J Mol Evol (1995) 41:597–603

JOURNAL OF MOLECULAR EVOLUTION

Neighboring Base Composition Is Strongly Correlated with Base Substitution Bias in a Region of the Chloroplast Genome

Brian R. Morton,* Michael T. Clegg

Department of Botany and Plant Sciences, University of California, Riverside, Riverside, CA 92521-0124, USA

Received: 6 August 1994 / Accepted: 10 March 1995

Abstract. Nucleotide sequence from a region of the chloroplast genome is presented for 12 species spanning four subfamilies of the grass family. The region contains the coding sequence for the *rbc*L gene and the intergenic spacer between the gene coding the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (*rbc*L) and the photosystem I gene *psa*I. This intergenic spacer contains a pseudogene for *rpl*23 as well as two noncoding segments with different A+T contents. Using the sequence of *rbc*L a chloroplast phylogeny of this family was constructed by parsimony. Variable sites of the two noncoding segments were traced onto the phylogeny to study the dynamics of base substitution. This was also performed for the fourfold-degenerate sites of the *rbc*L gene. A wide variation in transversion/transition is observed between the two noncoding segments and between the noncoding DNA and the fourfold-degenerate sites of *rbc*L. This variation is correlated with regional A+T content. As regional A+T content decreases, the ratio of transversions to transitions also decreases. Substitutions were then scored in relation to neighboring base composition. The composition of the two bases immediately flanking each substitution is highly correlated with the transversion/transition bias. When both the 5′ and 3′ flanking bases are an A or a T, transversions are observed 2.2 times as frequently as transitions. When either or both neighbors are a C or a G, the opposite trend is found; transitions are observed 1.5 times more frequently than transversions.

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

Correspondence to: Brian R. Morton

Key words: Nucleotide substitution — Transition — Transversion — Phylogenetic analysis — Neighbor composition

Introduction

The process of nucleotide substitution is one of the most fundamental features of molecular evolution. Understanding the process and pattern of substitution underlies the methods for estimating the number of substitutional events occurring along lineages, and it is fundamental to some methods for reconstructing phylogenies based on molecular data.

Substitutions often occur with unequal probability such that the pattern of substitution is biased. The pattern of nucleotide substitution can be affected at the level of spontaneous mutation or at the level of fixation by natural selection. At the mutational level, biases can arise by several means, such as thermodynamic stability of the mispaired intermediate (Topal and Fresco 1976) and interaction with neighboring bases (Topal et al. 1980; Radman and Wagner 1986; Jones et al. 1987; Mendelman et al. 1989). Efficiency of the proofreading mechanism of the replication apparatus and of mismatch repair can also affect the mutational pattern. Mutations in *Escherichia coli* that affect the repair mechanism have been observed to alter the frequency of occurrence of certain types of nucleotide substitutions (Nghiem et al. 1998; Cabrera et al. 1988; Michaels et al. 1990) or to be involved in the repair of specific mismatches (Kramer et al. 1984; Akiyama et al. 1989). Similar loci have been found in eu-

karyotic systems (Jiricny et al. 1988; Stephenson and Karran 1989).

The pattern of substitution between closely related sequences has been well studied, particularly in mammals (Brown et al. 1982; Gojobori et al. 1982; Li et al. 1984; Blake et al. 1992) and the mitochondrial genome of *Drosophila* (Wolstenholme and Clary 1985; Clary and Wolstenholme 1987; DeSalle et al. 1987; Satta et al. 1987; Nigro et al. 1991; Tamura 1992). Since the substitutions observed in gene-coding sequences have been "filtered" by natural selection, they do not necessary reflect the spontaneous mutational process. To more directly address the question of mutation dynamics some studies have focused on apparently neutral substitutions such as those observed in pseudogenes (Gojobori et al. 1982; Li et al. 1984; Blake et al. 1992) or at the third-position substitutions in fourfold-degenerate amino acids (Tamura 1992).

The studies on mammalian DNA, both mitochondrial DNA (mtDNA) and nuclear genome pseudogenes, revealed an excess of transitional substitutions as well as an increased rate of substitution in mtDNA relative to nuclear DNA (Brown et al. 1982; Li et al. 1984). This bias toward transition substitutions has had important implications for estimations of genetic divergence (Nei 1987) and is employed widely in weighting character-state changes in phylogenetic analyses (Swofford and Olsen 1990).

