J Mol Evol (1994) 39:34–46

Journal of Molecular Evolution

Nucleotide Distribution in Gymnosperm Nuclear Sequences Suggests a Model for GC-Content Change in Land-Plant Nuclear Genomes

S. Jansson[1], G. Meyer-Gauen[2], R. Cerff[2], W. Martin[2]

[1] Department of Plant Physiology, Umeå, Sweden
[2] Institut für Genetik, Technische Universität Braunschweig, Spielmannstr. 7, D-38023 Braunschweig, Germany

Received: 7 June 1993 / Revised: 31 October 1993

Abstract. Nuclear protein coding sequences from gymnosperms are currently scarce. We have determined 4 kb of nuclear protein coding sequences from gymnosperms and have collected and analyzed >60 kb of nuclear sequences from gymnosperms and nonspermatophytes in order to better understand processes influencing genome evolution in plants. We show that conifers possess both biased and nonbiased genes with respect to GC content, as found in monocots, suggesting that the common ancestor of conifers and monocots may have possessed both biased and nonbiased genes. The lack of biased genes in dicots is suggested to be a derived character for this lineage. We present a simple but speculative model of land-plant genome evolution which considers changes in GC bias and CpG frequency, respectively, as independent processes and which can account for several puzzling aspects of observed nucleotide frequencies in plant genes.

Key words: CpG suppression — GC content — Angiosperms — Isochores — GC bias — Mutational pressure — Error-prone repair — Transcriptionally coupled repair

Correspondence to: W. Martin

Abbreviations: GC = guanosine plus cytosine; *GapC* = glycolytic glyceraldehyde-3-phosphate dehydrogenase, EC 1.2.1.12; *GapA* = Calvin cycle glyceraldehyde-3-phosphate dehydrogenase, EC 1.2.1.13; O/E = ratio of observed-to-expected dinucleotide frequencies

Introduction

In the nuclear genomes of monocotyledons, two distinct types of genes are found: GC-biased genes, in which codon usage is strongly biased toward G and C in the third codon position, and nonbiased genes, in which codon usage is relaxed and the four nucleotides appear in approximately equal frequencies in the third codon position (Campbell and Gowri 1990; Brinkmann et al. 1987). In dicotyledons, GC-biased genes are very rare and the codon usage of most genes resembles that of nonbiased genes in monocots (Gardiner-Garden and Frommer 1992). At the whole genome level, both monocots and dicots typically show moderate levels of CpG suppression (Gruenbaum et al. 1981; Hepburn et al. 1987), the cumulative result of spontaneous deamination in 5-methylcytosine (^{5m}C) residues over evolutionary time (Bird 1986). In angiosperms, coding regions and their proximal sequences tend to show lower levels of CpG suppression than the total genome does. Furthermore, the "significant CpG-rich regions" of angiosperm genomes, which are typically associated with genes, contain even higher frequencies of CpG dinucleotides than would be expected on the basis of base composition (Gardiner-Garden and Frommer 1992). This is in contrast to the "CpG islands" of vertebrate genomes, which are characterized by a mere lack of CpG suppression and thus equal frequencies of CpG and GpG dinucleotides (Bird 1986). Interestingly, the differences between monocot and dicot genes with respect to GC content at third positions are very similar to the differences observed between genes of warm and cold-blooded vertebrates (Salinas et al. 1988; Matassi et al. 1989).

In the case of animals, it has been suggested that the ancestral vertebrate genome, or "paleogenome," consisted mostly of nonbiased genes and very few CpG clustered regions, as in cold-blooded vertebrates, whereas the "neogenome," consisting of GC-rich genes and frequently containing CpG islands, appeared recently in evolution (Bernardi 1989, 1993).

Thus, during evolution, genes and genomes in plants have clearly undergone changes with respect to nucleotide composition. We wished to learn something about the mechanisms which might be involved in plant genome evolution and reasoned that outgroup comparison might provide some insights into this problem. We therefore examined nuclear DNA sequences from nonangiosperms using three approaches. First, we determined some new sequences from gymnosperms by isolating and sequencing full-size cDNA clones encoding glyceraldehyde-3-phosphate dehydrogenases from the gymnosperms *Ginkgo biloba, Taxus baccata,* and *Pinus sylvestris.* Second, we screened the database for nuclear sequences from conifers, the only nonangiosperm taxon from which data sufficient to draw conclusions on genome structure has been collected. Third, we examined the scarce sequence data available from nonspermatophytes.

Our results show that gymnosperms possess both unbiased and highly biased genes, as do graminaceous monocots. The data suggest that GC-biased genes were present in ancestral spermatophyte genomes and that the lack of GC bias in dicot genes is a derived character for this lineage. We put forth a simple bimechanistic model for plant genome composition change. Under this model it is suggested (1) that a GC-error-prone DNA repair mechanism may be active in the vicinity of transcribed regions and could therefore be responsible for GC enrichment in biased genes, (2) that in dicots this repair mechanism, if present, is not GC-error prone, and (3) that this mechanism for the introduction of GC bias into plant genes operates independently from CpG suppression, which lowers GC content. This simple but speculative model can account for observed nucleotide patterns in plant genomes and is supported by recent findings (Bootsma and Hoeijmakers 1993; Downes et al. 1993; Selby and Sancar 1993; Schaeffer et al. 1993) which demonstrate the presence of transcriptionally coupled DNA repair systems in eukaryotes.

Materials and Methods

mRNA Isolation. PolyA+ mRNA was purified from endosperm of maturing *Ginkgo biloba* seeds through two rounds of oligo(dT) cellulose (Pharmacia Type 7) chromatography as described (Cerff and Kloppstech 1982) except that a 2.0 M LiCl precipitation (Maniatis et al. 1989) was performed prior to oligo(dT) cellulose binding and that all oligo(dT) cellulose binding, wash, and elution buffers contained 0.01% (w/v) Proteinase K (Merck). PolyA+ mRNA from *Taxus baccata* was isolated from whole immature seeds by the method described (Martin et al. 1993b). Eluates from the second oligo(dT) columns were phenolized, precipitated with ethanol, and collected by centrifugation. mRNA pellets were dissolved in 10 mM Tris-HCl (pH 8.0), 1 mM EDTA to 1 μg/μl, immediately frozen in liquid nitrogen, and stored at −80°C prior to cDNA synthesis.

cDNA Cloning and Sequencing. cDNA from all three gymnosperms was constructed using the Pharmacia kit with modifications as described (Martin et al. 1990). cDNA cloning into λnm1149 and screening of recombinants was performed as described (Schwarz-Sommer et al. 1984). The *Ginkgo biloba* and *Taxus baccata* libraries were screened with an *Eco*RI fragment containing the entire coding region of *GapC* cDNA from *Magnolia liliiflora* (Martin et al. 1989). Positively hybridizing cDNAs were subcloned into SK+ plasmids (Stratagene). Sequences of the plasmids pTBC1, containing a *Not*I insert encoding *GapC* of *Taxus,* and pGB1, containing a *Not*I insert encoding *GapC* of *Ginkgo,* were determined by the dideoxy method. For *Pinus sylvestris,* the cDNA library from 2-week-old seedlings previously described (Martin et al. 1993a) was screened. The *Pinus GapA* clone was identified by plaque hybridization to the *Pst*I insert of pS84b encoding *GapA* of mustard (Martin and Cerff 1986), subcloned into the *Not*I site of pSK+ to yield the plasmid pPSA1, the *Not*I insert of which was sequenced by the chemical degradation method (Maxam and Gilbert 1980).

Sequence Analysis. The 46 gymnosperm sequences used in this study are given in Table 1. Sequences were analyzed using the GCG package (Devereux et al. 1984). Short sequences were excluded from analysis in some figures. (See figure legends.) In total 37,004 bp from coding regions and 23,599 bp from noncoding regions were analyzed. CpG and GpC plots were generated with the MACMOLLY program (Soft Gene GmbH, Berlin, FRG). Codon usage tables for the genes are available from authors upon request. Expected CpG frequencies were calculated as the product of the relative frequencies of C and G for the gene region analyzed. For comparisons between homologous genes of conifers, monocots, and dicots, those genes with known coding regions over 400 bp in length in conifers having known homologous sequences from monocots and dicots were analyzed. When multiple monocot and dicot sequences were available, the tomato and rice sequences were chosen for comparison.

