Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 9717–9721, October 1995
Evolution

Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions

(nucleotide substitution/neighboring base effect)

BRIAN R. MORTON*

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

Communicated by Michael T. Clegg, University of California, Riverside, CA, July 12, 1995

ABSTRACT **The correspondence between the transversion/transition ratio and the neighboring base composition in chloroplast DNA is examined. For 18 noncoding regions of the chloroplast genome, alignments between rice (*Oryza sativa*) and maize (*Zea mays*) were generated by two different methods. Difficulties of aligning noncoding DNA are discussed, and the alignments are analyzed in a manner that reduces alignment artifacts. Sequence divergence is <10%, so multiple substitutions at a site are assumed to be rare. Observed substitutions were analyzed with respect to the A+T content of the two immediately flanking bases. It is shown that as this content increases, the proportion of transversions also increases. When both the 5′- and 3′-flanking nucleotides are G or C (A+T content of 0), only 25% of the observed substitutions are transversions. However, when both the 5′- and 3′-flanking nucleotides are A or T (A+T content of 2), 57% of the observed substitutions are transversions. Therefore, the influence of flanking base composition on substitutions, previously reported for a single noncoding region, is a general feature of the chloroplast genome.**

An understanding of nucleotide substitution dynamics is of great importance to many molecular evolutionary studies. In analyses of substitutions, including studies of pseudogenes and noncoding DNA, nonrandom frequencies of transitions and transversions have been observed, most frequently with an overrepresentation of transitions (1–3). Substitution models are used widely in molecular evolutionary studies, including molecular systematics studies (4), and are frequently based on this observed bias toward transitions. Of particular interest here is the chloroplast ribulose bisphosphate carboxylase/oxygenase gene (*rbcL*), which has been widely used in plant molecular systematic studies (5, 6). An important feature of all current models is that they assume that the substitution model for a site is independent of neighboring sites.

In a detailed study of the noncoding region immediately downstream of *rbcL* in the chloroplast genome of grasses, it was found that this independence does not hold. Substitutions were observed to be highly dependent upon the composition of flanking bases (7). In this noncoding region, it was noted that there were two distinct sections that had substantially different A+T contents as well as different transversion (Tv)/transition (Ts) biases. The Tv/Ts ratio was observed to be much lower in the section with a lower A+T content (7). Furthermore, the fourfold degenerate sites of the neighboring *rbcL* gene had an even lower Tv/Ts, and the A+T content of the gene was noted to be lower than the noncoding sections. This correlation between A+T content and Tv/Ts was a significant point, leading to the observation that the composition of the two nucleotides immediately flanking a substitution was highly correlated with the Tv/Ts bias of substitutions in the noncoding region. It was shown that, as the A+T content of the two nucleotides flanking the substitutions increased from zero to two, the ratio of Tv/Ts increased from 0.67 (16/24) to 2.2 (87/39). Therefore, the difference in Tv/Ts observed between the two noncoding sections was simply due to the different frequency with which substitutions were flanked by a particular A+T content in the two sections as a result of their different overall A+T contents.

In considering the *rbcL* gene, it was noted that most fourfold degenerate codon groups have a G or C at the second position. Thus, the low Tv/Ts observed at fourfold degenerate sites of *rbcL* may also be the result of this bias in neighboring base composition. This, however, remains to be tested since, due to the genetic code structure noted, a significant number of substitutions with a flanking nucleotide A+T content of two was not observed within the *rbcL* gene of grasses (7).

The assumption of site independence is clearly violated for the noncoding region downstream of *rbcL*. Substitutions are strongly influenced by neighboring base composition. Such influences are potentially of great importance for molecular evolutionary studies. However, it remains to be determined whether this is something unique to the single region studied or a feature of the entire chloroplast genome. The chloroplast genome has been completely sequenced for rice (8), and there are many chloroplast sequences from another grass, maize, many of which contain full or partial noncoding regions. These two species are only 8% diverged in the noncoding region downstream of *rbcL*. Given this low divergence, multiple hits at a site should be rare, and the substitutions in these noncoding regions, which are presumably neutral substitutions, can be used to investigate neighboring base effects. By using substitution data from 19 noncoding regions, this study investigates whether or not neighboring base influences are present throughout the entire chloroplast genome.