Neighboring bases may also influence the pattern of nucleotide substitution. Evidence for such an influence has been observed in both *E. coli* (Jones et al. 1987; Mendelman et al. 1989) and in mammalian systems (Blake et al. 1992). Studies on mismatch repair in *E. coli* have shown that the transitional mismatches G:T and A:C are repaired most efficiently (Radman and Wagner 1986). The efficiency of this repair increases as the G+C content of the neighboring four to ten bases on both sides increases, suggesting that the repair is more efficient in increasingly stable double helixes, particularly for transition intermediates (Radman and Wagner 1986). The result is that the local conformation of the helix can affect the pattern of substitution through repair efficiency. The pattern observed has led Radman and Wagner (1986) to refer to mismatches leading to a transversion, particularly G:A and C:T, in A+T-rich regions as nonrepairable mismatches in *E. coli*.

Although the chloroplast genome is widely used as a marker for phylogenetic studies (Clegg 1993), the dynamics of substitution have not been studied in great detail. Examination of the *rbc*L coding sequence and the upstream noncoding region showed that there was a bias toward transitions in the coding sequence but an equal frequency of transitions and transversions in the intergenic spacer (Zurawski et al. 1984). The current study focuses on a small region of the chloroplast genome in members of the grass family. Sequence data have been obtained over a wide representation of the family from the first base of the gene *rbc*L to the first base of the gene *psa*I. The two genes are separated by a spacer region ranging in size from about 900 bases to 1,700 bases and containing a pseudogene, ψ*rpl*23 (Morton and Clegg 1993). Using *rbc*L, a phylogeny of the chloroplast genome in this family was constructed and then used as a basis for studying nucleotide substitutions in various noncoding segments and the *rbc*L gene. The transversion/transition ratio varies widely among the various DNA segments and between noncoding DNA and fourfold-degenerate sites of *rbc*L. This variation is well correlated with the difference in A+T content of the segments and appears to be the result of an influence on mutational dynamics of the composition of the nucleotides flanking each substitution.

Materials and Methods

Genomic DNA was prepared by the method of Doyle and Doyle (1987) and the region from the first base of *rbc*L to *psa*I amplified by the PCR method and sequenced directly as described previously with primers PsaI (Morton and Clegg 1993) and Z1 (Doebly et al. 1990). The species sequenced are listed in Table 1. The amplification products from *Bambusa multiplex* and *Erioneuron nealegis* were cloned into the TA cloning vector from InvitroGen (San Diego, CA) following the protocol supplied. Positive clones were PCR amplified with Z1 and PsaI primers and sequenced directly. Sequence alignment was done using the BestFit program of the GCG package (Devereaux et al. 1984) for pairwise comparisons between a new sequence and the closest relative in the alignment previously presented (Morton and Clegg 1993). Each sequence was added to the multiple alignment in this manner. For all sequencing, one strand of the PCR product was sequenced twice; the second strand was sequenced at least once.

Phylogenetic analyses were done using PAUP 3.0s (Swofford 1991). The *rbc*L analysis was performed on bases 1 to 1428 of the coding sequence. Sequence of *rbc*L is available from several species for which noncoding sequence was not obtained (Table 1). These extra sequences were used in the phylogeny construction.

Equal weighting was used for all character-state changes. Searches performed were heuristic with a random input order. Ten repetitions of each search were performed. Bootstrapping on the *rbc*L data set was performed for 100 replicates, each replicate consisting of three repetitions of random-order-input heuristic searches.

To examine substitutions the most parsimonious tree from the PAUP analysis on *rbc*L was input into MacClade (Maddison and Maddison 1992) along with the data set from each noncoding region of interest and the Character Trace function was performed on each variable site. Most substitutions could be scored unambiguously in this manner. Substitutions ambiguous in terms of direction were excluded from the substitutional matrix. These substitutions were, however, scored for calculation of transversion-to-transition ratio. In *rbc*L, four sites (279, 282, 1095, and 1425) were completely ambiguous and were not scored. Two indel mutations in the noncoding region scored were ambiguous in terms of the end point in that two different substitution events could be observed depending on how gaps were introduced. In each of the cases the substitutions were excluded. Substitutions within the *rpl*23 pseudogene were excluded from the analysis due to the observation that gene conversion between ψ*rpl*23 and the functional gene has occurred at least twice during the divergence of the grasses (Morton and Clegg 1993).