Results and Discussion

New cDNA Clones for Nuclear Encoded Genes from Gymnosperm

Gymnosperm nuclear sequences are currently scarce in the database. Previously, we have analyzed several GAPDH sequences from angiosperms (Martin et al. 1989, 1993a; Brinkmann et al. 1989). Here we chose to investigate their homologs in gymnosperms. Higher plants possess two distinct, nuclear-encoded glyceraldehyde-3-phosphate dehydrogenase proteins, a Calvin cycle enzyme (GAPA) active within chloroplasts, and a glycolytic enzyme (GAPC) active within the cytosol (Cerff 1979; Martin and Cerff 1986; Martin et al. 1993c). We constructed libraries, and isolated and sequenced cDNAs for *GapA* from *Pinus sylvestris* and *GapC* from *Ginkgo biloba* and *Taxus baccata. GapA* and *GapC* are highly expressed genes in both angiosperms and gymnosperms. As measured by the frequency of positively hybridizing clones per recombinant

Table 1. Gymnosperm nuclear sequences surveyed

Species/gene	Gene product	Reference
Pinus sylvestris		
Chs	Chalcone synthasse	Fliegmann et al. 1992
Cp-Sod	Chloroplast superoxide dismutase	Karpinski et al. 1992
Cyt-Sod	Cytosolic superoxide dismutase	Karpinski et al. 1992
GapA	Chloroplast glyceraldehyde-3-phosphate dehydrogenase	This paper
GapC	Cytosolic glyceraldehyde-3-phosphate dehydrogenase	Martin et al. 1993a
*Lhca1*1*	Type I chlorophyll *a/b*-binding protein of LHC I	Jansson and Gustafsson 1991
*Lhca1*2*	Type I chlorophyll *a/b*-binding protein of LHC I	Jansson and Gustafsson 1991
*Lhca2*1*	Type II chlorophyll *a/b*-binding protein of LHC I	Jansson and Gustafsson 1991
*Lhca3*1*	Type III chlorophyll *a/b*-binding protein of LHC I	Jansson and Gustafsson 1991
*Lhca4*1*	Type IV chlorophyll *a/b*-binding protein of LHC I	Jansson (unpublished)
*Lhcb1*1*	Type I chlorophyl *a/b*-binding protein of LHC II	Jansson and Gustafsson 1990
*Lhcb1*2*	Type I chlorophyl *a/b*-binding protein of LHC II	Jansson and Gustafsson 1990
*Lhcb2*1*	Type II chlorophyl *a/b*-binding protein of LHC II	Jansson and Gustafsson 1990
*Lhcb5*1*	Chlorophyll *a/b*-binding protein of CP26	Jansson (unpublished)
Pdc	Pyruvate decarboxylase	Jansson (unpublished)
Sts	Stilbene synthase	Fliegmann et al. 1992; Schwekendiek et al. 1992
Pinus taeda		
Adh	Alcohol dehydrogenase	D. Harry (unpublished)
Lpcr	Protochlorophyllide oxidoreductase	Spano et al. 1992b
Pal	Phenylalanine-ammonium lyase	Whetten and Sederoff 1992
Xsp	Xylem-specific protein	Loopstra (unpublished)
Pinus thunbergii		
*Lhcb2*1*	Type II chlorophyll *a/b*-binding protein of LHC II	Yamamoto et al. 1998a; Kojima et al. 1992
RbcS	Ribulose bisphosphate carboxylase/oxygenase	Yamamoto et al. 1988b
Pinus radiata		
Adh1	Alcohol dehydrogenase	Kinlaw et al. 1990
Adh2	Alcohol dehydrogenase	Kinlaw et al. 1990
Pinus strobus		
Glbn1	Globulinlike protein 1	EMBL Z11486
Glbn2	Globulinlike protein 2	EMBL Z11487
Alb1	Albuminlike storage protein 1	EMBL X62433
Alb2	Albuminlike storage protein 2	EMBL X62434
Alb3	Albuminlike storage protein 3	EMBL X62435
Pinus contorta		
Act	Actin	Kenny et al. 1988
Picea abies		
Cinalda	Cinnamyl alcohol dehydrogenase	EMBL X72675
Gluc	β-1,3-glucanase-like protein	Sundås et al. 1992
H2A	Histone 2A	Sundås (unpublished)
Hbk1	Knotted homologous gene 1	Sundås (unpublished)
Hbk2	Knotted homologous gene 2	Sundås (unpublished)
*Lhcb1*1*	Type I chlorophyl *a/b*-binding protein of LHC II	Sundås (unpublished)
Rps12	Chloroplast ribosomal protein S12	Sundås (unpublished)
Tuba	α-tubulin	Sundås (unpublished)
Ubiq	Polyubiquitin	Sundås et al. 1992
Picea glauca		
Leg	Leguminlike storage protein	EMBL X63192
Vic	Vicilinlike storage protein	Newton et al. 1992
2S	2S-like storage protein	EMBL X63193
Pseudotsuga menziessi		
Lge	Leguminlike storage protein	Leal and Misre 1993
Larix larcinia		
*Lhcb1*1*	Type I chlorophyll *a/b*-binding protein of LHC II	Hutchisson (unpublished)
RbcS	Ribulose bisphosphate carboxylase/oxygenase	Hutchisson et al. 1990
Taxus baccata		
GapC	Cytosolic glyceraldehyde-3-phosphate dehydrogenase	This paper
Ginkgo biloba		
GapC	Cytosolic glyceraldehyde-3-phosphate dehydrogenase	This paper

phage in our cDNA libraries from several gymnosperms and angiosperms, we estimate that *GapA* constitutes about 0.5% of the clonable polyA+ mRNA in light-grown spermatophyte tissues and *GapC* about 0.1%.

Clone pPSA1 contains the entire coding region of the mature GAPA chloroplast GAPDH subunit (337 amino acids) from *Pinus sylvestris* and an N-terminal 74-amino-acid putative transit peptide. The deduced amino acid sequence (Fig. 1) has 81% and 78% identity to its pea and maize counterparts, respectively. The clone pTBC1 contains the entire coding region (341 amino acids) of *GapC* from *Taxus* and the clone pGBC1 contains the coding region for *GapC* from *Ginkgo*. The deduced amino acid sequences of the pTBC1 and pGBC1 clones share 91% identity; these share 91% and 89% identity to *GapC* from Scots pine (Martin et al. 1993a), respectively. These sequences have been deposited in the database under the accession numbers L26923 (pPSA1), L26924 (pGBC1), and L26922 (pTBC1).

Dinucleotide Profiles for Coding Regions

To examine the nucleotide composition fine structure within coding regions, we plotted CpG and GpC profiles for the sequences of gymnosperm *GapC, Lhcb1,* and *GapA* in the context of angiosperm homologs and, in the case of *GapC,* a bryophyte (*Physcomitrella*) outgroup (Fig. 1). In Fig. 1a it can be seen that the GpC profiles for the gymnosperm *GapC* sequences are remarkably similar to their angiosperm and bryophyte homologs, whereas the CpG profiles (and O/E ratios) differ considerably. In the coding regions of *GapC* (Fig. 1a) there is some variation across gymnosperms with respect to GC content. Both *Ginkgo* and *Taxus* show stronger CpG suppression than the homolog from *Pinus* although the third position GC contents in the gymnosperm *GapC* sequences are quite similar (42–47%). *GapC* in both conifers and the dicot is poor in overall GC content and also shows CpG suppression. Since their divergence from a common ancestral gene, CpG frequency in *GapC* coding regions shown in the figure has changed much more dramatically than GpC frequency has. The pattern observed in Fig. 1a favors the view that the degree or rate of CpG suppression for a given gene can vary across plant taxa. As shown in Fig. 1c, *GapA* of *Pinus* is quite GC-poor (42% at third codon positions) and shows slightly more pronounced CpG suppression than the dicot homologue (O/E ratio 0.35 vs 0.52), whereas the maize homolog has both high third-position GC content (97%) and CpG frequency (O/E ratio 1.19). An even more dramatic change in profile can be observed for *Lhcb1* (Fig. 1b), where the rice homolog has a CpG O/E ratio of 1.59 as compared to 0.32 for tobacco. Ratios of observed-to-expected (O/E) frequencies for GpC dinucleotides are close to unity for the sequences investigated in Fig. 1 and, furthermore, for all gymnosperm sequences surveyed here (data not shown), similar to the situation found in angiosperm genes (Gardiner-Garden et al. 1992).

We note that two findings in Fig. 1 would be quite difficult to achieve simply through the effects of lacking CpG suppression in these sequences: (1) the CpG O/E ratio of $\sim$1.6 for rice *Lhcb1* and (2) third-position GC contents in excess of 90% for monocot *Lhcb1* and *GapA*. These straightforward data indicate that some mechanism(s) other than CpG suppression (or lack thereof) may be affecting GC bias in these sequences.