This type of analysis, however, is not straightforward. The difficulty is that noncoding regions of the chloroplast, like most noncoding sequence, undergo insertion/deletion mutations (indels) at a high rate. This can make alignments difficult in some areas, and inferred alignments can sometimes differ depending upon the parameters set for common alignment algorithms. The most serious problem, however, is when indels overlap in the two species being examined. In the noncoding region 3′ of *rbcL*, there is a large deletion in the Panicoid grasses, which includes maize (9). Although rice does not share this deletion, it has a shorter deletion that overlaps the larger Panicoid deletion as diagrammed in Fig. 1. The result is that 60 bases of nonhomologous sequence are present in a homologous position in a comparison of rice and maize. This situation was detected only by sequence data from several species (9). With just two sequences, the only way to detect when this has happened is that a section of much decreased similarity exists. As a result of these potential alignment

Abbreviations: NW, Needleman–Wunsch; ML, maximum likelihood; Tv, transversion(s); Ts, transition(s).

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

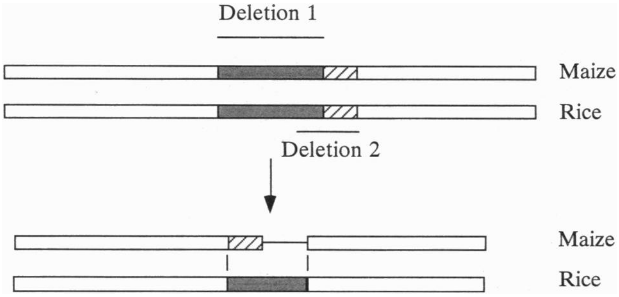

FIG. 1. Overlapping deletions from the noncoding region downstream of *rbcL* (9). Original sequences are drawn such that homologous sequences are aligned. Deletions are noted. Similar shading or hatching represents homology between the two sequences. As can be seen, after the deletions, nonhomologous sections are in a homologous position. Most alignment algorithms will force an alignment between these sections. The result is a stretch of very low similarity in the alignment.

difficulties, an accurate analysis cannot simply align the two sequences and score all differences as substitutions.

This difficulty is well illustrated in a previous study of the noncoding region downstream of *rbcL* that compared only rice and maize (10). The section of decreased similarity was noted and used to draw incorrect conclusions concerning recombination based on the assumption that homologous sequences were being compared (see ref. 9). Therefore, it is essential that alignment artifacts and comparisons of nonhomologous sequences are eliminated or minimized so as not to obscure the true substitution patterns.

The problem of alignment artifacts was dealt with by two methods, each of which was performed under two models. Results from these four approaches could then be compared. The goal of the models was to keep the process as objective as possible; subjective elimination of apparent alignment artifacts was avoided. The first method was to align the noncoding regions by the Needleman–Wunsch (NW) algorithm (11) with two different gap weights. Three *a priori* rules were designed to eliminate sections of extremely low similarity in the alignment. The substitutions remaining after the application of these rules to the alignments were then analyzed. The second method was to use a maximum likelihood (ML) alignment algorithm with an arc probability of 95% (12, 13).

The approach adopted in the present analysis results in the study of only those substitutions occurring within areas where the alignment has strong statistical support and that are likely to represent actual substitution events as opposed to alignment artifacts. Substitution data from the different methods are compared. All four data sets are very similar, and all show the same trend as was observed in the noncoding region 3′ of *rbcL*. The Tv/Ts of a region is correlated with A+T content; regions with a higher A+T content are observed to have a greater proportion of transversions. This correlation appears to be the result of neighboring base influence, which was tested by scoring substitutions with respect to the base composition of the two flanking nucleotides. It is shown that as the A+T content of the 5′- and 3′-flanking nucleotides increases, the Tv/Ts increases significantly. The strong influence of neighboring bases on substitutions is a general feature of the chloroplast genome of the grasses.

MATERIALS AND METHODS

Noncoding sequence data were taken from GenBank. For rice, 18 noncoding regions of the complete genome, all >300 bases in length, were used. These regions and the defining sequences are shown in Table 1. They are numbered by order in the genome using the numbering from the sequence file (8). Noncoding regions for maize were taken from various GenBank files; the accession numbers are given in Table 1. Substitution data from a 19th noncoding region were taken from the multiple alignment presented for the region upstream of *rbcL*, flanked on the other end by *atpB*, in ref. 14. This region is region number 12 in Table 1, again numbered by location in the rice genome.

