between T-DNA and tobacco gene sets there are no differences in the G+C content at RS sites ($P = 0.23$), while the differences at silent sites are significant ($P < 10^{-4}$). The differences are also significant when G+C content at RS in T-DNA is compared with that at RS in non-T-DNA ($P < 10^{-3}$) and when G+C content at RS in non-T-DNA gene sets is compared with that at RS in plant ($P < 10^{-2}$) gene sets.

Codon-Usage Differentiation

Codon usages of T-DNA and non-T-DNA genes have been summarized elsewhere (Winans et al. 1987), but, to our knowledge, no comparison among these two gene sets and the potential host-plant genomes has been made.

Figure 1 shows the results of the correspondence analysis applied to the 59 codon-usage frequencies of T-DNA, non-T-DNA, and tobacco genes. Axis 1 roughly corresponds to G+C content at silent codon positions, which agrees with the results found by Grantham et al. (1980) for mRNA sequences from diverse sources. The rightward genes of figure 1 (T-DNA and non-T-DNA genes) often use G or C at SS, while the richest AT genes (i.e., the plant genes) are situated leftward.

Correspondence analysis discriminates between plasmid and plant gene sets, indicating that global codon usage is more similar within plasmid genes than between plasmid and plant genes. It is interesting that T-DNA genes are closer to plant genes than are non-T-DNA genes, thus pointing to some influence of plant genomic conditions on the compositional features of T-DNA genes. Using the codon-usage data tabulated by Wada et al. (1990) and performing similar correspondence analyses, we extended the comparison of T-DNA and non-T-DNA codon usages to other *Agrobacterium* natural host-plant species: *Arabidopsis thaliana* (23 genes), pea (21 genes), tomato (15 genes), and soybean (34 genes). In every case, T-DNA genes were also grouped with non-T-DNA genes, while plant genes stood apart. Furthermore, T-DNA genes were always nearer to plant genes than were the non-T-DNA genes (results not shown).

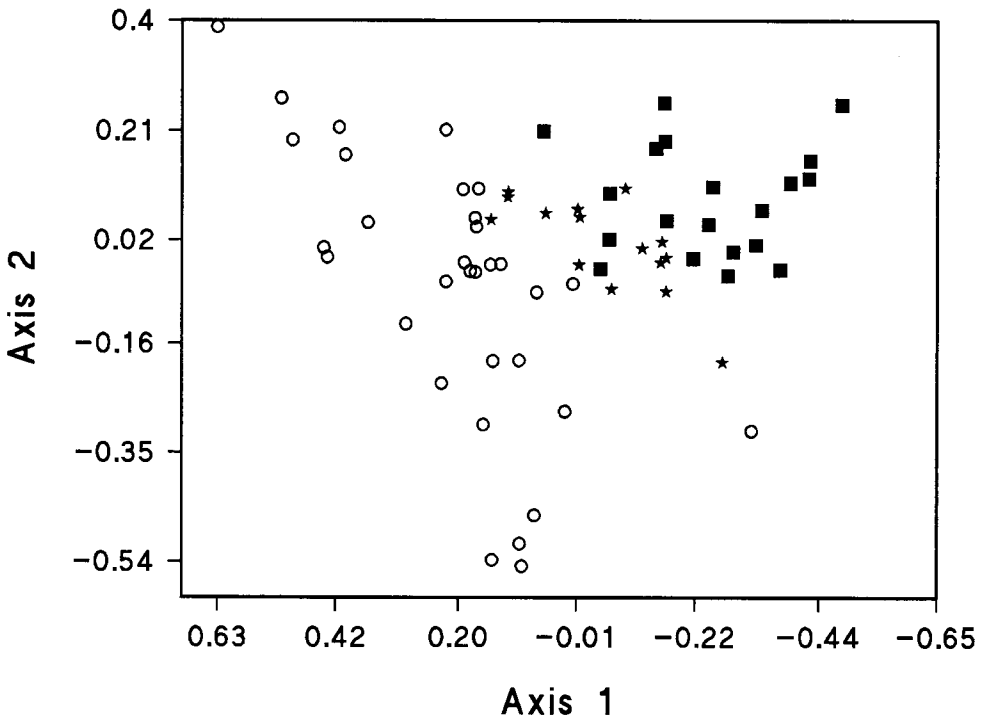

FIG. 1.—Correspondence analysis on codon-usage frequencies in T-DNA (*), non-T-DNA (■), and tobacco (○) genes. Axis 1 denotes the 17.9% variability in codon usage exhibited by these genes, and axis 2 denotes the 9.6% variability.