Conifer Nuclear Genes are Heterogeneous with Respect to Third-Position GC Content

There was a large degree of heterogeneity in the third-codon-position GC content of the gymnosperm genes analyzed (Fig. 2). Notwithstanding our small sample, the gymnosperm nuclear genes can very roughly be divided into two classes—a GC-rich class with >60% GC at third positions and a GC-poor class with <60% GC at third positions. This distribution is more or less intermediate with regard to the distributions found in monocots and dicots, respectively (Salinas et al. 1988; Campbell and Gowri et al. 1990). Both the GC-rich and the GC-poor group of conifer genes are somewhat lower in third-position GC content than their graminaceous counterparts. There is a general correlation between decreasing GC content and increasing CpG suppression in gymnosperm nuclear sequences surveyed (Fig. 3; $r^2 = 0.737$). The GC-rich conifer genes have a frequency of CpG close to the expected values whereas the GC-poor genes are depleted in CpG, similar to the situation in monocot genomes. The unusually high frequency of CpG in the gymnosperm storage protein genes (open squares in Fig. 3) can be explained solely by the extremely high arginine content of the encoded proteins (data not shown). The O/E CpG value for the complete set of conifer coding sequences is 0.66, which is intermediate between the values for monocots (0.86) and dicots (0.56) (Gardiner-Garden et al. 1992). We have not attempted to strictly identify "significant CpG-rich regions" using the same criteria as Gardiner-Garden and Frommer (1992) did, but we note that many of the gymnosperm sequences with high-third-position GC content match their criteria. Inspection of the conifer data revealed no correlation between expression level and GC content at third positions nor between gene copy number and GC content at third positions; in addition, both housekeeping and tissue-specific genes were found among the GC-rich and GC-poor conifer genes. These findings suggest the GC bias for a given gene is independent of the nature of the gene product.

With the exception of storage protein genes, we found no gymnosperm sequences with a third-position GC content between 53 and 60% (Table 2). Both GC-rich and GC-poor gymnosperm genes have lower GC content than their monocot homologs, the GC-rich genes

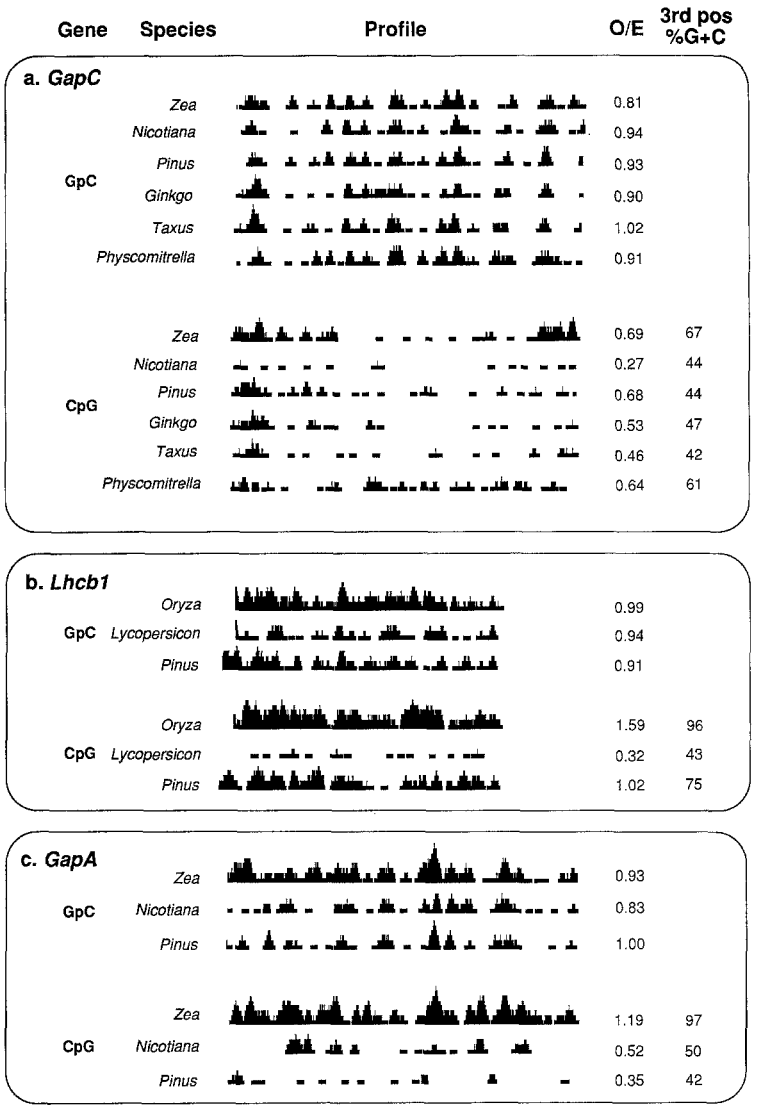

Fig. 1. CpG and GpC profiles of spermatophyte sequences. **a** Coding regions for *GapC* from *P. sylvestris* (Martin et al. 1993a), *T. baccata* (this paper), *G. biloba* (this paper), *P. patens* (Martin et al. 1993a), *Z. mays* (Brinkmann et al. 1987). **b** Mature subunits of *Lhcb1*1* from *P. sylvestris*, rice and tomato. (For references to sequences see Materials and Methods.) **c** Mature subunits for *GapA* from maize, tobacco (for references to sequences see Materials and Methods) and *P. sylvestris* (this paper). One narrow *vertical bar* in dinucleotide profiles corresponds to one dinucleotide per 21-bp window; *vertical increments* correspond to one dinucleotide each. Ratios of observed-to-expected dinucleotide frequencies (O/E) were calculated as described in Materials and Methods.

in gymnosperms having roughly >60% GC at third positions as compared to >90% in monocots and the GC-poor gymnosperm genes having <50% GC at third positions compared to roughly <70% for monocots. These data suggest the existence of an approximately bipartite genome structure in gymnosperms, although with an increasing gene sample, the division between the two types of genes may be expected to become less distinct. It should also be noted that genes can be biased in one end and nonbiased in the other (Martinez et al. 1989; Gardiner-Garden and Frommer 1992).

Comparison of Conifer Coding and Noncoding Regions

In samples of angiosperm nuclear sequences, there is a correlation between the GC content of the coding and noncoding regions of individual genes (Gardiner-Garden and Frommer 1992), in agreement with an isochore organization for angiosperm genomes (Salinas et al. 1988; Matassi et al. 1989). By contrast, if the third-position GC content of the conifer genes is plotted against the GC content of noncoding regions (Fig. 3) we see no such correlation, although our analysis is limited by the small number of nuclear sequences in the sample. The overall GC content of the conifer coding regions is 51%, whereas noncoding regions contain an average of 36% GC. The difference (15%) is very similar to the corresponding values for monocot and dicot genes (17%: Gardiner-Garden et al. 1992).

Comparisons of Homologous Genes from Conifers, Monocots, and Dicots

We compared the third-position GC content in homologous genes from monocots, dicots, and conifers (Table 3). We found that in most cases GC-rich conifer genes (>60% GC at third positions) were highly biased in monocots and that GC-poor conifer genes were low in GC in monocots. A portion of our sample (four out of 11 genes: *GapA, Lhcb5, Pcr,* and *Pal*) was, notably, GC-poor in conifers, yet highly biased in monocots. Thus, homologous functional genes can be either GC-rich or GC-poor, suggesting that transitions between genome

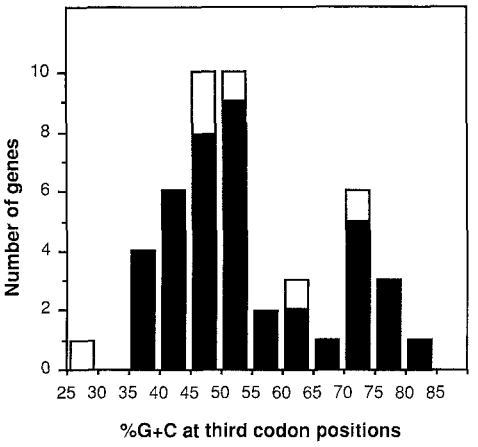

Fig. 2. Histogram depicting the number of gymnosperm nuclear genes with the indicated third-position GC content. Data were taken from Table 2. *White columns:* Sequences with less than 100 codons. *Black columns:* Sequences with over 100 codons.