Composition of the flanking bases in the lineage in which a substitution occurred was also recorded. In the few cases where two substitutions occurred along a single branch at neighboring sites, the

Table 1. Representatives of the grass family and sequence source

Species	*rbc*L	Noncoding region
Aegilops crassa	Terachi et al. 1987	Ogihara et al. (1988)
Avena sativa	L15300[a]	—
Bambusa multiplex	Duvall et al. (1993a)	This study
Cenchrus setigerus	Doebley et al. (1990)	—
Eragrostis japonica	This study	This study
Erioneuron nealegis	This study	This study
Hordeum vulgare	Zurawski et al. (1984)	Morton and Clegg (1993)
Joinvillea plicata	Duvall et al. (1993a)	—
Muhlenbergia setarioides	This study	This study
Neurachne tenuifolia	Hudson et al. (1990)	—
Oryza sativa	Hiratsuka et al. (1989)	Hiratsuka et al. (1989)
Pennisetum glaucum	Doebley et al. (1990)	Morton and Clegg (1993)
Puccinella distans	Doebley et al. (1990)	—
Sorghum bicolor	Doebley et al. (1990)	Morton and Clegg (1993)
Tripsacum dactyloides	This study	Morton and Clegg (1993)
Triticum aestivum	Terachi et al. (1987)	—
Zea mays	Gaut et al. (1992)	Morton and Clegg (1993)
Zizania texana	Duvall et al. (1993b)	This study

[a] Unpublished sequence. The GenBank accession number is given

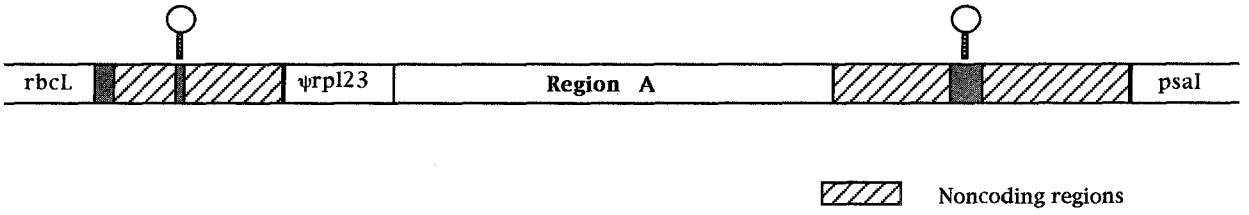

Fig. 1. Diagram of the region from *rbc*L to *psa*I in the large single-copy region of the chloroplast genome showing region A and the *rpl*23 pseudogene. The noncoding regions combined in this study are indicated by *hatching*. Three short noncoding regions excluded from the analysis are indicated by *shading*. (See text.) Inverted repeat regions discussed in Morton and Clegg (1993) are indicated by a *hairpin*.

substitution was excluded from the neighboring bases analysis. Composition of the flanking bases was done by recording the A+T content of the two nucleotides flanking the substitution (zero, one, or two bases A or T).

Results and Discussion

The region of the chloroplast genome examined in this study is diagrammed in Fig. 1. The region contains a pseudogene for *rpl*23 as well as two inverted repeats and a large deletion in the members of the Panicoideae subfamily, which included *P. glaucum* and *Zeu mays,* spanning the region called region A in Fig. 1 (Morton and Clegg 1993). This deletion, beginning within ψ*rpl*23 26 bases from the end of ψ*rpl*23 and *rpl*23 homology, is also found in this study to be present in the subfamily Chloridoideae represented by *Muhlenbergia setarioides, Eragrostis japonica,* and *Erioneuron nealegis.* Region A is defined in this study as the deleted region minus the 26 bases at the 3′ end of ψ*rpl*23. The complete noncoding region is about 1.6 kb in length in *Oryza sativa* and about 0.95 kb in *Z. mays*. Polarity is defined in this study by the *rbc*L gene. The noncoding region diagrammed is 3′ to *rbc*L.