Table 1. Noncoding regions

Region	Definition*	Accession no(s).
1	3,939–4,486	X60823
2	5,554–6,614	X60823
3	11,591–11,936	X04184
4	13,753–15,059	X05296
5	18,130–19,213	X17318
6	31,245–32,038	X52270
7	35,938–36,308	Y00359
8	41,251–41,850	X58080
9	43,838–44,437	X58080
10	47,498–47,983	X17438, V00174
11	49,663–50,366	X17438
12	53,311–54,094	†
13	60,564–61,564	J04502
14	62,335–63,530	J04502
15	64,304–64,621	J04502
16	64,757–65,197	J04502, X56673
17	68,289–68,798	X05422
18	99,288–100,817	M16907, M16908, X52614
19	108,266–108,711	X13159

*First and last base of each noncoding region from the rice genome sequence.
†From ref. 14.

The 18 noncoding regions, with the exclusion of region 12, were aligned by using the GAP program of the Genetics Computer Group package (15), which uses the NW algorithm. Two alignments were performed. The first used gap weight 3; the second used a gap weight of 5. Both used a zero gap length weight.

Because of the difficulty of alignment artifacts in noncoding DNA, three rules were applied to exclude sections of very low similarity, particularly around gaps. These rules are based on the observation that known homologous regions are >90% similar. The purpose of the rules is to exclude sections of alignment that, given an expectation of 90% similarity, are highly improbable. The rules are

(*i*) Any stretch of 5 or more consecutive bases with no match is excluded.

(*ii*) Any stretch not containing a section excluded by rule *i* that begins and ends in a mismatch, is 10 bases or greater in length, and is 50% similar or less is excluded. The binomial distribution with $P = 0.9$ gives

$$P(10 \text{ occurrences with 5 or less matches}) < 0.002.$$

This probability decreases as the length increases, but similarity is held to 50%.

(*iii*) Gap overhangs are excluded since they depend upon which end of the gap the algorithm places the mismatch. This rule is extended to exclude all aligned bases at the end of gaps prior to a match of at least 2 bases. The beginning and end of the alignment are considered gaps for this purpose.

An example of alignment difficulties encountered is the beginning of region 4 as shown in Fig. 2. A 111-base gap has been introduced to bring the sequences into alignment. However, given an expectation of 90% similarity, it is doubtful that the differences observed at the start of the alignment represent substitutions but are more likely to be artifacts. Applications of the exclusion rules are indicated in the figure.

It is not claimed that only computer artifacts will be excluded by these rules. However, by applying the rules objectively, the

```
               333333   111111111      2 2222  22
  1 -----------GTACAGTTCCCATAAATCCCTTTCTTCCTTTTCTTTTTTGTTCAGAATTGAACAAAGAAATTGCGAAAGA...
                    | |||         |||||| |    ||  ||||||||||||||||||||||||| ||| | ||||
101 CAGTTAGCGGTTGGTACTTCGATCGCGGGCCTTTCCTTTACTTTCTTTTTTGTTCAGAATTGAACAAAGAATTTGGGGAAGA...
```

FIG. 2. The 5′ end of the alignment from region 4. A section of extremely low similarity, which is likely to be an alignment artifact, is evident. The sections excluded by the rules discussed in the text are indicated by the rule number above the excluded difference.

substitutions remaining will not be biased by the exclusion rules and will be in sections of high similarity such that there is a high likelihood that they are real substitution events. If areas of very low similarity are not alignment artifacts but, rather, sections of hypermutability, the difficulty is that multiple substitutions per site render Tv/Ts uninformative. In either case, inclusion of such sections could obscure the real substitutions.

The 18 regions were also aligned by an ML algorithm with a 95% arc probability (12, 13). This method includes only those sections (or arcs on a directed graph) for which the sum of the probabilities of all alignments that contain that arc is >95% of the sum of probabilities of all possible alignments (13). The probabilities themselves are conditional on ML estimates of model parameters. Two different models were used. The first constrained substitution dynamics to be homogeneous across the sequence; the second allowed for heterogeneity across the sequence.

For region 12, substitutions in the multiple sequence alignment were scored with multiple hits inferred from the phylogeny of the grass family (7, 16). Therefore, only a single Tv/Ts was calculated for this region.