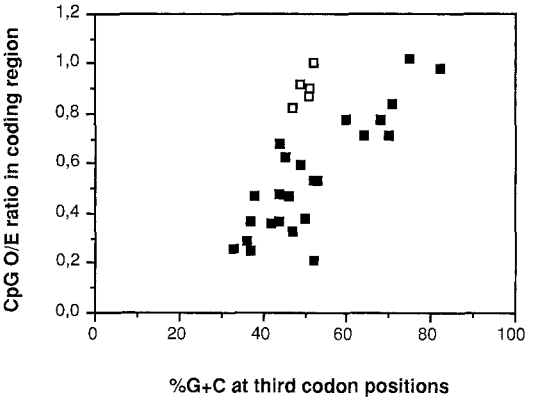

Fig. 3. Observed-to-expected (O/E) ratios of CpG frequency in gymnosperm nuclear genes plotted against the third-position GC content of the gene. Data were taken from Table 2. Coding sequences shorter than 200 codons were excluded. *Open squares:* genes encoding storage proteins. *Filled squares:* genes encoding nonstorage proteins.

compartments may occur independently for individual genes within a bipartite genome. It is possible that coincidence between GC bias of a given gene in monocots and conifers simply reflects the ancesteral state of the gene in the antecedent of the sequences studied.

Analysis of Nonspermatophyte Genes

We gathered and analyzed available sequences from nonspermatophytes (Table 4). Although the sample is extremely small, several points are notable. The data, albeit scarce, suggest that there is no correlation between third-position GC content and CpG frequencies in fern genomes. Also, functional fern *Lhcb* genes show moderate CpG suppression, whereas two of the pseudogenes show a very strong CpG avoidance (Table 3). This suggests that a mechanism exists in ferns, in addition to monocots (Quigley et al. 1989), which maintains CpG content within functional genes, and that pseudogenes become disengated from this mechanism. This would be consistent with the finding that CpG methylation existed in the antecedent of vascular plants (Belanger and Hepburn 1990) and that fern genomes show overall CpG methylation levels similar to angiosperm genomes (McGrath and Pichersky 1991). The difference in third-position GC content between moss *Lhcb* and the other moss sequences (Table 4) hints at a heterogeneous nature of the nuclear genome of mosses. Moreover, the PhyCer sequence appears to be heterogenous in itself; the 5′ portion of the gene reveals a much lower degree of CpG suppression than the 3′ region does.

A Model for the Evolution of Land-Plant Nuclear Genomes

We have here been able to demonstrate similarities and differences in the genome structure of higher land-plant taxa. We have provided evidence (1) that biased/CpG-rich and unbiased/CpG-poor genes are present not only in graminaceous monocot (Salinas et al. 1988) but also in conifer genomes, (2) that transitions may occur from a biased to a non-biased state (or vice versa) for individual genes within a given genome (e.g., *GapA* of *P. sylvestris* vs *GapA* of maize), (3) that, in contrast to angiosperms, there is currently no clear correlation between the GC content of coding and overall noncoding regions of conifer genes, although this preliminary finding is limited by the small sample size and strong underrepresentation of 5′ noncoding sequences in the data set analyzed, (4) that homologues of GC-rich conifer genes are also GC-rich in monocots, (5) that the degree of bias for a particular gene does not appear to relate to the expression characteristics of the gene, (6) that a mechanism maintaining high CpG content of active genes appears to be present in ferns (similar to the situation in monocots: Quigley et al. 1989), and (7) that bryophyte genomes appear to contain both biased and unbiased genes, as do gramniaceous monocots and conifers.

Taken in the context of previously published data, these findings provide enough information to put forth a hypothesis for land-plant nuclear genome evolution. We consider two independent mechanisms: (1) one which introduces GC bias into plant nuclear DNA and, as a mere consequence, CpG dinucleotides, and (2) a second (CpG suppression) which depletes DNA of CpG dinucleotides and, as a mere consequence, GC content. As put forth in the following, this bimechanistic model can account for observed base composition patterns in plants.

The latter mechanism is relatively well characterized. CpG suppression is found to varying extents in different plant genomes (Hepburn et al. 1987; Quigley et al. 1989; Campbell and Gowri 1990; Gardiner-Garden et al. 1992). Lack of suppression has been attributed to both protection against methylation (Bird 1986) as well as specific G-T mismatch repair (Hepburn et al. 1987); for the purpose of our model, a distinction between these

Table 2. Base composition and O/E ratios of CpG frequencies in gymnosperm nuclear sequences

	Coding region				5′ noncoding			3′ noncoding			Introns		
Species/ gene	No. of codons	Total G + C (%)	3rd pos. G + C (%)	CpG O/E ratio	bp	G + C (%)	CpG O/E ratio	bp	G + C (%)	CpG O/E ratio	bp	G + C (%)	CpG O/E ratio
Pinus sylvestris													
Chs	397	54	64	0.71	33	48	0	68	41	1.4	109	35	0.6
Cp-Sod	141	49	36	0.29	—	—	—	250	33	0.6	—	—	—
Cyt-Sod	155	49	33	0.26	112	41	0.4	227	35	0.6	—	—	—
GapA	412	46	42	0.36	33	48	0	185	25	0	—	—	—
GapC	341	46	44	0.68	58	33	0.7	195	33	0.8	—	—	—
*Lhcal*1*	247	49	38	0.47	85	45	0.2	351	36	1.3	—	—	—
*Lhcal*2*	208	48	37	0.37	—	—	—	302	42	0.8	—	—	—
*Lhca2*1*	279	53	52	0.21	46	43	0.5	214	48	0.9	—	—	—
*Lhca3*1*	287	49	47	0.33	75	37	0	219	28	0.2	—	—	—
*Lhca4*1*	251	56	68	0.78	4	25	0	253	34	0	—	—	—
*Lhcb1*1*	279	61	75	1.02	55	53	1.1	191	42	0.3	—	—	—
*Lhcb1*2*	275	59	71	0.85	48	46	0.9	133	47	0.7	—	—	—
*Lhcb2*1*	151	58	72	0.59	—	—	—	128	33	0.3	—	—	—
*Lhcb5*1*	303	50	44	0.37	48	48	0.9	236	37	0.4	—	—	—
Pdc	59	49	46	0.47	—	—	—	—	—	—	—	—	—
Sts	394	53	60	0.78	70	40	0	178	29	0	567	31	1.0
Pinus taeda													
Adh	377	46	44	0.48	352	39	0.8	109	36	0.6	1,326	33	0.3
Lpcr	401	44	37	0.35	1114	39	0.6	308	33	0.5	1,805	34	0.2
Pal	748	49	52	0.53	—	—	—	233	37	0.4	—	—	—
Xsp	218	63	82	0.98	53	53	0.5	300	33	1.1	—	—	—
Pinus thunbergii													
*Lhcb2*1*	267	58	70	0.71	2146	34	0.4	2350	32	0.3	122	43	1.1
RbcS	172	56	72	0.73	32	44	0	299	40	0.5	—	—	—
Pinus radiata													
Adh1	175	48	44	0.34	13	46	0	—	—	—	—	—	—
Adh2	133	47	42	0.51	—	—	—	—	—	—	—	—	—
Pinus strobus													
Glbn1	489	53	52	1.00	119	39	1.3	363	33	0.4	—	—	—
Glbn2	411	51	49	0.92	171	35	0.4	335	32	0.6	—	—	—
Alb1	191	54	51	1.01	262	44	0.3	214	35	0.8	—	—	—
Alb2	172	56	55	1.01	—	—	—	193	37	0.7	—	—	—
Alb3	173	53	50	1.05	—	—	—	163	34	0.8	—	—	—
Pinus contorta													
Act	162	43	39	0.32	—	—	—	125	36	0.6	134	31	0.3
Picea abies													
Cinalda	358	47	49	0.39	57	42	0.4	147	28	0.7	—	—	—
Gluc	53	45	49	0.87	—	—	—	117	44	0.9	—	—	—
H2A	139	50	45	0.63	73	38	0	196	35	1.0	891	34	0.9
Hbk1	98	43	46	0.46	—	—	—	668	38	0.3	—	—	—
Hbk2	168	43	42	0.44	—	—	—	229	32	1.2	—	—	—
Lhcb1	86	59	72	0.82	—	—	—	—	—	—	—	—	—
Rps12	990	41	27	1.12	—	—	—	—	—	—	—	—	—
Tuba	244	56	49	0.60	—	—	—	—	—	—	—	—	—
Ubiq	88	50	60	0.39	—	—	—	—	—	—	—	—	—
Picea glauca													
Leg	510	52	51	0.90	14	64	1.4	177	36	1.3	—	—	—
Vic	449	49	47	0.82	11	45	0	223	37	0.9	—	—	—
2S	173	54	58	1.01	56	41	0.8	120	38	0.7	—	—	—
Pseudotsuga menziessi													
Leg	528	52	51	0.87	40	48	0.7	90	34	1.6	—	—	—
Larix larcinia													
*Lhcb*1*	192	61	77	0.71	—	—	—	7	57	0	—	—	—
RbcS	190	55	73	0.81	1614	39	0.4	1293	34	0.2	203	27	0.5
Taxus baccata													
GapC	341	45	46	0.47	1	0	0	91	30	0.5	—	—	—
Ginkgo biloba													
GapC	341	46	53	0.53	22	40	0	146	36	0.6	—	—	—