Phylogenetic Analysis

The phylogenetic analysis of the grass species using *rbc*L sequence data yielded a single most parsimonious tree which is shown in Fig. 2. The deletion event of region A discussed earlier is indicated on the phylogeny. Given that the phylogeny is employed in an examination of substitutions, the character-change weighting scheme used in tree construction call influence the substitutions inferred from the tree topology. Since an a priori distribution is required, the uniform distribution (equal weighting) was used, as it makes the smallest number of assumptions regarding substitution dynamics. However, the substitution data for the *rbc*L gene reveal a large number of transitions relative to transversions. Therefore, tree construction was repeated using a weighting scheme under which transitions are twice as likely as transversions. The same topology is obtained for this weighting scheme as with equal weighting. The bootstrap results for both weighting schemes are shown in Fig. 2.

Previous phylogenetic analysis of this family using *rbc*L yielded ambiguous results (Doebley et al. 1990). With the addition of the members of the Chloridoideae in

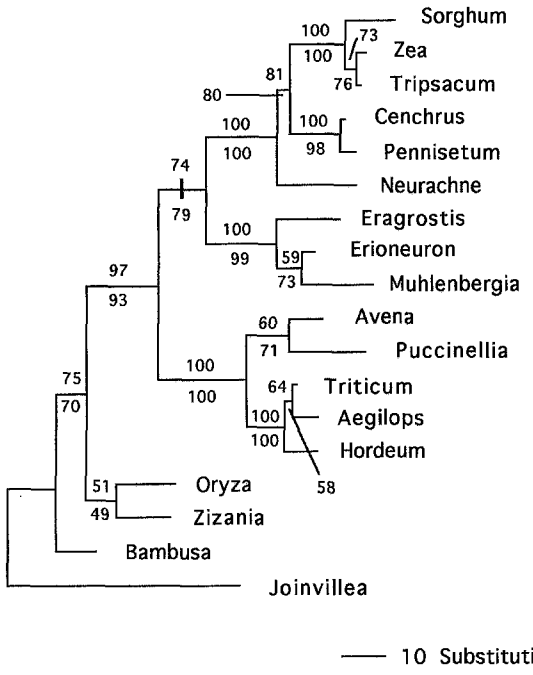

Fig. 2. Most parsimonious tree of the grass family using *rbc*L. The deletion of noncoding region A (see Fig. 1) is indicated by a *bar*. Branch lengths are shown to scale based on the number of substitutions along that edge. Results of 100 bootstrap replicates are indicated for all clades. The bootstrap results for equal weighting are indicated above the branches; those for the weighting scheme favoring transitions (see text) are indicated below. The *scale bar* represents ten substitution events without accounting for possible multiple substitutions at a site.

this study a single most parsimonious tree is obtained which agrees very well with traditional grass systematics (Clayton and Renvoize 1986), suggesting that *rbc*L might be a useful systematic tool for this family.

Patterns of Nucleotide Substitution

The noncoding region downstream of *rbc*L was sequenced for only 12 of the 18 species in Fig. 2 (see Table 1) but substitutions can still be investigated by examining the variable sites in light of this phylogeny. The substitutional matrix for the noncoding regions, excluding region A, is shown in Table 2. The data are from the combination of the noncoding regions indicated by hatching in Fig. 1. These are the noncoding regions that can be aligned with enough certainty to examine nucleotide substitutions. The three sections excluded are a short hypervariable region immediately following the *rbc*L gene, an 18-base loop from a stem-loop structure approximately 150 bases downstream of *rbc*L, and an inverted repeat which is evolving by complex rearrangements (Morton and Clegg 1993). Substitutions within region A are shown separately in Table 3.

The substitutional matrix for the fourfold-degenerate sites of *rbc*L sequence are shown in Table 4. All species in the phylogeny presented in Fig. 2 were used in this analysis. One consistent feature of substitutions in the chloroplast can be seen in Tables 2, 3, and 4 by comparing the compositional frequency to the frequency at which each base is replaced. This demonstrates that C and G are replaced at a rate that is higher than the com-

Table 2. Substitutions in the noncoding region 3′ of *rbc*L

From–to	G[a] (0.13)	A (0.42)	T (0.35)	C (0.12)	Total
G		9	13	8	30
A	14		16	15	45
T	6	5		16	27
C	3	18	12		33
Total	23[b] (0.15)	32 (0.23)	41 (0.32)	39 (0.30)	135

[a] Average composition by base is given in parentheses
[b] Proportion of total substitutions given in parentheses