Substitutions were scored as Tv or Ts, assuming a single substitution at a site. Neighboring base composition was scored in terms of the number of As and Ts at the two sites. This number was scored as 0, 0.5, 1, 1.5, or 2. The 0.5 and 1.5 were scored when A+T content was conserved at one of the flanking sites but not the second as a result of neighboring substitution events.

RESULTS

As observed in the region downstream of *rbcL* (7), the A+T content of a region is correlated positively with the Tv/Ts ratio in that region. For regions with 15 or more substitutions, a plot of Tv/Ts from the ML alignment, with rate heterogeneity, against the A+T content of the region is shown in Fig. 3. The two variables are correlated (r = 0.47). When the two sections from the 3′ region of *rbcL* as well as the fourfold degenerate sites of *rbcL* (7) are included, the correlation is stronger (r = 0.72), and these new regions fall on the slope from the previous data. The results from the ML alignments when rate homogeneity is modeled are the same.

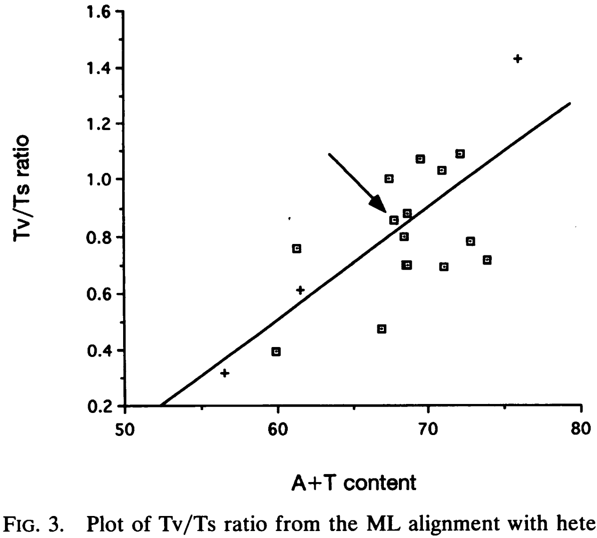

FIG. 3. Plot of Tv/Ts ratio from the ML alignment with heterogeneous substitution dynamics against the A+T content of the region for those regions where at least 15 substitutions were observed. Included are the three regions discussed in the text. These three regions are plotted with a "+." Region 12 is indicated by the arrow.

The results from scoring neighboring base composition and Tv/Ts were totaled over all 19 regions for the NW method. These are given in Table 2. The results of the ML approach, without the addition of the data from region 12, are given in Table 3. All give the same general result and are consistent with the results from the previous study on one region. As the A+T content of the two flanking bases increases, Tv/Ts increases. This increase is significant, from 25% to 57% of the substitutions. By testing the ML method (heterogeneity model) as a 5 × 2 table, the χ^2 is 29.4 ($P < 0.001$). The other three tables are also highly significant (data not shown).

DISCUSSION

The substitution data from the alignments of noncoding chloroplast DNA of rice and maize are strong evidence that neighboring base composition has an influence on mutational dynamics at a site. The Tv/Ts ratio increases steadily as the observed A+T content of the two flanking nucleotides increases (Tables 2 and 3). Therefore, the site dependence previously observed in a single region (7) is present throughout the chloroplast genome. This site dependence violates assumptions made in many types of molecular evolutionary studies. It is apparent from Tables 2 and 3 that the influence of neighboring base composition on substitutions is quite strong. Therefore, this site dependence could have very important implications in molecular evolution.

Several questions that are raised by this study remain to be investigated. One is to determine whether or not the composition of more than simply the immediate neighboring bases influences substitutions. Also, all regions studied are from the chloroplast genome of the grass family. Studies on another family, such as a family of dicot plants, will help to determine how widespread this effect is in plant chloroplasts. The existence in other genomes, such as nuclear and mitochondrial, is also a possibility.