Distribution of CpG Dinucleotides

Table 3 shows the ratios of the observed number versus the expected number of CpG dinucleotides found in T-DNA, non-T-DNA, and tobacco gene sets. All pairwise comparisons among the three gene sets, as assayed by means of the Mann-Whitney nonparametric test, were statistically significant (results not shown). Tobacco shows CpG dinucleotide avoidance at all three pairs of codon positions tested, which was in agreement with the well-known CpG avoidance in plant genomes (Gruenbaum et al. 1981). Such avoidance was strongest at codon positions II–III. Non-T-DNA genes do not show CpG avoidance, while T-DNA genes show a shortage of CpG dinucleotides at codon positions II–III, having CpG ratios intermediate between those found in non-T-DNA genes and those found in plant genes.

The distribution of CpG doublets for two homologous genes (*tzs* and *tmr;* Beaty et al. 1986; Weiler and Schröder 1987) that occur in non-T-DNA and T-DNA regions, respectively, is also shown in table 3. They have similar nucleotide and amino acid sequences, providing an opportunity to compare CpG content with nucleotide substitutions. The alignment between *tmr* and *tzs* (Beaty et al. 1986) shows that base substitutions have affected replacement positions: 12 of 20 arginine CGN codons in *tzs* are aligned with nonarginine codons in *tmr.* Similarly, for silent positions, the number of NCG codons in *tzs* is double that in *tmr.*

The variation of CpG content in the different gene sets is also illustrated by the coding indexes given in table 4. The first index is the percentage of arginine residues encoded. The second is the ratio of the number of arginine residues encoded by the quartet, codons beginning with CG, to that encoded by the duet, beginning with AG. The increase in this index is known to go along with the increase in G+C content in other eukaryotic sequences (Marín et al. 1989). The strength of CpG doublet avoidance in codon positions II–III can be measured by computing the ratio of XCG to XCC (Grantham 1978). Finally, the ratio of arginine to lysine has been computed. The content of lysine decreases as G+C content increases, whereas arginine increases and compensates for the decrease of lysine (Hanai and Wada 1990). T-DNA and non-T-DNA gene sets show different values for all coding indexes. T-DNA always shows values intermediate between those of non-T-DNA genes and those of plant genes, except for the low arginine content, in which respect it is similar to plants.

Discussion

Origin of T-DNA Genes

We have found that plant and T-DNA genes have (i) similar base composition at RS and (ii) similar strong avoidance of CpG at codon positions II–III. These results support an ancestral plant origin for T-DNA genes. This ancestral origin had been suggested elsewhere, because the regulatory sequences of T-DNA genes are typical of eukaryotic genes and are functional in plant cells (Chilton et al. 1977; Bevan and Chilton 1982) and because the G+C content of TL and TR regions of octopine Ti plasmids are more similar to plant than to bacterial G+C content (Barker et al. 1983). Arguments favoring a prokaryotic origin of T-DNA genes (Weiler and Schröder 1987) have been based on the discovery of a T-DNA related gene in *Pseudomonas savastoni* (Powell and Morris 1986). However, in light of the promiscuity of bacterial plasmids (for a review, see Mazodier and Davies 1991), the spreading of *Agrobacterium* T-DNA genes to other bacteria could be a simpler explanation. This seems to be the case for a sequence from *Bradyrhizobium japonicum* showing extensive similarity to

Table 3
Ratios of Observed versus Expected Number of CpG Dinucleotides in Three Codon-defined Positions of Plasmid and Plant Genes