Table 3. Substitutions in the noncoding region A

From–to	G[a] (0.22)	A (0.32)	T (0.30)	C (0.17)	Total
G	—	10	8	3	21
A	23	—	4	4	31
T	7	4	—	21	32
C	2	8	7	—	17
Total	32[b] (0.29)	22 (0.22)	19 (0.20)	28 (0.29)	101

[a] Average composition by base is given in parentheses
[b] Proportion of total substitutions given in parentheses

Table 4. Substitution matrix for *rbc*L fourfold-degenerate sites

From–to	G[a] (0.10)	A (0.30)	T (0.48)	C (0.12)	Total
G	—	15	3	4	22
A	20	—	6	5	31
T	2	4	—	24	30
C	2	1	18	—	21
Total	24[b] (0.22)	20 (0.18)	27 (0.26)	33 (0.33)	104

[a] Composition of *O. sativa* is given in parentheses
[b] Proportion of total substitutions given in parentheses

Table 5. Base composition and transversion/transition ratio for segments of chloroplast DNA

Region	Tv/Ts[a]	A + T content
Noncoding[b]	1.43	76.0
Region A	0.61	61.6
*rbc*L 4-fold[c]	0.32	56.5

[a] Transversion/transition ratio
[b] Noncoding regions 3′ of *rbc*L except region A as discussed in text
[c] Substitutions are from the fourfold-degenerate sites. A + T content is given for the entire gene

Table 6. Composition of nucleotides flanking substitutions in noncoding DNA

	Noncoding[a]			Region A			Cumulative		
A + T Content[b]	Tv[c]	Ts	Tv/Ts	Tv	Ts	Tv/Ts	Tv	Ts	Tv/Ts
0	5	4	1.3	11	20	0.55	16	24	0.67
1	24	34	0.71	24	41	0.59	48	75	0.64
2	63	22	2.86	24	17	1.41	87	39	2.23

[a] Noncoding region with the exclusion of region A

[b] Number of the two nucleotides flanking the substitution which are either A or T

[c] Number of transversions (Tv) or transitions (Ts) observed with the given flanking A + T content

positional frequency while A and T are replaced at a lower rate. The equilibrium frequencies for G, A, T, and C are calculated from this matrix by the method given in Tajima and Nei (1982) are 0.29, 0.38, 0.20, and 0.13.

Many previous studies of substitutional patterns have noted a bias toward transitional substitutions. This bias may be a result of thermodynamic stability of intermediates (Topal and Fresco 1976, Brown et al. 1982) or a difference in repair efficiency (Brown et al. 1982). There have been few studies of chloroplast mutational dynamics but in this genome a bias toward transitions has also been observed in coding sequence (e.g., Zurawski et al. 1984). The transversion/transition ratios for the three matrices presented in Tables 2–4 are shown in Table 5. It is apparent that the ratio varies widely between the noncoding DNA and the fourfold-degenerate sites of *rbc*L. Furthermore, there is a significant difference between region A and the rest of the noncoding DNA.

An effect of neighboring bases on substitutions has been observed in other systems (Radman and Wagner 1986; Jones et al. 1987; Mendelman et al. 1989; Blake et al. 1992), and this could be responsible for the variation observed within this region. The A+T contents of the different regions are also given in Table 5. As is apparent, the transversion/transition ratio decreases as the regional A+T content decreases. Therefore, the composition of neighboring bases may be influencing mutations in some manner such that transversions are more frequent in A+T-rich areas.

To test this possibility further, the composition of bases directly flanking each substitution were recorded. The results are given in Table 6 for substitutions in the noncoding region 3′ of *rbc*L excluding region A, and for substitutions in region A. The cumulative results are also given.

The transversion/transition ratio is significantly different for substitutions that are flanked on both the 5′ and 3′ side by an A and/or a T and substitutions that have a G or a C at either flanking position. In the first case, transversions occur more than twice as frequently as transitions. Conversely, when a G or C flanks the substitution, transitions are almost twice as frequent as transversions. Testing for the effect of individual bases either 5′ or 3′ on substitutions revealed no significant effect (data not shown).

This difference exists in both of the noncoding regions 3′ of *rbc*L. The difference between region A and the rest of the noncoding sequence in terms of transversion/transition ratio as noted in Table 5 is only the result of the frequency with which substitutions are flanked by either a G or a C. Due to the lower A+T content of region A, a greater proportion of substitutions are biased toward transitions as a result of neighboring base composition.