Table 3. Third codon position GC content of homologous genes[a]

Gene	Taxon		
	Monocot	Gymnosperm	Dicot
Lhcb1	100	73	47
RbcS	90	72	47
Lhcb2	93	71	38
Chs	94	64	43
Pal	94	52	39
Adh	65	44	38
GapC	67	44	44
Lhcb5	90	44	48
GapA	96	42	50
Act	68	39	39
Pcr	96	37	40
Cyt-Sod	41	33	32

[a] The angiosperm sequences analyzed for comparison to gymnosperm homologues were *Act* from rice (McElroy et al. 1990) and potato (Drouin and Dover 1990), *Adh* from rice (Xie and Wu 1989) and tomato (Van der Straeten et al. 1991), *Chs* from barley (Rhode et al. 1991) and tomato (O'Neill et al. 1990), *Cyt-Sod* from rice (Sakamoto, 1992) and tomato (Perl-Treves et al. 1988), *GapA* from maize (Brinkmann et al 1987) and tobacco (Shih et al. 1986), *GapC* from maize (Brinkmann et al. 1987) and tobacco (Shih et al. 1986), *Lhcb1* from rice (*Lhcb1*1* or 2120, Matsouka 1990) and tomato (*Lhcb1*2* or *Cab-1B*, Pichersky et al. 1985), *Lhcb2* from rice (*Lhcb2*1* or 2123, Matsouka 1990) and tomato (Pichersky et al. 1987), *Lhcb5* from barley (Sørensen et al. 1992) and tomato (Pichersky et al. 1991), *Pal* from rice (Minami et al. 1989) and tomato (EMBL M90692), *Pcr* from barley (Schultz et al. 1989) and pea (Spano et al. 1992a), and *RbcS* from rice (Matsuoka et al. 1988) and tomato (*RbcS2:* Sugita et al. 1987)

is not necessary. Hypermutability of 5-methylcytosine obviously can account both for low O/E CpG ratios observed in land plant genes and for the fact that GpC frequencies in plant genes show no significant fluctuation (Gardiner-Garden et al. 1992; see also Fig. 1). The fact that some genes of gymnosperms (*Lhcb1* and *Xsp,* Table 1) and lower land plants (*Lhcb* and *Phy1,* Table 4) have O/E CpG ratios close to unity whereas their homologs in other taxa may be depleted suggests to us that the ancestral state of land plant genomes was CpG rich. Consistent with this is the finding that the nuclear genes of green algal outgroups such as *Chlamydomonas reinhardtii* show little or no CpG suppression (O/E ratios for CpG of 0.7–1.1, data not shown). Thus, CpG suppression is a mechanism which may reduce GC content and CpG frequencies without heavily influencing GpC frequencies.

Now we consider the more tedious problem of a mechanism which may introduce GC bias into DNA. It is well known that many spermatophyte genes contain up to 100% GC at third codon positions (Table 4; Niesbach-Klösgen et al. 1987; Brinkmann et al. 1987; Salinas et al. 1988). We find it difficult to imagine negative selective pressures which would be sufficiently strong to eradicate all alleles in a plant population which possesses adenine or thymidine substitutions at one single third position of such a coding region; we find it equally difficult to imagine positive selection pressures which would fix guanosine or cytosine substitutions with equal tenacity. Anticodon-condon coevolution as a selective mechanism can be excluded since (1) no known anticodon base pairs with only G or C and (2) GC-richness often extends several hundred base pairs into promotor regions of highly biased genes (e.g., Fig. 4). Fluctuating dNTP pools during DNA replication (Wolfe et al. 1989) were put forth as a nonselective mechanism responsible for introduction of high GC content into DNA. Yet, were this mechanism active in plant genes, one would expect to observe roughly homogeneous GC content across plant replicons or at least strong CpG suppression in those GC-poor regions flanking extremely GC-rich genes; neither the former nor the latter is observed in analyses of dinucleotide frequencies in plant sequences with large flanking regions for study (Fig. 4; see also Gardiner-Garden and Frommer 1992), making it very unlikely in our opinion that mutational bias through replication timing causes high GC content in biased plant genes. Near-perfect G/T mismatch repair and/or low levels of CpG methylation in premeiotic cells of angiosperms were suggested as possible explanations for the lack of CpG suppression in their GC-rich genes (Gardiner-Garden et al. 1992), but neither of these mechanisms could account for an enrichment in G and C residues reaching 100% at third positions in some plant genes.

We suggest that a (selectively neutral) "GC-error-prone" repair synthesis machinery which may be active (1) to varying degrees for different genes and (2) in the vicinity of expressed genes provides the mutational pressure which is responsible for GC enrichment in biased genes of plants. This mechanism of mutational bias is similar to yet more explicit than that postulated by Sueoka (1988)[1] from his analyses of bacterial and animal genomes. Although Wolfe et al. (1989) mentioned the possibility of error-prone repair as a means of differentially increasing GC content within regions of the genome, they strongly favored the replication timing model. Though it has been known for some time that DNA repair in prokaryotes (Mellon and Hanawalt 1989) and eukaryotes (Bohr et al. 1985) is more efficient in transcribed than in nontranscribed DNA, unequivocal evidence for the existence of transcriptionally coupled DNA repair synthesis in yeast and animals has only recently emerged. (See Downes et al. 1993 and Hoeijmakers 1993; see also Bootsma and Hoeijmakers 1993.)

[1] "DNA replication and DNA repair synthesis may make replication errors differently and the extent of DNA repair synthesis may vary among different domains of the chromatin because of the different susceptibility of DNA to damage and repair due to differences in chromatin structure"

Table 4. G + C content and CpG frequencies in nonspermatophyte nuclear coding sequences

Division/species Gene	bp coding	Total G + C (%)	3rd pos. G + C (%)	CpG O/E ratio	Reference
Pteridophyta/*P. munitum*					
Lhcb	807	55	58	0.48	Pichersky (unpublished)
LhcbI	792	57	60	0.58	Pichersky (unpublished)
Lhcb F3	798	57	63	0.51	Pichersky et al. 1990
Lhcb F6Ψ	1,020	42	—	0.18	Pichersky et al. 1990
Lhcb F7Ψ	830	56	—	0.60	Pichersky et al. 1990
Lhcb F8Ψ	499	42	—	0.29	Pichersky et al. 1990
Lycopodiphyta/*S. martensii*					
Phy1	3,402	59	71	0.81	EMBL X61458
Bryophyta/*P. patens*					
GapC	1,038	52	61	0.64	Martin et al. 1993a
Lhcb	810	62	83	0.90	Long et al. 1989
myb-related1	1,351	49	55	0.79	Leech et al. 1993
myb-related2	1,266	51	56	0.66	Leech et al. 1993
Bryophyta/*C. purpureus*					
PhyCer	3,909	44	42	0.60	Thümmler et al. 1992
Exon I	2,035	47	—	0.75	Thümmler et al. 1992
Exon II	300	47	—	0.37	Thümmler et al. 1992
Exon III	1,574	40	—	0.38	Thümmler et al. 1992

There is thus no doubt that DNA repair in the vicinity of active genes differs from that for bulk DNA in these organisms. The biochemistry of transcriptionally coupled repair synthesis in plants has not been studied in detail, although its presence in *E. coli,* humans, and yeast suggests that the basic mechanism may be common to all organisms (Sweder and Hanawalt 1992). Spermatophytes may also possess this DNA repair system. GC-error-proneness may be a variable, if not reversible, property of transcriptionally coupled repair across higher eukaryotic taxa.