Gene or Gene Set	No. of Codons	Codon Positions I–II				Codon Positions II–III				Codon Positions III–I			
		No. of CpG Dinucleotides				No. of CpG Dinucleotides				No. of CpG Dinucleotides			
		Observed	Expected	Ratio	χ^2	Observed	Expected	Ratio	χ^2	Observed	Expected	Ratio	χ^2
Non-T-DNA	5,749	406	298	1.36	39.1***	449	409	1.10	3.9*	606	557	1.09	4.3*
T-DNA	4,726	179	192	0.93	0.9	222	267	0.83	7.6**	411	435	0.94	1.3
Tobacco	11,632	274	429	0.64	56.0***	246	576	0.43	189.1***	482	858	0.56	164.8***
tzs	244	20	16	1.25	1.0	22	16	1.38	2.3	24	21	1.14	0.4
tmr	241	10	12	0.83	0.3	12	15	0.80	0.6	19	17	1.12	0.2

NOTE.—Data are presented for 15 T-DNA genes, 20 non-T-DNA genes, and 32 tobacco genes. The observed and expected values in each gene were pooled to get the ratios in each gene set. The ratios found in *tzs* and *tmr* are also shown. We closely followed Shpaer and Mullins (1990) to compute the expected values. For example, there are a total of 241 codons in *tmr;* 28% of 241 codons start with "C," and 18% have "G" in the second position, so the expected frequency of CpG in codon positions I–II is $0.28 \times 0.18 \times 241 = 12$. There are, however, only 10 CpG dinucleotides in codon positions I–II of this gene, so the ratio is $10/12 = 0.83$.

* $P < 0.05$.

** $P < 0.01$.

*** $P < 0.001$.

Table 4
Coding Indexes Illustrating Variation in CpG Dinucleotide Content in Different Gene Sets and in *tzs* and *tmr*

	Non-T-DNA	T-DNA	Tobacco	*tzs*	*tmr*
Arginine content	8.6	5.9	4.8	9.50	5.40
Arg4/Arg2	4.6	1.8	0.9	6.67	3.33
XCG/XCC	1.1	0.8	0.4	1.57	0.75
Arg/Lys	2.4	1.4	0.9	3.83	1.30

the T-DNA of *A. rhizogenes:* several lines of evidence indicate that this sequence is contained within a typical prokaryotic insertion sequence (Ramseier and Göttfert 1991).

Evolutionary Forces Driving Base Composition

From the results noted above, we interpret the similarities between T-DNA and plant genes as the result of the selective maintenance of features that were common in origin. Therefore, by analyzing the base-composition features that have been maintained and those that are now more similar to the bacterial counterparts, we will be able to distinguish between the effect of replication/repair bias and gene expressivity selection on base composition.

During DNA Replication/Repair

At the replication/repair level, the most conspicuous effect of the bacterial mutational bias on T-DNA nucleotide composition is the similar higher G+C content at SS of both T-DNA and non-T-DNA genes, compared with that in plant genes (tables 1 and 2 and fig. 1). These results support the fundamental role of mutational bias in driving the overall composition of DNA sequences (Sueoka 1962, 1992; Jukes and Bhushan 1986; Filipski 1987; Wolfe et al. 1989). The strength of mutational bias in homogenizing base composition suggests that compositional features that remain unchanged between T-DNA genes and plant genes will be so because there is a selective constraint to the compositional change.

During Gene Expression

The influence of selective expression constraints on T-DNA base composition, coming from the plant environment, is apparent from the following: (i) the similarity, in G+C composition at RS, shown by T-DNA and plant genes (tables 1 and 2); (ii) the codon-usage correspondence graph (fig. 1), showing T-DNA genes to be closer to the plant genes than are the non-T-DNA genes; and (iii) the relative avoidance of CpG dinucleotides found in the T-DNA gene set, as well as the distribution of this avoidance among the different codon positions (tables 3 and 4). All these features differentiate T-DNA from non-T-DNA and make T-DNA genes more similar to plant genes. The fact that all these features are maintained in T-DNA genes suggests that they play a significant role in producing a high expression efficiency in the plant cell.