It remains to be determined if the contextual effects observed present difficulties for weighing schemes in phylogenetic analyses or for the estimation of sequence divergence. It is certainly possible that simple weighting schemes place too much importance on presumably less probable substitutions that in some contexts are actually highly probable. In addition, the previous study that demonstrated an influence of flanking bases on substitutions in the chloroplast genome also analyzed noncoding sequences. It was observed in that study that fourfold degenerate sites of the *rbcL* gene in grasses have a

Table 2. Tv/Ts ratio and A+T content of neighboring bases for the NW alignments

A+T	Gap weight = 3			Gap weight = 5		
content	Tv	Ts	Tv/Ts	Tv	Ts	Tv/Ts
0	19	57	0.33	16	55	0.29
0.5	18	30	0.60	16	33	0.48
1	125	206	0.61	120	191	0.63
1.5	57	72	0.84	47	60	0.78
2	148	104	1.42	143	92	1.55

Table 3. Tv/Ts ratio and A+T content of neighboring bases for the ML alignments

A+T	Heterogeneous			Homogeneous		
content	Tv	Ts	Tv/Ts	Tv	Ts	Tv/Ts
0	16	49	0.33	16	48	0.33
0.5	13	27	0.48	14	29	0.48
1	93	166	0.56	90	166	0.54
1.5	45	53	0.85	37	51	0.73
2	98	75	1.31	95	73	1.30

fairly strong bias toward transitions (7). Of the eight fourfold degenerate groups, six have a G or C at the second position. As a result, neighboring base composition will be highly biased toward one or more G or C. Therefore, the low Tv/Ts at fourfold degenerate sites could be a result of this bias in neighboring base composition. However, an influence of neighboring base composition on substitutions in chloroplast coding sequences has yet to be examined.

The mechanism for the influence of neighboring bases on substitutions is unknown. One possibility is that misincorporation and/or proofreading by the DNA polymerase is influenced by the nucleotide 5′ to the site of incorporation. In this case, the observations can be explained by this influence acting on both strands. Sites with a G and/or C both 5′ and 3′ would be biased to transitions during polymerization using either strand as a template. Sites with a G or C on one side and an A or T on the other are biased to transitions in one direction but not the other. Therefore, Tv/Ts would be increased at these sites relative to the first case as is observed. The second possibility is that repair efficiency is influenced by neighboring base composition. In studies on *Escherichia coli*, repair of transversional intermediates is very inefficient in high A+T content regions (17, 18). As of now it is unknown which of these two factors, or both, is responsible. It may be possible to differentiate between the two by a detailed comparison of rate of substitution and neighboring base composition.

One consequence of the neighboring base influence is that regions with a higher A+T content have a higher proportion of transversions than regions with lower A+T content (Fig. 3). This suggests that neighboring influences can be tested by such a correlation. Examination of the correlation in Fig. 3 shows that the four points that were generated using multiple alignments fall almost directly on a straight line. The other points are more widely dispersed. This difference may result from the fact that a multiple alignment was used in the one case as opposed to a pairwise alignment in the others. In the pairwise comparisons, where substitutions are scored at only 8.7% of the homologous sites, the regional A+T content is not necessarily representative of the A+T content of the bases flanking the substitutions. Therefore, a correlation between Tv/Ts and regional A+T content might be weak or nonexistent in pairwise comparisons. In multiple alignments, the proportion of sites where a substitution is observed is much higher, and the two values of A+T content converge. It may be that the correlation would be much stronger if a multiple alignment were used for each region. Therefore, if neighboring influences are examined by using a correlation, multiple sequence alignments are preferable to pairwise alignments, especially when alignment difficulties are taken into consideration as discussed below.

This study also demonstrates some of the difficulties in analyzing noncoding DNA. As discussed earlier, the indels that occur over evolution of noncoding sequences raise problems in terms of analyzing substitutions. Obviously, it is essential to compare only homologous sequences. However, many computer algorithms now available will force an alignment even if overlapping indels have occurred. In these cases, a significant number of incorrect substitutions will be postulated.

A second difficulty was also observed. In some cases, a different gap weight resulted in a different alignment in sections. In some cases, the alignment is almost certainly incorrect. To simply alter the alignments introduces a serious degree of subjectivity into the procedure. On the other hand, to accept all differences given by the computer alignment will certainly lead to incorrect conclusions.