The plots of dinucleotide frequencies in plant genes shown in Fig. 4 are readily interpretable in light of this model. In the case of *Lhcb2*1* of pine, the active gene is clearly enriched in GC content relative to the flanking regions, which are furthermore heavily CpG suppressed, likely due to nonprotection from methylation. In maize *GapA1,* the promoter, exons, and introns are quite GC-rich, which we interpret to be the result of GC-error-prone repair synthesis on an active gene; the flanking regions are CpG suppressed but also lower in overall GC content. When we plotted the sequence of a recently inactivated (Quigley et al. 1989) maize *GapA* pseudogene *ΨGapA1* (Fig. 4), we were surprised to find a very GC-rich region in the 3′ noncoding flank. A database search revealed that this region was not really noncoding as we believed, but rather contained the remnant of a once-active gene on the opposite strand: an interrupted reading frame for an ATP-dependent RNA helicase domain with 50% amino acid identity to the yeast *suv3* gene product. This congruence between high GC content and coding capacity supports the notion of transcriptionally coupled, GC-error-prone repair synthesis as a plausible means of introducing GC bias into plant genes. In maize *GapC1,* the promoter and first-five-plus-last-exon are more GC-rich than the 5′ flank, consistent with the repair synthesis model, but the region from intron 5 through exon 10 is CpG suppressed. Thus, in *GapC1* we either observe suppression superimposed upon repair synthesis within the transcribed region or a restriction of repair synthesis to the promoter-proximal region of the gene.

The data suggest that "GC-error-prone" or "GC-biased" transcriptionally coupled DNA repair synthesis should exist in conifers and graminaceous monocots, since they possess biased genes in which the bias is localized to the immediate vicinity of transcribed regions (Fig. 4). This mechanism appears to have been almost entirely lost or heavily modified in dicots. Since some nuclear genes of bryophytes are moderately biased (Table 4) and those of green algal outgroups such as *Chlamydomonas reinhardtii* are extremely GC-biased (Campbell and Gowri 1990; both *GapA* and *GapC* of *C. reinhardtii* have >90% GC at third codon positions, Kersarnach et al. 1994), one might speculate that biased repair synthesis may have existed in ancestral chlorophytes and that they therefore may have possessed GC-rich genes. We offer no explanation for the variation in efficiency of "GC-error-prone repair" machinery across genes within a biased genome (Table 3) other than the postulate that some genes may lose their affinity as substrates for error-prone repair synthesis. Indeed, in animals there is no simple correlation between rates of transcription and repair synthesis, and the relationship between chromatin structure and rate of repair synthesis for active genes is also unclear (Downes et al. 1993).

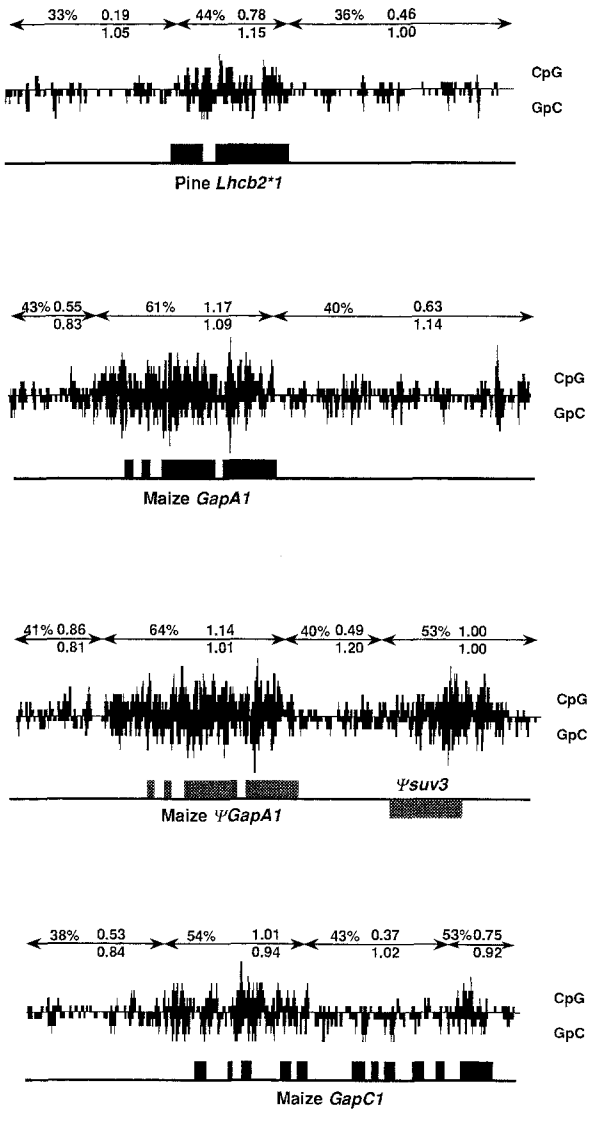

Fig. 4. Dinucleotide distributions plotted against intron-exon structure. Dinucleotides were plotted as in Fig. 1. Vertical increments for CpGs and GpCs were plotted back to back along the horizontal axis. Exons of active are indicated as *black rectangles;* exons of pseudogenes are indicated as *gray rectangles. Double-headed arrows* above plots indicate regions for which GC content (shown in %) and O/E ratios for CpG (shown above the *arrow*) and GpC (shown below the *arrow*), respectively, were calculated. Sources of sequences are pine *Lhcb2*1,* Kojima et al. (1992); maize *GapC1,* Martinez et al. (1989); maize *GapA1,* Quigley et al. (1988); maize *ΨGapA1,* Quigley et al. (1989). The region designated *Ψsuv3* located 5′ of the *ΨGapA1* pseudogene contains an interrupted reading frame (five dispersed single base indels over an 825-bp region of homology, data not shown) which we identified through database search as having 50% amino acid identity across 275 residues to the C-terminal ATP-dependent RNA helicase domain of the yeast *suv3* protein (Stepien et al. 1992). The *suv3* homologous region is located on the lower strand, opposite *ΨGapA1.* Although the *Ψsuv3* sequence is probably inactive, since the reading frame is not contiguous, the degree of homology to yeast *suv3* suggests that it at one time was. Maize *GapA1* and *ΨGapA1* sequences shown share no significant homology (<40%) outside of the promoter and exon regions; the GC-rich regions (61% and 64%, respectively) share 88% nucleotide identity. The *scale bar* at the lower left indicates 1 kb.

Nonetheless, factors which infuence the rate of repair synthesis in plants should be able to evoke transitions from the biased to nonbiased state and vice versa as observed for *GapA, Lhcb5, Pcr,* and *Pal* in monocot-gymnosperm comparisons (Fig. 1 and Table 4).

During evolution, many plant genes, both biased and unbiased, clearly underwent CpG suppression. In the case of dicotyledons surveyed to date, it appears that only few regions of the genome, the significant CpG-rich regions (Gardiner-Garden and Frommer 1992), escaped suppression. Many genes of monocots, conifers, bryophytes, and algae escaped CpG suppression and have also maintained GC bias (i.e., *Lhcb1*). In the case of dicots, the GC-error-prone nature of repair synthesis appears to have been modified (or perhaps lost altogether in some taxa), since heavily biased dicot genes have not been observed to date. The fact that third positions in dicot coding sequences are somewhat more GC-rich than flanking regions (Gardiner-Garden and Frommer 1992) may be a reflection of residual GC-error-prone repair synthesis, the effects of which in proximal flanking regions may be overridden by non-error-prone mutation events. The fact that GC-rich isochores contain most of the genes in plant genomes which possess bias (Salinas et al. 1988) would be explained under our model through the cumulative effect of error-prone repair synthesis in linked genes across isochore-length stretches of DNA, rather than through Darwinian selection. Of course, the GC-error-prone repair synthesis postulated here, though indirectly evidenced for plants in Fig. 4, is a hypothetical model which remains to be demonstrated by direct biochemical means.

In summary, we have generated some new nuclear sequences from gymnosperms and have collected and analyzed a number of gymnosperm and nonspermatophyte nuclear sequences in order to better understand processes influencing genome evolution in plants. We have presented a simple but speculative model of land-plant genome evolution which considers GC bias and CpG frequency as independent phenomena and which can account for observed GC bias, CpG suppression, and isochore structure in plant genomes. Our model is compatible with the finding that the majority of expressed plant genes are located in GC-rich isochores (Montero et al. 1990; Matassi et al. 1991), but could also account for the fact that GC content tends to be higher in the vicinity (~1 kb) of expressed genes than in more distal flanking sequences (Martinez et al. 1989; Gardiner-Garden and Frommer 1992). If early angiosperms contained a GC-rich class of genes as extant monocots and gymnosperms do, bias may have been retained in some dicotyledonous genomes and, if found, could help to identify dicot lineages which branched off early in angiosperm evolution.