These conclusions, reached at the entire-gene-set level, also explain the results on the evolution of the individual gene sequences *tmr* (T-DNA gene) and *tzs* (non-T-DNA gene). Consequently, because they are replicated and repaired in the same bacterial plasmid, they show similar G+C content at SS (table 1). However, in spite

of their common evolutionary origin, they differ in (i) G+C content at RS (table 1), (ii) levels of CpG avoidance, and (iii) coding indexes (table 4). All these differential features can reasonably be related to differential gene expression optimization pressures. Given the eukaryotic ancestral origin of T-DNA (see above), perhaps *tzs* moved out of the T-DNA region and ceased to be expressed in the plant cells. Once *tzs* was released from constraints for plant cell expression (namely, CpG avoidance), it could accept nucleotide substitutions, which increased the frequency of CpG doublets.

Acknowledgments

Thanks are due to Drs. L. M. Corrochano, G. Visedo, and J. Salinas for helpful comments. This work was partially supported by INIA grant 7556 and DGICYT grant PB90-0847 of the Spanish government.

LITERATURE CITED

BARKER, R. F., K. B. IDLER, D. V. THOMPSON, and J. D. KEMP. 1983. Nucleotide sequence of the T-DNA region from the *Agrobacterium tumefaciens* octopine Ti plasmid pTi15955. Plant Mol. Biol. **2**:335–350.

BEATY, J. S., G. K. POWELL, L. LICA, D. A. REGIER, E. M. S. MACDONALD, N. G. HOMMES, and R. O. MORRIS. 1986. *Tzs*, a nopaline Ti plasmid gene from *Agrobacterium tumefaciens* associated with *trans*-zeatin biosynthesis. Mol. Gen. Genet. **203**:274–280.

BENNETZEN, J. L., and B. D. HALL. 1982. Codon selection in yeast. J. Biol. Chem. **257**:3026–3031.

BENZÉCRI, J. P. 1976. L'analyse des données. Dunod, Paris.

BEVAN, M. W., and M. D. CHILTON. 1982. T-DNA of the *Agrobacterium* Ti and Ri plasmids. Annu. Rev. Genet. **16**:357–384.

BULMER, M. 1988. Are codon usage patterns in unicellular organisms determined by selection-mutation balance? J. Evol. Biol. **1**:15–16.

CHILTON, M.-D., M. H. DRUMMOND, D. J. MERLO, D. SCIAKY, A. L. MONTOYA, P. M. GORDON, and E. W. NESTER. 1977. Stable incorporation of plasmid DNA into higher plant cells: the molecular basis of crown gall tumorigenesis. Cell **11**:263–274.

COWE, E., and P. M. SHARP. 1991. Molecular evolution of bacteriophages: discrete patterns of codon usage in T4 genes are related to the time of gene expression. J. Mol. Evol. **33**:13–22.

DATA, N. 1985. Plasmids as organisms. Pp. 3–16 *in* D. R. HELINSKI, S. N. COHEN, D. B. CLEWELL, D. A. JACKSON, and A. HOLLAENDER, eds. Plasmids in bacteria. Plenum, New York and London.

DIXON, W. J., and M. B. BROWN. 1979. BMDP-79 Biomedical computer programs, P series. University of California Press, Berkeley.

EYRE-WALKER, A. C. 1991. An analysis of codon usage in mammals—selection or mutation bias. J. Mol. Evol. **33**:442–449.

FILIPSKI, J. 1987. Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett. **217**:184–186.

GOUY, M., and C. GAUTIER. 1982. Codon usage in bacteria: correlation with gene expressitivity. Nucleic Acids Res. **10**:7055–7074.

GRANTHAM, R. 1978. Viral, prokaryote and eukaryote genes contrasted by mRNA sequence indexes. FEBS Lett. **95**:1–11.

GRANTHAM, R., C. GAUTIER, M. GOUY, R. MERCIER, and A. PAVE. 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. **8**:r49–r62.

GRUENBAUM, Y., T. NAVEH-MANY, H. CEDAR, and A. RAZIN. 1981. Sequence specificity of methylation in higher plant DNA. Nature **292**:860–862.

HANAI, R., and A. WADA. 1990. Doublet preference and gene evolution. J. Mol. Evol. **30**:109–115.

HILL, M. O. 1974. Correspondence analysis: a neglected multivariate method. Appl. Stat. **23**:340–354.

IKEMURA, T. 1981. Correlation between the abundance of *E. coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. **146**:1–21.

———. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. J. Mol. Biol. **158**:573–597.

JUKES, T. H., and V. BHUSHAN. 1986. Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. J. Mol. Evol. **24**:39–44.