The solutions offered here are intended to keep the procedure as objective as possible as well as easily repeatable. The ML procedure allows one to assign a threshold value, here 95%, such that aligned regions of low likelihood are excluded. The second attempt was to generate rules *a priori*, in this case based on a very simple binomial calculation. By assigning the rules prior to the analysis, no subjective decision is made about any section; the application of the rules is performed objectively. The rules themselves are subjective and can be altered, just as gap weight can be altered *a priori*, and the procedure is repeated. However, some degree of subjectivity is inevitable as in all alignment methodology, since parameters must always be chosen. By assigning the rules *a priori* to exclude sections of alignment with low probability based on known rate of divergence, no bias is introduced by subjectivity at the level of exclusion of particular sections.

The rules proposed are based on the observation that known homologous noncoding sequences are <10% divergent. They are designed to eliminate sections that are highly improbable given the expectation of 90% similarity. It should be noted that over all of the regions used in the present study, the level of divergence is similar to that observed in the previous study: only 8.7% of all sites show a difference in the heterogeneous ML alignment (data not shown). This divergence level also supports the assumption that multiple hits are rare.

While the rules are designed to exclude aligned sections that are likely to be alignment artifacts, it is possible that some sections of "true" alignment will be eliminated. However, no bias should be introduced into the analysis. Since exclusion is independent of the specific differences in that section, the substitutions scored should be unbiased representatives of the substitution process. The only way in which a bias could have been introduced is if highly dissimilar sections are simply hypermutable areas and the substitution dynamics of these sections are different than those in the highly similar sections. This case, however, is highly unlikely. There is no evidence that substitution dynamics vary in such a manner. There is, however, much evidence for alignment artifacts, which seem to be much more likely as a result. Even in the unlikely case of a variation in substitution dynamics, it is apparent that the highly similar sections have a strong neighboring base influence.

The data presented here are strong evidence that the neighboring base influence observed previously in a single noncoding region (7) is present throughout the entire genome. It is seen that the Tv/Ts ratio varies substantially among sites and that this is correlated with flanking base composition. These results indicate that the molecular evolutionary dynamics of chloroplast DNA is much more complicated than previously supposed.

I greatly appreciate the help of Jeff Thorne, who performed the ML alignments. I also thank Richard Morton, Brandon Gaut, Michael Clegg, and Richard Whitkus for discussion. Ken Wolfe provided a list of accession numbers for noncoding sequences of maize. This work was supported in part by National Institutes of Health Grant GM 45144.

1. Brown, W. M., Prager, E. M., Wang, A. & Wilson, A. C. (1982) *J. Mol. Evol.* **18,** 225–239.
2. Gojobori, T., Li, W. H. & Graur, D. (1982) *J. Mol. Evol.* **18,** 360–369.
3. Li, W. H., Wu, C. I. & Luo, C. C. (1984) *J. Mol. Evol.* **21,** 58–71.
4. Swofford, D. L. & Olsen, G. J. (1990) in *Molecular Systematics,* eds. Hillis, M. & Moritz, C. (Sinauer, Sunderland, MA), pp. 411–501.

5. Clegg, M. T. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 363–367.
6. Chase, M. W., Soltis, D. E., Olmstead, R. G., Morgan, D., Les, D. H. *et al.*, (1993) *Ann. Mo. Bot. Gard.* **80,** 528–580.
7. Morton, B. R. & Clegg, M. T. (1995) *J. Mol. Evol.*, in press.
8. 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) *Mol. Gen. Genet.* **217,** 185–194.
9. Morton, B. R. & Clegg, M. T. (1993) *Curr. Genet.* **24,** 357–365.
10. Bowman, C. M., Barker, R. F. & Dyer, T. (1988) *Curr. Genet.* **14,** 127–136.
11. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48,** 443–453.
12. Thorne, J. L., Kishino, H. & Felsenstein, J. (1991) *J. Mol. Evol.* **33,** 114–124.
13. Thorne, J. L. & Churchill, G. A. (1995) *Biometrics* **51,** 100–113.
14. Golenberg, E. M., Clegg, M. T., Durbin, M. L., Doebley, J. & Ma, D. P. (1993) *Mol. Phylogenet. Evol.* **2,** 52–64.
15. Devereux, J., Haeberli, H. & Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387–395.
16. Duvall, M. R. & Morton, B. R. (1995) *Mol. Phylogenet. Evol.*, in press.
17. Radman, M. & Wagner, R. (1986) *Annu. Rev. Genet.* **20,** 523–538.
18. Jones, M., Wagner, R. & Radman, M. (1987) *Genetics* **115,** 605–610.