Acknowledgments. We wish to thank Dave Harry, University of Illinois at Urbana–Champaign; Keith Hutchison, University of Maine;

Carol Loopstra and Ross Whetten, North Carolina State University; Fran Pichersky, University of Michigan; and Annika Sundås, University of Uppsala, for communication of unpublished sequences. Also Steffen Nock and Klaus-Peter Häger (Bayreuth) for help in preparing the cDNA libraries; Giorgio Matassi and Nicolas Carels for critical reading of the manuscript; the Gesellschaft für Biotechnologische Forschung for use of their computer facilities; and Petter Gustafsson for general support. This work was supported in part by the Swedish Council for Forestry and Agricultural Research and by grant Ma 1426/1-1 from the Deutsche Forschungsgemeinschaft to W.M.

References

Belanger FC, Hepburn AG (1990) The evolution of CpNpG methylation in plants. J Mol Evol 30:26–35

Bernardi G (1989) The isochore organization of the human genome. Annu Rev Genet 23:637–661

Bernardi G (1993) The vertebrate genome: Isochores and evolution. Mol Biol Evol 10:186–204

Bird AP (1986) CpG-rich islands and the function of DNA methylation. Nature 321:209–213

Bohr VA, Smith CA, Okumoto DS, Hanawalt PC (1985) DNA repair in an active gene: removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than in the genome overall. Cell 40:359–369

Bootsma D, Hoeijmakers JHJ (1993) Engagement with transcription. Nature 363:114–115

Brinkmann H, Martinez P, Quigley F, Martin W, Cerff R (1987) Endosymbiontic origin and codon bias of the nuclear gene for chloroplast glyceraldehyde-3-phosphate dehydrogenase from maize. J Mol Evol 26:320–328

Brinkmann H, Cerff R, Salomon M, Soll J (1989) Cloning and sequence analysis of cDNAs encoding the cytosolic precursors of subunits GapA and GapB of chloroplast glyceraldehyde-3-phosphate dehydrogenase from pea and spinach. Plant Mol Biol 13:81–94

Campbell WH, Gowri G (1990) Codon usage in higher plants, green algae, and cyanobacteria. Plant Physiol 92:1–11

Cerff R (1979) Quaternary structure of higher plant glyceraldehyde-3-phosphate dehydrogenases. Eur J Biochem 94:243–247

Cerff R, Kloppstech K (1982) Structural diversity and differential light control of mRNAs coding for angiosperm glyceraldehyde-3-phosphate dehydrogenases. Proc Natl Acad Sci USA 79:7624–7628

Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12:387–395

Downes CS, Anderson JR, Johnson RT (1993) Fine tuning of DNA repair in trancribed genes: mechanisms, prevalence and consequences. Bioessays 15:209–216

Drouin G, Dover GA (1990) Independent gene evolution in the potato actin gene family demonstrated by phylogenetic procedures for resolving gene conversions and the phylogeny of angiosperm actin genes. J Mol Evol 31:132–150

Fliegmann J, Schröder G, Schanz S, Britsch L, Schröder J (1992) Molecular analysis of chalcone and dihydropinosylvin synthase from Scots pine (*Pinus sylvestris*), and differential regulation of these and related enzyme activities in stressed plants. Plant Mol Biol 18:489–503

Gardiner-Garden M, Frommer M (1992) Significant CpG-rich regions in angiosperm genes. J Mol Evol 14:231–245

Gardiner-Garden M, Sved JA, Frommer M (1992) Methylation sites in angiosperm genes. J Mol Evol 34:219–230

Gruenbaum Y, Naveh-Many T, Cedar H, Razin A (1981) Sequence specificity of methylation in higher plant DNA. Nature 292: 860–862

Hepburn A, Belanger F, Mattheis J (1987) DNA methylation in plants. Dev Genet 8:475–493

Hoeijmakers JHJ (1993) Nucleotide excision repair I: from *E. coli* to yeast. Trends Genet 9:173–177

Hutchison KW, Harvie PD, Singer PB, Brunner AF, Greenwood MS (1990) Nucleotide sequence of the small subunit of ribulose-1,5-bisphosphate carboxylase from the conifer *Larix laricina*. Plant Mol Biol 14:281–284

Jansson S, Gusafsson P (1990) Type I and Type II genes for the chlorophyll *a/b*-binding protein in the gymnosperm *Pinus sylvestris* (Scots pine): cDNA cloning and sequence analysis. Plant Mol Biol 14:287–296

Jansson S, Gustafsson P (1991) Evolutionary conservation of the chlorophyll a/b-binding proteins: cDNAs encoding Type I, II, and III LHC I polypeptides from the gymnosperm Scots pine. Mol Gen Genet 229:67–76

Karpinski S, Wingsle G, Olsson O, Hällgren JE (1992) Characterization of cDNA encoding CuZn-superoxide dismutases in Scots pine. Plant Mol Biol 18:545–555

Kenny JR, Dancik BP, Florence LZ, Nargang FE (1988) Nucleotide sequence of the carboxy-terminal portion of a lodgepole pine actin gene. Can J For Res 18:1595–1602

Kersarnach R, Brinkmann H, Liaud M-F, Zhang D-X, Martin W, Cerff R (1994) Five identical intron positions in ancient duplicated genes of eubacterial origin. Nature 367:387–389

Kinlaw CS, Harry DE, Sederoff RR (1990) Isolation and characterization of alcohol dehydrogenase cDNAs from *Pinus radiata*. Can J For Res 20:1343–1350

Kojima K, Yamamoto N, Sasaki S (1992) Structure of the pine (*Pinus thunbergii*) chlorophyll a/b-binding protein gene expressed in the absence of light. Plant Mol Biol 19:405–410

Leal I, Misre S (1993) Molecular cloning and characterisation of a legumin-like storage protein cDNA of douglas fir seeds. Plant Mol Biol 21:709–715

Leech MJ, Martin CM, Wang TL (1993) Regulation of *myb*-related genes in the moss *Physcomitrella patens*. Plant J 3:51–61

Long Z, Wang SY, Nelson N (1989) Cloning and nucleotide sequence analysis of genes coding for the major chlorophyll-binding protein of the moss *Physcomitrella patens* and the halotolerant alga *Dunaliella salina*. Gene 76:299–312

Martin W, Cerff R (1986) Prokaryotic features of a nucleus encoded enzyme: cDNA sequences for chloroplast and cytosolic glyceraldehyde-3-phosphate dehydrogenases from mustard (*Sinapis alba*). Eur J Biochem 159:323–331

Martin W, Gierl A, Saedler H (1989) Molecular evidence for pre-Cretaceous angiosperm origins. Nature 339:46–48

Martin W, Lagrange T, Li Y-F, Bisanz-Seyer C, Mache R (1990) Hypothesis for the evolutionary origin of the chloroplast ribosomal protein L21 of spinach. Curr Genet 18:553–556

Martin W, Lydiate D, Brinkmann H, Forkmann G, Saedler H, Cerff R (1993a) Molecular phylogenies in angiosperm evolution. Mol Biol Evol 10:140–163

Martin W, Nock S, Meyer-Gauen G, Häger K-P, Jensen U, Cerff R (1993b) A method for isolation of cDNA-quality mRNA from immature seeds of a gymnosperm rich in polyphenolics. Plant Mol Biol 22:555–556

Martin W, Brinkmann H, Savonna C, Cerff R (1993c) Evidence for a chimeric nature of nuclear genomes: Eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. Proc Natl Acad Sci USA 90:8692–9696

Martinez P, Martin W, Cerff R (1989) Structure, evolution and anaerobic regulation of a nuclear gene encoding cytosolic matrix dehydrogenase from maize. J Mol Biol 208:551–565

Matassi G, Montero LM, Salinas J, Bernardi G (1989) The isochore organisation and compositional distribution of homologous cod-

ing sequences in the nuclear genome of plants. Nucleic Acids Res 17:5273–5290

Matassi G, Melis R, Mecaya G, Bernardi G (1991) Compositional bimodality of the nuclear genome of tobacco. Nucleic Acids Res 19:3561–3567

Matsuoka M (1990) Classification and characterization of cDNA that encodes the light-harvesting chlorophyll a/b binding protein of photosystem II from rice. Plant Cell Physiol 31:519–526

Matsuoka M, Kano-Murakami Y, Tanaka Y, Ozeki Y, Yamamoto N (1988) Classification and nucleotide sequence of cDNA encoding the small subunit of ribulose-1,5-bisphosphate carboxylase from rice. Plant Cell Physiol 29:1015–1022

Maxam AM, Gilbert W (1980) Sequencing end-labelled DNA with base-specific chemical cleavages. Methods Enzymol 65:499–560