The same analysis was performed on the fourfold-degenerate sites of the *rbc*L gene. Due to the structure of the genetic code, most fourfold-degenerate sites are flanked by a G or C on the 5′ side which is the second codon position. As a result, almost all substitutions are flanked by at least one G or C; only two of the 107 substitutions are flanked on both sides by an A and/or T. Therefore, it is not possible to establish a difference between the two classes as was done for the noncoding sequence. It remains a possibility that neighboring base influences are the same as in the noncoding region and that the low transversion/transition ratio of the fourfold-degenerate sites is due to the high G+C content of second codon position of fourfold-degenerate groups.

The influence of neighboring bases on the mutational dynamics observed in this region may be due either to an effect on misincorporation by the polymerase or to an effect of local stability on the repair of the mispaired intermediate. Both of these effects could themselves be influenced in more complex ways than simply neighboring base. It is possible that the influence of neighboring bases is itself a complex function of the composition of a larger number of surrounding nucleotides. Such possible influences remain to be examined.

Conclusions

Within a small region of the chloroplast genome a wide range of substitution dynamics, in terms of transversion/transition ratio, is observed. Fourfold-degenerate sites of *rbc*L show a much lower proportion of transversions than neighboring noncoding regions and variation is observed between two noncoding segments with differing base compositions. In the noncoding DNA there is a very

strong correlation between flanking base composition and transversion/transition bias. This appears to be responsible for the variation in transversion/transition ratio throughout this region of the chloroplast genome. The mechanism by which neighboring base composition influences substitution bias is unknown. Further, it remains to be determined whether this is a general feature of chloroplast DNA and what other factors influence substitution dynamics throughout the entire genome. One important point is that, in the absence of selection, the substitution dynamics at fourfold-degenerate sites of coding sequences can differ substantially from the dynamics of noncoding DNA.

Acknowledgements. The authors thank Mel Duvall for the *B. multiplex* DNA and helpful discussion of grass taxonomics as well as Paul Peterson for generous donation of *Erioneuron, Eragrostis,* and *Muhlenbergia* plant material. This research was supported by NIH grant GM 45144 to M.T. Clegg.

References

Akiyama M, Maki H, Sekiguchi M, Horiuchi T (1989) A specific role of MutT protein: to prevent dG · dA mispairing in DNA replication. Proc Natl Acad Sci USA 86:3949–3952

Blake RD, Hess ST, Nicholson-Tuell J (1992) The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. J Mol Evol 34:189–200

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225–239

Cabrera M, Nghiem Y, Miller JH (1988) *mut*M, a second mutator locus in Escherichia coli that generates G · C-T · A transversions. J Bacteriol 170:5405–5407

Clary DO, Wolstenholme DR (1987) *Drosophila* mitochondrial DNA: conserved sequences in the A+T-rich region and supporting evidence for a secondary structure model of the small ribosomal RNA. J Mol Evol 25:116–125

Clayton WD, Renvoize SA (1986) Genera Graminum. Grasses of the world. Her Majesty's Stationery Office, London.

Clegg MT (1993) Chloroplast gene sequences and the study of plant evolution. Proc Natl Acad Sci USA 90:363–367

DeSalle R, Freedman T, Prager EM, Wilson AC (1987) Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila.* J Mol Evol 26:157–164

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

Doebley J, Durbin M, Golenberg EM, Clegg MT, Ma DP (1990) Evolutionary analysis of the large subunit of carboxylase (*rbc*L) nucleotide sequence among the grasses (Gramineae). Evolution 44: 1097–1108

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf material. Phytochem Bull 19:11–15

Duvall MR, Learn GH, Eguiarte LE, Clegg MT (1993a) Phylogenetic analysis of *rbc*L sequence identifies *Acorus calamus* as the primal extant monocotyledon. Proc Natl Acad Sci USA 90:4641–4644

Duvall MR, Chase MW, Soltis DE, Clegg MT (1993b) Phylogenetic hypotheses for the monocotyledons constructed from *rbc*L sequence data. Ann Miss Bot Gard 80:607–619

Gaut BS, Muse SV, Clark WD, Clegg MT (1992) Relative rates of nucleotide substitution at the *rbc*L locus of monocotyledonous plants. J Mol Evol 35:292–303