LÉBART, L., and J. P. FENELON. 1975. Statistique et informatique appliquées. Dunod, Paris.

LÉBART, L., A. MORINEAU, and K. A. WARWICK. 1984. Multivariate descriptive statistical analysis. John Wiley & Sons, New York.

MARÍN, A., J. BERTRANPETIT, J. L. OLIVER, and J. R. MEDINA. 1989. Variation in G+C content and codon choice: differences among synonymous codon groups in vertebrate genes. Nucleic Acids Res. **17**:6181–6189.

MAZODIER, P., and J. DAVIES. 1991. Gene transfer between distantly related bacteria. Annu. Rev. Genet. **25**:147–171.

OLIVER, J. L., A. MARÍN, and J. M. MARTÍNEZ-ZAPATER. 1990. Chloroplast genes transferred to the nuclear plant genome have adjusted to nuclear base composition and codon usage. Nucleic Acids Res. **18**:65–73.

OSAWA, S., T. OHAMA, F. YAMAO, A. MUTO, T. H. JUKES, H. OZEKI, and K. UMESONO. 1988. Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. Proc. Natl. Acad. Sci. USA **85**:1124–1128.

POWELL, G. K., and R. O. MORRIS. 1986. Nucleotide sequence and expression of a *Pseudomonas savastanoi* cytokinin biosynthetic gene: homology with *Agrobacterium tumefaciens tmr* and *tzs* loci. Nucleic Acids Res. **14**:2555–2565.

RAMSEIER, T. M., and M. GÖTTFERT. 1991. Codon usage and G+C content in *Bradyrhizobium japonicum* genes are not uniform. Arch. Microbiol. **156**:270–276.

SHARP, P. M., and W.-H. LI. 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res. **14**:7734–7749.

SHARP, P. M., M. S. ROGERS, and D. J. MCCONNELL. 1985. Selection pressures on codon usage in the complete genome of bacteriophage T7. J. Mol. Evol. **21**:150–160.

SHIELDS, D. C. 1990. Switches in species-specific codon preferences: the influence of mutation biases. J. Mol. Evol. **31**:71–80.

SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS, and F. WRIGHT. 1988. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5**:704–716.

SHPAER, E. G., and J. I. MULLINS. 1990. Selection against CpG dinucleotides in lentiviral genes: a possible role of methylation in regulation of viral expression. Nucleic Acids Res. **18**:5793–5797.

STOEHR, P. J., and G. N. CAMERON. 1991. The EMBL data library. Nucleic Acids Res. Suppl. **19**:2227–2230.

SUEOKA, N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA **48**:582–592.

———. 1992. Directional mutation pressure, selective constraints, and genetic equilibria. J. Mol. Evol. **34**:95–114.

WADA, A., A. SUYAMA, and R. HANAI. 1991. Phenomenological theory of GC/AT pressure on DNA base composition. J. Mol. Evol. **32**:374–378.

WADA, K.-N., S.-I. AOTA, R. TSUCHIYA, F. ISHIBASHI, T. GOJOBORI, and T. IKEMURA. 1990. Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res. **18**:2367–2411.

WEILER, E. W., and J. SCHRÖDER. 1987. Hormone genes and crown gall disease. Trends Biochem. **12**:271–275.

WINANS, S. C. 1992. Two-way chemical signaling in *Agrobacterium*-plant interactions. Microbiol. Rev. **56**:12–31.

WINANS, S. C., P. ALLENZA, S. E. STACHEL, K. E. MCBRIDE, and E. W. NESTER. 1987. Characterization of the *virE* operon of the *Agrobacterium* Ti plasmid pTiA6. Nucleic Acids Res. **15**:825–837.

WOLFE, K. H., P. M. SHARP, and W.-H. LI. 1989. Mutation rates differ among regions of the mammalian genome. Nature **337**:283–285.

ZAMBRYSKI, P., J. TEMPE, and J. SCHELL. 1989. Transfer and function of T-DNA genes from *Agrobacterium* Ti and Ri plasmids in plants. Cell **56**:193–201.

WALTER M. FITCH, reviewing editor

Received February 25, 1992; revision received July 1, 1992

Accepted July 1, 1992