McElroy D, Rothenberg M, Wu R (1990) Structural characterization of a rice action gene. Plant Mol Biol 14:163–171

McGrath JM, Pichersky E (1991) 5-methylcytosine content in homosporous ferns. Abstract 1822, Third International Congress of ISPMB, Tucson

Mellon I, Hanawalt PC (1989) Induction of the *Escherichia coli* lactose operon selectively increases repair of its transcribed strand. Nature 342:95–97

Minami E, Ozeki Y, Matsuoka M, Koizuka N, Tanaka Y (1989) Structure and some characterization of the gene for phenylalanine ammonia-lyase from rice plants. Eur J Biochem 185:19–25

Montero LM, Salinas J, Matassi G, Bernardi G (1990) Gene distribution and isochore organisation in the nuclear genome of plants. Nucleic Acids Res 18:1857–1867

Newton CH, Flinn BS, Sutton BCS (1992) Vicilin-like seed storage proteins in the gymnosperm interior spruce (*Picea glaucalengelmanii*). Plant Mol Biol 20:315–322

Niesbach-Klösgen U, Barzen E, Bernhardt J, Rohde W, Schwarz-Sommer Zs, Reif HJ, Wienand U, Saedler H (1987) Chalcone synthase genes in plants: a tool to study evolutionary relationships. J Mol Evol 26:213–225

O'Neill SD, Tong Y, Spörlein B, Forkmann G, Yoder JI (1990) Molecular genetic analysis of chalcone synthase in *Lycopersicon esculentum* and an anthocyanin-deficient mutant. Mol Gen Genet 224:279–288

Perl-Treves R, Nacmias B, Aviv D, Zeelon EP, Galun E (1988) Isolation of two cDNA clones from tomato containing two different superoxide dismutase sequences. Plant Mol Biol 11:609–623

Pichersky E, Bernatzky R, Tanksley SD, Breidenbach B, Kausch AP, Cashmore AR (1985) Molecular characterization and genetic mapping of two clusters of genes encoding chlorophyll *al/b*-binding proteins in *Lycopersicon esculentum* (tomato). Gene 40:247–258

Pichersky E, Hoffman NE, Malik VS, Bernatzky R, Tanksley SD, Szabo L, Cashmore AR (1987) The tomato *Cab*-4 and *Cab*-5 genes encode a second type of CAB polypeptides localized in photosystem II. Plant Mol Biol 9:109–120

Pichersky E, Soltis D, Soltis P (1990) Defective chlorophyll a/b-binding protein genes in the genome of a homosporous fern. Proc Natl Acad Sci USA 87:195–199

Pichersky E, Subramaniam R, White MJ, Reid J, Aebersold R, Green BR (1991) Chlorophyll a/b binding polypeptides of CP29, the internal chlorophyll a/b complex of PSII: Characterization of the tomato gene encoding the 26 kDa (type I) polypeptide, and evidence for a second CP29 polypeptide. Mol Gen Genet 227:277–284

Quigley F, Martin W, Cerff R (1988) Intron conservation across the prokaryote-eukaryote boundry: structure of the nuclear gene for chloroplast glyceraldehyde-3-phosphate dehydrogenase from maize. Proc Natl Acad Sci USA 85:2672–2676

Quigley F, Brinkmann H, Martin W, Cerff R (1989) Strong functional GC-pressure in a light regulated maize gene encoding chloroplast GAPA: implications for the evolution of GAPA pseudogenes. J Mol Evol 29:412–421

Rohde W, Dörr S, Salamini F, Becker D (1991) Structure of a chalcone synthase gene from *Hordeum vulgare*. Plant Mol Biol 16: 1103–1106

Sakamoto A, Ohsuga H, Tanaka K (1992) Nucleotide sequences of two cDNA clones encoding different Cu/Zn-superoxide dismutases expressed in developing rice seed (*Oryza sativa* L.). Plant Mol Biol 19:323–327

Salinas J, Matassi G, Montero LM, Bernardi G (1988) Compositional compartmentalization and compositional pattern in the nuclear genomes of plants. Nucleic Acids Res 16:4269–4285

Schaeffer L, Roy R, Humbert S, Moncollin V, Vermeulen W, Hoeijmakers JHJ, Chambon P, Egly J-M (1993) DNA repair helicase: a component of BTF2 (TFIIH) basic transcription factor. Science 260:58–63

Schultz R, Steinmüller K, Klaas M, Forreiter C, Rasmussen S, Hiller C, Apel K (1989) Nucleotide sequence of a cDNA coding for the NADPH-protochlorophyllide oxidoreductase (PCR) of barley (*Hordeum vulgare* L.) and its expression in *Escherichia coli*. Mol Gen Genet 217:355–361

Schwarz-Sommer Z, Gierl A, Klösgen RB, Wienand U, Petersen PA, Saedler H (1984) The Spm (En) transposable element controls the excision of a 2-kb DNA insert at the wx^{m-8} allele of *Zea mays*. EMBO J 3:1021–1028

Schwekendiek A, Pfeffer G, Kindl H (1992) Pine stilbene synthase cDNA, a tool for probing environmental stress. FEBS Lett 301:41–44

Selby CP, Sancar A (1993) Molecular mechanism of transcription repair coupling. Science 260:53–58

Shih M-C, Lazar G, Goodman HM (1986) Evidence in favor of the symbiotic origin of chloroplasts: primary structure and evolution of tobacco glyceraldehyde-3-phosphate dehydrogenases. Cell 47:73–80

Sørensen AB, Lauridsen BF, Gausing K (1992) Barley (Hordeum vulgare) gene for CP29, a core chlorophyll a/b binding protein of photosystem II. Plant Physiol 98:1538–1540

Spano AJ, He Z, Michel H, Hunt DF, Timko MP (1992a) Molecular cloning, nuclear gene structure and developmental expression of NADPH-protochlorophyllide oxidoreductase in pea (*Pisum sativum* L.). Plant Mol Biol 18:967–972

Spano AJ, He Z, Timko MP (1992b) NADPH:protochlorophyllide oxidoreductases in white pine (*Pinus strobus*) and loblolly pine (*P. taeda*). Mol Gen Genet 236:86–95

Stepien PP, Margossian SP, Landsman D, Butow RA (1992) The yeast nuclear gene *suv3* affecting mitochondrial post-transcriptional processes encodes a putative ATP-dependent RNA helicase. Proc Natl Acad Sci USA 89:6813–6817

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657

Sugita M, Manzara T, Pichersky E, Cashmore A, Gruissem W (1987) Genomic organization, sequence analysis and expression of all five genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase from tomato. Mol Gen Genet 209:247–256

Sundås A, Tandre K, Holmstedt E, Engström P (1992) Differential gene expression during germination and after the induction of adventitious bud formation in Norway spruce embryos. Plant Mol Biol 18:713–724

Sweder KS, Hanawalt PC (1992) Preferred repair of cyclobutane pyrimidine dimers in the transcribed strand of a gene in yeast chromosomes and plasmids is dependent on transcription. Proc Natl Acad Sci USA 89:10696–10700

Thüummler F, Dufner M, Kreisl PP, Dittrich P (1992) Molecular cloning of a novel phytochrome gene of the moss *Ceratodon purpureus* which encodes a putative light-regulated protein kinase. Plant Mol Biol 20:1003–1017

Van Der Straeten D, Rodrigues-Pousada RA, Gielen J, Van Montagu M (1991) Tomato alcohol dehydrogenase: Expression during fruit ripening and under hypoxic conditions. FEBS Lett 295:39–42

Whetten RW, Sederoff RR (1992) Phenylalanine ammonia-lyase from loblolly pine: Purification of the enzyme and isolation of complementary DNA clones. Plant Physiol 98:380–386

Wolfe KH, Sharp P, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Xie Y, Wu R (1989) Rice alcohol dehydrogenase genes: anaerobic induction, organ specific expression and characterization of cDNA clones. Plant Mol Biol 13:53–68

Yamamoto N, Kano-Murakami Y, Matsuoka M, Ohashi Y, Tanaka Y (1988a) Nucleotide sequence of a full length cDNA clone of ribulose bisphosphate carboxylase small subunit gene from green dark-grown pine (*Pinus tunbergii*) seedling. Nucleic Acids Res 16:11830

Yamamoto N, Matsuoka M, Kano-Murakimi Y, Tanaka Y, Ohashi Y (1988b) Nucleotide sequence of a full length cDNA clone of light harvesting chlorophyll a/b binding protein gene from dark-grown pine (*Pinus thunbergii*) seedlings. Nucleic Acids Res 16:11829