Gojobori T, Li WH, Graur D (1982) Patterns of nucleotide substitution in pseudogenes. J Mol Evol 18:360–369

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun C-R, Meng B-Y, Li Y-Q, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol Gen Genet 217:185–194

Hudson GS, Mahon JD, Anderson PA, Gibbs MJ, Badger MR, Andrews TJ, Whitfield PR (1990) Comparison of *rbc*L genes for the large subunit of ribulose-bisphosphate carboxylase from closely related C3 and C4 plant species. J Biol Chem 265:808–814

Jiricny J, Hughes M, Corman N, Rudkin BB (1988) A human 200-kDa protein binds selectively to DNA fragments containing G · T mismatches. Proc Natl Acad Sci USA 85:8860–8864

Jones M, Wagner R, Radman M (1987) Repair of a mismatch is influenced by the base composition of the surrounding nucleotide sequence. Genetics 115:605–610

Kramer B, Kramer W, Fritz HJ (1984) Different base/base mismatches are corrected with different efficiencies by the methyl-directed DNA mismatch-repair system of *E. coli.* Cell 38:879–887

Li WH, Wu CI, Luo CC (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Maddison WP, Maddison DR (1992) MacClade: analysis of phylogeny and character evolution, version 3.0. Sinauer Associates, Sunderland, MA

Mendelman LV, Boosalis MS, Petruska J, Goodman MF (1989) Nearest neighbor influences on DNA polymerase insertion fidelity. J Biol Chem 264:14415–14423

Michaels ML, Cruz C, Miller JH (1990) *mut*A and *mut*C: two mutator loci in *Escherichia coli* that stimulate transversions. Proc Natl Acad Sci USA 87:9211–9215

Morton BR, Clegg MT (1993) A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near *rbc*L in the grass family (Poaceae). Curr Genet 24:357–365

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nghiem Y, Cabrera M, Cupples CG, Miller JH (1988) The *mut*Y gene: a mutator locus in *Escherichia coli* that generates G · C-T · A transversions. Proc Natl Acad Sci USA 85:2709–2713

Nigro L, Solignac M, Sharp PM (1991) Mitochondrial DNA sequence divergence in the melanogaster and oriental species subgroup of *Drosophila.* J Mol Evol 33:156–162

Radman M, Wagner R (1986) Mismatch repair in *Escherichia coli.* Annu Rev Genet 20:523–538

Satta Y, Ishiwa H, Chigusa SI (1987) Analysis of nucleotide substitutions of mitochondrial DNAs in *Drosophila melanogaster* and its sibling species. Mol Biol Evol 4:638–650

Stephenson C, Karran P (1989) Selective binding to DNA base pair mismatches by proteins from human cells. J Biol Chem 264:21177–21182

Swofford DL (1991) PAUP: phylogenetic analysis using parsimony, version 3.0s. Computer program distributed by the Illinois Natural History Survey, Champaign

Swofford DL, Olsen GJ (1990) Phylogeny reconstruction. In: Hillis DM, Moritz C (eds) Molecular systematics. Sinauer Associates, Sunderland, MA, pp 411–501

Tajima F, Nei M (1982) Biases of the estimates of DNA divergence

obtained by the restriction enzyme technique. J Mol Evol 18:115–120

Tamura K (1992) The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. Mol Biol Evol 9:814–825

Terachi T, Ogihara Y, Tsunewaki K (1987) The molecular basis of genetic diversity among cytoplasms of *Triticum* and *Aegilops*. 8. Complete nucleotide sequences of the *rbc*L genes encoding H- and L-type rubisco large subunits in common wheat and *Ae crassa* 4X. Jpn J Genet 62:375–387

Topal MD, Fresco JR (1976) Complementary base pairing and the origin of substitution mutations. Nature 263:285–289

Topal MD, DiGuiseppi SR, Sinha NK (1980) Molecular basis for substitution mutations. J Biol Chem 255:11717–11724

Wolstenholme DR, Clary DO (1985) Sequence evolution of *Drosophila* mitochondrial DNA. Genetics 109:725–744

Zurawski G, Clegg MT, Brown AHD (1984) The nature of nucleotide sequence divergence between barley and maize chloroplast DNA. Genetics 106:735–749